OPEN ACCESS
Received: 22 -08 -2026
Accepted: 24- 09-2026

الآداب
للدراسات اللغوية والأدبية

The Impact of Professional Experience on Simultaneous Interpreting Fluency and Terminology Fidelity: An Experimental Study within the Context of Saudi Vision 2030 Events

Dr. Mansour Ali Al-Maswari *
Mansoury2000@yahoo.com

Dr. Redhwan Qasem Ghaleb Rashed **
ReRashed@uoh.edu.sa

Dr. Munir Ahmed Al-Aghberi ***
maghberi@yahoo.com

Abstract

Saudi Vision 2030 positions the Kingdom as a regional hub for international conferences, increasing demand for high-quality Arabic–English simultaneous interpreting. This study examines the effects of professional experience and elapsed time on interpreters' performance, focusing on fluency, terminology fidelity, comprehension equivalence, and delivery naturalness. A quasi-experimental design was employed with 60 conference interpreters: 30 experienced professionals with at least three years of elite conference experience and 30 novices with less than one year of postgraduate conference experience. Participants completed a 25-minute simultaneous interpreting task based on technology and investment conference scenarios, evaluated independently by five certified interpreters using a standardized 100-point rubric. Repeated-measures ANOVA revealed significant main effects of professional experience and elapsed time, alongside a large experience-by-time interaction ($F(1, 58) = 118.76$, $p < .001$, $\eta p^2 = .672$). Experienced interpreters maintained consistent fluency, whereas novices exhibited a 41.6% decline in fluency during the second half of the task, accompanied by lexical deterioration. Hierarchical multiple regression further indicated that booth experience and proactive cognitive strategies jointly explained 81.2% of the variance in sustained performance ($R^2 = .812$). These findings underscore the importance of structured practical training, strategic cognitive preparation, and curriculum reform to enhance interpreting competence and support national workforce localization objectives.

Keywords: Interpreter experience, Simultaneous interpreting fluency, Terminology fidelity, Conference interpreting, Saudi mega-events.

* Assistant Professor of Comparative Literature, Department of English, College of Arts & Applied Sciences, Amran University, Amran, Yemen.

** Associate Professor of English Literature & Comparative Studies, Department of English, College of Arts, University of Ha'il, Ha'il, Saudi Arabia.

*** A professor of English Literature, Department of English, Albaydha University, Al Bayda, Yemen.

Acknowledgement: This research received grant no. (968) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

Cite this article as: Al-Maswari, M. A. Rashed, R. Q. G. Al-Aghberi, M. A. (2026) The Impact of Professional Experience on Simultaneous Interpreting Fluency and Terminology Fidelity: An Experimental Study within the Context of Saudi Vision 2030 Events, *Arts for Linguistic & Literary Studies*, *8*(4): 222 -261 https://doi.org/10.53286/3wbesf52

© This material is published under the license of Attribution 4.0 International (CC BY 4.0), which allows the user to copy and redistribute the material in any medium or format. It also allows adapting, transforming or adding to the material for any purpose, even commercially, as long as such modifications are highlighted and the material is credited to its author.

الآداب
للدراسات اللغوية والأدبية

OPEN ACCESS
تاريخ الاستلام: 2026/08/22م
تاريخ القبول: 2026/09/24م

أثر الخبرة المهنية على طلاقة الأداء والدقة الاصطلاحية في الترجمة الفورية: دراسة تجريبية في سياق الفعاليات لرؤية السعودية 2030

د. منصور علي المسوري [*]
Mansoury2000@yahoo.com

د. رضوان قاسم غالب راشد [**]
ReRashed@uoh.edu.sa

د. منير أحمد الأغبري [***]
maghberi@yahoo.com

الملخص:

تهدف هذه الدراسة إلى تحليل أثر الخبرة المهنية والوقت المنقضي في أداء المترجمين الفوريين، مع التركيز على الطلاقة والدقة الاصطلاحية والتكافؤ الدلالي وجودة الإلقاء. إذ في ضوء رؤية السعودية 2030، تتزايد الحاجة إلى الترجمة الفورية في المؤتمرات الدولية الكبرى، مما يستدعي تعزيز كفاءة المترجمين وضمان دقة التواصل. اعتمدت الدراسة تصميمًا شبه تجريبي عامليًا مختلطًا (2×2)، بمشاركة 60 مترجمًا فوريًا، موزعين بالتساوي بين خبراء يمتلكون خبرة لا تقل عن ثلاث سنوات ومبتدئين تقل خبرتهم عن عام واحد. خضع المشاركون لمحاكاة ترجمة فورية من العربية إلى الإنجليزية استمرت 25 دقيقة، وقُيّم أداؤهم بواسطة خمسة محكمين معتمدين وفق مقياس معياري من 100 نقطة. أظهر تحليل التباين للقياسات المتكررة تأثيرات دالة إحصائيًا للخبرة المهنية والوقت، إلى جانب تأثير تفاعلي كبير $F(1, 58) = 118.76$، $p < .001$، $\eta p^2 = .672$. حافظ المترجمون الخبراء على استقرار الطلاقة، بينما انخفضت طلاقة المبتدئين بنسبة 41.6% خلال النصف الثاني من المهمة، مصحوبة بتراجع اصطلاحي ملحوظ. كما فسرت الخبرة العملية في كبائن الترجمة واستراتيجيات المراقبة المعرفية الاستباقية 81.2% من التباين في استدامة الأداء ($R^2 = .812$). وتؤكد النتائج أهمية تطوير التدريب العملي والمناهج الجامعية بما يعزز الكفاءة المهنية ويدعم مستهدفات التوطين في قطاع الترجمة.

الكلمات المفتاحية: خبرة المترجم، طلاقة الترجمة الفورية، الدقة الاصطلاحية، ترجمة المؤتمرات، الفعاليات الكبرى في السعودية.

---

[*] أستاذ مساعد في الأدب المقارن، قسم اللغة الإنجليزية، كلية الآداب والعلوم التطبيقية، جامعة عمران، عمران، اليمن.

[**] أستاذ مشارك في الأدب الإنجليزي والدراسات المقارنة، قسم اللغة الإنجليزية، كلية الآداب، جامعة حائل، حائل، المملكة العربية السعودية.

[***] أستاذ في الأدب الإنجليزي، قسم اللغة الإنجليزية، جامعة البيضاء، اليمن.

شكر وتقدير: حصل هذا البحث على منحة رقم (968) من المرصد العربي للترجمة (التابع للمنظمة العربية للتربية والثقافة والعلوم - ألكسو)، الذي يحظى بدعم هيئة الأدب والنشر والترجمة في المملكة العربية السعودية.

للاقتباس: المسوري، م. ع. راشد، ر. ق. غ. الأغبري، م. أ. (2026). أثر الخبرة المهنية على طلاقة الأداء والدقة الاصطلاحية في الترجمة الفورية: دراسة تجريبية في سياق الفعاليات لرؤية السعودية 2030، *الآداب للدراسات اللغوية والأدبية*، *8*(4): 222-261 https://doi.org/10.53286/3wbesf52

© نُشر هذا البحث وفقًا لشروط الرخصة Attribution 4.0 International (CC BY 4.0)، التي تسمح بنسخ البحث وتوزيعه ونقله بأي شكل من الأشكال، كما تسمح بتكييف البحث أو تحويله أو الإضافة إليه لأي غرض كان، بما في ذلك الأغراض التجارية، شريطة نسبة العمل إلى صاحبه مع بيان أي تعديلات أُجريت عليه.

1. Introduction

1.1 Contextual Foundation: The Strategic Pivot to Mega-Conferences

The Kingdom of Saudi Arabia is in the midst of an ambitious economic, institutional, and cultural restructuring based on Saudi Vision 2030 (Kingdom of Saudi Arabia, 2016). One of the primary objectives of this national strategy is to make Riyadh the top conference, exhibition, and diplomatic hub in the Middle East in addition to other well-known global commercial hubs like Singapore and Dubai (Al-Suhaim, 2024). The Kingdom hosts a growing schedule of international events to attract foreign direct investment, foster technology transfer, and grow regional diplomacy through capital deployment managed by the Saudi Conventions and Exhibitions General Authority.

Within this conference ecosystem, flagship forums such as the LEAP technology exhibition, the Future Investment Initiative (FII), and multilateral Hajj coordination summits serve as critical empirical settings not merely due to their economic scale, but because they introduce an acute psycholinguistic challenge for Arabic–English simultaneous interpreters (Kingdom of Saudi Arabia, 2016; Al Zahran, 2021; Al-Suhaim, 2024). Consequently, the central research problem of this study is to determine how accumulated professional expertise protects simultaneous interpreters against temporal fluency decay and lexical breakdown when processing structurally divergent Arabic-to-English speech embedded with newly institutionalized, non-standardized terminology. The exact gap in previous interpreting research is the lack of controlled experimental studies that quantify micro-temporal acoustic degradation (across sustained 25-minute turns) alongside domain-specific lexical retrieval accuracy in Arabic–English simultaneous interpreting. Beyond simply comparing experienced and novice practitioners in a Saudi setting, this study contributes a predictive psycholinguistic model demonstrating how automated institutional schemas and proactive syntactic strategies prevent working memory saturation under sustained temporal pressure.

1.2 The Empirical Reality and Institutional Problem

The international conference industry is an important sector of the national economy, but there is a major mismatch in the capacity of the local labor market and the demand for the sector. About 800 students are awarded degrees each year in translation and language studies in 14 specialized academic departments in Saudi universities. However, pedagogical assessments indicate that the vast majority of these graduates complete their degrees without systematic, practical exposure to simultaneous interpreting in an ISO-standard booth (Al-Suhaim, 2024; Aluthman & Al-Buraidi, 2024). From the beginning of the undergraduate programs in the Kingdom, the focus is mainly on written translation, literary analysis, and basic consecutive skills, which is a sharp contrast to the necessary skills in conferences (Aluthman & Al-Buraidi, 2024). Summarizing the results of

the analysis presented in Table 1, there is a disparity between institutional preparation and industry needs, which gives rise to systemic operational bottlenecks in various economic and performance indicators.

Table 1. Structural Disparities Between Domestic Training Baselines and Conference Market Demands in Saudi Arabia

Operational Metric	Domestic Baseline	International / Market Requirement	Resulting Structural Deficit
Annual University Output	A round of 800 translation majors from 14 departments	High market demand for conference interpreters	72% graduate without simultaneous booth exposure
Sector Revenue Generation	Written translation: SAR 800 per diem	Simultaneous booth: SAR 3,000 per diem	82% of industry revenues stem from conference booths
Workforce Sourcing	Local graduates face 72% agency rejection rate	Qualified conference-ready booth professionals	65% of conference personnel imported at 5x cost
Technical Infrastructure	SAR 120M invested in booths and headsets	Full deployment across recurring mega-events	28% infrastructure underutilization due to talent shortages
Event Operational Risk	Novice fluency breakdown and terminology errors	Seamless, accurate Arabic-to-English delivery	SAR 2.5M in lost credibility and cancellations per event
Automated Translation (AI)	62% accuracy on local institutional terms	95%+ required for high-stakes investment plenaries	AI requires experienced human oversight to avoid failure

Note. Illustrative sector benchmarks and industry estimates compiled from regional policy reports, commercial translation agency surveys, and empirical interpreting audits: Kingdom of Saudi Arabia (2016); Aluthman and Al-Buraidi (2024); Al-Suhaim (2024); Alharbi (2025); and NES Fircroft (2025).

This pedagogical conflict is directly opposed to commercial realities. Simultaneous conference interpreting accounts for about 82% of the total revenue generated in the translation industry in the country, and the daily price of this interpretation service is SAR 3,000, whereas the price of the standard written service is SAR 800 (Al-Suhaim, 2024; NES Fircroft, 2025). In the domestic commercial sector, private translation agencies reject approximately 72% of local graduate candidates because they lack the cognitive stamina,

delivery fluency, and terminological competence required in live conference settings (Aluthman & Al-Buraidi, 2024; NES Fircroft, 2025). This means that for the organizers to access a pool of simultaneous interpreters, they need to recruit at least 65% of their staff from European and Asian markets at five times the local cost of compensation, putting them at a severe disadvantage in achieving the national goal of Saudization (NES Fircroft, 2025).

In addition, there are still huge amounts of investment in the physical conference facilities that are underutilized. The Saudi Event Authority and associated parties have spent more than SAR 120 million to acquire top-of-the-line soundproof booths and wireless conferencing headsets, but with a shortage of qualified conference interpreters in the country, it is estimated that 28% of this equipment is currently unused. According to commercial event sector estimates, when inexperienced interpreters are deployed in high-stakes environments, rapid cognitive fatigue frequently triggers hesitation chains and delivery breakdowns, risking substantial financial and reputational costs for event organizers (Al-Suhaim, 2024; NES Fircroft, 2025).

This gap has not been filled by automated translation services. Artificial Intelligence engines and automated speech-to-text tools have been tested in exhibition spaces; however, empirical studies showed that current AI systems are not 100% accurate in processing the local Saudi governmental, socio-economic, and technical lexicon (Alharbi, 2025). The regional administrative acronyms, terms used for hijri calendar, and Vision 2030 program labels are typical examples of such systems, where the translation errors must be corrected quickly by experienced human interpreters. Absent experienced experts managing these processes, automated solutions can create operational risks in critical conferences (Alharbi, 2025).

1.3 Theoretical Gaps in Interpreting Studies

The cognitive aspects of simultaneous interpretation have been extensively studied with various models, such as Gile's Effort Models (Gile, 2016) and Baddeley's multicomponent working memory model (Baddeley, 2012). But there are some empirical and conceptual gaps in the literature, especially for the Arabic-English language pair in the context of Middle Eastern conferences.

While the general directional superiority of experienced practitioners over novice interpreters may appear intuitively predictable, interpreting literature lacks rigorous empirical quantification of performance degradation curves, acoustic metrics, and cognitive saturation thresholds under high-stress regional conference conditions (Ericsson, 2006; Kööp, 2019). Previous studies have largely relied on qualitative self-reports rather than controlled laboratory simulations tracking micro-temporal fluency decay (e.g., phonation-time ratios and pause distributions) and precise lexical retrieval deficits across specialized institutional domains (Seeber, 2011; Timarova et al., 2014). Establishing exact empirical baselines, such as the proportional collapse of novice fluency during sustained 25-minute intervals and regression-tested variance models, is vital for moving beyond

intuitive assumptions to provide actionable, evidence-based frameworks for university curriculum restructuring, national labor Saudization mandates, and the optimization of multi-million-riyal conference infrastructure.

First, the research that has been conducted on conference interpreting in the Arab region is mainly based on subjective self-reporting surveys, post hoc satisfaction questionnaires, and qualitative case reports (Al-Suhaim, 2024; Al Zahran, 2021). Only a few studies have carried out laboratory-controlled experimental simulations, while keeping acoustic delivery patterns and silent pauses constant and tracking the performance degradation under standard booth conditions.

Secondly, translating from Arabic into English is a very challenging task owing to the differences in word order between the two languages (Verb-Subject-Object), morpho-syntactic density, and the presence of sudden shifts in the register and language used from Modern Standard Arabic to the special vocabulary used in administration (Aal-Hajiahmed, 2023). In the West, the cognitive requirements for sustained simultaneous interpreting during conferences have been the subject of well-defined empirical baselines (Timarova et al., 2014); however, the cognitive capacity imposed on Arabic-to-English simultaneous interpreting has not been well documented.

Thirdly, the specific terms of Saudi Vision 2030 have not been sufficiently studied in the field of psycholinguistics. Specialized contextual knowledge is required for institutional concepts like Nitaqat employment quotas, the milestones of the National Transformation Program, capital allocations from the Public Investment Fund, and targets for the Vision Realization Program. There has been little experimental examination of how such specialized lexicon is retrieved, stored, and expressed in the context of high cognitive load, as a function of professional experience.

Lastly, all international conferences require 20- to 30-minute rotations of the booths to prevent cognitive overload, but there are not many experimental studies on this topic that have systematically monitored the influence of continuous 25-minute rotations on performance between untrained and experienced interpreters in this region.

1.4 Research Objectives and Questions

To respond to these empirical and pedagogical questions, this study will compare the performance of qualified full-time conference interpreters with that of young graduates from the University who are new to the field, using a controlled simulation in a laboratory setting to create a representative LEAP and FII to which the participants will be compared.

The main goal is to investigate the effect of professional experience on the fluency of delivery, the accuracy of the terminology, semantic accuracy, and quality of presentation when cognitive loads are sustained. The secondary goal is to analyze specific cognitive processing and self-monitoring strategies of skilled practitioners, based on which the reform of higher education curricula and training of domestic workers is done empirically. Rather than treating temporal fatigue, Arabic–English structural asymmetry, Saudi Vision 2030 terminology, and coping strategies as isolated topics, this study integrates them into a single unified psycholinguistic mechanism: in Arabic-to-English simultaneous interpreting, intrinsic structural load (Verb-Subject-Object to Subject-Verb-Object reordering) continuously taxes working memory; when compounded by newly coined, non-standardized Saudi Vision 2030 institutional neologisms (e.g., Nitaqat, VRPs, and sovereign fund structures that lack established equivalents in traditional bilingual dictionaries) over a sustained 25-minute turn, cognitive saturation occurs unless the interpreter deploys proactive coping strategies. Consequently, the Saudi Vision 2030 context is not merely decorative background, but a high-density lexical stressor essential to testing domain-specific schema retrieval. Furthermore, moving beyond the predictable finding that experts outperform novices, this study identifies where performance diverges most sharply, specifically, during the second temporal interval (T2) and within culturally embedded labor-governance terminology. Finally, because mental processes are not measured via direct neurophysiological instrumentation in this protocol, the coping strategies examined (anticipation, syntactic chunking, and tactical condensation) are rigorously framed as observable linguistic-behavioral indicators and retrospective metacognitive accounts rather than direct neural measurements.

This study investigates three specific research questions:

- Research Question 1: To what extent does professional interpreter experience determine the sustainability of simultaneous interpreting fluency, measured by unfilled pause distribution, hesitation markers, and phonation-time ratios, over a continuous 25-minute conference simulation replicating LEAP and FII sessions?
- Research Question 2: How significantly does professional experience differentiate terminology fidelity when rendering specialized Saudi business, governmental, and technical discourse from Arabic into English?
- Research Question 3: Which self-reported coping strategies and observable target-text restructuring behaviors (such as prospective anticipation, syntactic chunking, and tactical condensation) distinguish experienced conference interpreters from novices during sustained interpreting intervals?

2. Theoretical Framework and Literature Review

2.1 Theoretical Foundations

2.1.1 Gile's Effort Models and the Dynamics of Cognitive Saturation

When it comes to psycholinguistic implementation of simultaneous interpreting, there are strict processing restrictions that are traditionally described with the use of Daniel Gile's Effort Models (Gile, 2016). Gile defines the overall cognitive load of simultaneous interpreting through the foundational equation:

$$SI = L + M + P + C$$

In this formulation, simultaneous interpreting (SI) represents the coordinated sum of four non-automatic, resource-consuming components:

1. **The Listening and Analysis Effort (*L*):** Encompasses auditory perception, phonological decoding, syntactic parsing, and semantic extraction from the incoming source language stream
2. **The Short-Term Memory Effort (*M*):** Controls the temporary storage of information for the latency interval (ear-to-voice span [EVS]) between the time the source is heard and the time the target is voiced.
3. **The Speech Production Effort (P):** Involves Lexical retrieval, syntactic reformulations, phonetic planning, prospective target monitoring, and articulatory output.
4. **The Coordination Effort (*C*):** Serves as an executive metacognitive governor that dynamically shifts attention across the competing demands of *L*, *M*, and *P*.

Gile (2016) states that simultaneous interpreting works close to cognitive saturation thresholds in accordance with his "tightrope hypothesis. The total required cognitive capacity (TR) should not exceed the total available processing capacity (TA), that is:

$$TR = L + M + P + C \leq TA$$

Simultaneously, each individual component must not exceed its dedicated capacity threshold ($L \leq AL$, $M \leq AM$, $P \leq AP$, $C \leq AC$).

The allocation demands for L and M increase quickly when there is any specialized high-density terminology in the incoming speech, such as the institutional programs of Saudi Vision 2030 or financial instruments. When the interpreter does not have automated lexical associations, the Coordination Effort (C) uses an excessive cognitive load to perform semantic disambiguation and memory maintenance. In turn, resources for the Production Effort (P) become lower than the requirements for an operational system. This cognitive overload causes breakdowns in operations and manifests itself as filled hesitations, lexical substitutions, silent pauses longer than the prosodic norm, or dropping the message

2.1.2 Working Memory Architecture and Cognitive Load Theory

The cognitive processes that underlie Gile's formulation are in line with the multicomponent working memory model proposed by Alan Baddeley and the Cognitive Load Theory (CLT) by John Sweller (Uçar, 2025). Baddeley views working memory as an attentional control system, with subordinate slave systems that are the phonological loop, visuospatial sketchpad, and episodic buffer.

In simultaneous conference interpreting, the central executive coordinates the attentional focus and helps to prevent acoustic interference and task switching in the comprehension and speech production process. During the ear-voice span, the phonological loop stores the phonetic trace of the source discourse, and the episodic buffer combines information from LTM schemas with real-time situational discourse (Uçar, 2025). Cognitive Load Theory categorizes processing load into three distinct channels:

- **Intrinsic Cognitive Load:** The baseline complexity inherent to the semantic and structural features of the source text.
- **Extraneous Cognitive Load:** Mental strain generated by external presentation conditions, such as rapid speaker delivery rates, acoustic reverberation, booth ambient noise, or poor audio fidelity.
- **Germane Cognitive Load:** Mental resources dedicated to constructing, refining, and automating long-term cognitive schemas for rapid task resolution.

The intrinsic cognitive load (ICL) is high when translating between Arabic and English due to linguistic directionality (Aal-Hajiahmed, 2023). The modern standard version of Arabic often uses the VSO (Verb-Subject-Object) sentence structure, Root-and-pattern morphology, and coordinate clausal parataxis (waw al-hal, waw al-atf). English, on the other hand, uses analytic syntax, subordinate hypotaxis, and Subject-Verb-Object (SVO) ordering. During the translation of Arabic to English, an interpreter is required to retain the initial verb complexes in working memory until the markers of the agent and/or patient have been provided, which increases the ear-voice span and overloads the phonological loop (Al Zahran, 2021). When unfamiliar institutional jargon adds to the intrinsic load, and rapid delivery adds to the extraneous load, working memory can become saturated and cause instability of both target fluency and lexical fidelity.

2.1.3 Deliberate Practice and Expertise Theory

The differences in cognitive performance between novice and expert interpreters can be explored with the help of K. Anders Ericsson's model on the development of deliberate practice and expertise. Elite performance isn't the result of passive, length-of-service, but of extended periods of well-structured, effortful practice, focused on specific performance problems, says Ericsson. The 10,000-hour milestone is a threshold that helps distinguish between more advanced practitioners and less advanced ones, as it refers to the

automation or reduction of the complexity of more basic linguistic tasks, as well as the consolidation of more complex cognitive schemas, as noted by Al Zahran (2021).

Experienced interpreters automate the morphological parsing and repetitive syntactic restructuring tasks during the conference through deliberate practice, thus saving mental resources during the Listening (L) and Production (P) process (Gile, 2016). These structural automatisms leave space for central executive resources that supports prospective monitoring of the text, syntactic anticipation, contextual modifications, and discourse-level coherence. Novice interpreters, who generally complete a limited number of booth hours throughout a typical undergraduate degree program, use non-automated translation strategies that are effortful (Rendón & Moser-Mercer, 2000; Setton & Dawrant, 2016). Novices very quickly exhaust their thinking capacity when called upon by sustained mental demands, resulting in performance degradation and communication failure.

2.2 Empirical Literature Review

2.2.1 Fluency Sustainability and Temporal Performance Decay

One of the key indicators of cognitive balance and professionalism in simultaneous interpreting is speech fluency. When interpreting studies, the following acoustic parameters are usually used to represent delivery fluency: length of silent pauses (silent pauses below 3 seconds are considered normal while those above 3 seconds are disruptive); phonation-time ratios; speech delivery rates; and filled pauses or "umms" (self-repairs, false starts, "uh"s, "um"s) (Wenda & Nur Salam, 2025).

Empirical studies demonstrate that expert conference interpreters maintain a significantly more consistent acoustic output throughout sustained working turns, whereas novices exhibit progressive temporal degradation over time (Moser-Mercer et al., 1998; Timarova et al., 2014). Timarova et al. (2014) demonstrated that experienced practitioners exhibit superior working-memory executive control, particularly in updating and inhibiting interference, which correlates directly with fewer disfluencies and higher prospective monitoring during continuous conference tasks. Skilled interpreters regulate their latency adaptively, maintaining an ear-voice span of approximately 2.5 to 4.0 seconds to resolve syntactic ambiguities without disrupting speech flow (Setton & Dawrant, 2016; Kööp, 2019).

Conversely, novice interpreters display erratic ear-voice spans, frequently lagging behind the speaker until their phonological loop reaches saturation, which results in disruptive silent pauses exceeding 3 seconds and extended hesitation chains (Tissi, 2000; Kööp, 2019). Recent scholarship by Uçar (2025) and Dong and Chen (2025) further confirms that working-memory regulation and metacognitive strategy deployment account for substantial variance in translation and interpreting fluency under sustained cognitive load, with professional practitioners deploying anticipatory and monitoring strategies far more systematically than

trainees. However, few empirical studies have traced the exact micro-temporal degradation of fluency in continuous Arabic-to-English conference interpreting, necessitating controlled experimental verification.

2.2.2 Terminology Fidelity in Specialized Institutional Contexts

Terminology fidelity indicates an interpreter's competence to accurately capture the meaning of specialized, domain-specific terms; to retain the nuances of meaning; and to maintain semantic consistency and legal/economic correspondence within the target linguistic register. In special conference settings, terminological imprecision can lead to misunderstanding of institutional requirements or adversely impact cross-border transactions (Al Zahran, 2021).

Psycholinguistic research shows that lexical retrieval under time pressure depends on automated semantic networks (Liu et al., 2004; Aal-Hajiahmed, 2023). Low-frequency institutional terms are frequently postponed when working memory is filled with structural syntactic transformations. Lexical choice becomes more complex in the case of Arabic-to-English interpreting, as there is culturally specific and government-dictated institutional terminology as well (Aal-Hajiahmed, 2023).

In the Saudi Vision 2030 paradigm, the conference interpreter is confronted with a special lexicon that integrates the administrative policy lexicon, sovereign investment structures, and giga-project lexicon (Al-Suhaim, 2024). Key examples include:

- Labor governance programs such as *Nitaqat* (Saudization quota bands) and the *Ajeer* portal;
- Public financial instruments including Public Investment Fund (PIF) Tier-1 asset deployments and Vision Realization Program (VRP) allocations;
- Technical terminology linking NEOM digital infrastructure, the OXAGON manufacturing grid, and THE LINE urban design;
- Temporal references tied to the Hijri-Gregorian dual administrative calendar.

Empirical work by Aal-Hajiahmed (2023) demonstrated that novices experience substantially elevated cognitive load and diminished terminological precision when processing dense morpho-syntactic structures in the Arabic-to-English direction. Avoiding this breakdown is a method that experienced practitioners use, and involves having a set of semantically equivalent phrases ready to use and dividing up the text into chunks, retaining the specificity of the terms even in the face of a dense presentation.

Likewise, Wenda and Nur Salam (2025) indicated that as the speaker is an inexperienced one, his fidelity will be significantly reduced when he uses administrative acronyms and rapid transitions. When there are no institutional equivalents that are automated, Aluthman and Al-Buraidi (2024) noted that novice interpreters often use literal calques or phonetic approximations, which may alter the institution's meaning.

2.2.3 AI Augmentation and Machine Translation in the Conference Booth

The integration of automated speech recognition (ASR), large language models (LLMs), and computer-assisted interpreting (CAI) tools has introduced new workflows into the conference ecosystem (Alharbi, 2025). Emerging platforms promise real-time automated transcription, instant terminology lookups, and direct artificial intelligence captioning.

However, the local institutional discourse has been found to pose some performance constraints in the processing of the discourse by automated systems in recent evaluations (Alharbi, 2025). In the controlled Future Investment Initiative simulations, the speech-to-text (STT) systems were only able to attain 62% in terms of terminology fidelity for the specialized Saudi business-government lexicon (Alharbi, 2025). Automated engines do not perform as well for Arabic regional phonetics, code-switching within Arabic and foreign business languages, or Arabic institutional acronyms.

If the automated systems receive inaccurate source information and the information is not corrected by humans (human practitioners), the errors will be carried over to the target outputs (Alharbi, 2025). In addition, handling automated CAI interfaces creates an additional visual channel, adding to extraneous cognitive load. Automated glossaries are used by experienced interpreters when needed, without impacting acoustic tracking; inexperienced interpreters, on the other hand, often find themselves overwhelmed by the amount of information on the screen and have to pause their output to look at the screen and consult the glossary (Alharbi, 2025).

2.2.4 Interpreting Pedagogy and the Middle Eastern Training Gap

This is due to the lack of conference interpreters in the Middle East, which is a structural gap in university curricula in the field of translation (Setton & Dawrant, 2016; Aluthman & Al-Buraidi, 2024). The International Association of Conference Interpreters (AIIC) sets a minimum of 400 hours of supervised booth experience for baseline professional certification, but in Saudi Arabia, the average number of booth hours of training is 28 hours of total experience (Setton & Dawrant, 20216) .

Undergraduate translation programs across the 14 specialized departments in Saudi universities primarily emphasize written translation, comparative stylistics, and literary analysis rather than intensive booth training (Al-Suhaim, 2024; Aluthman & Al-Buraidi, 2024). Usually, when oral interpreting is being taught, it is limited to consecutive interpreting without taking notes or sight translating non-specialized general texts (Aluthman & Al-Buraidi, 2024).

Therefore, students leave the market without acquiring the automated cognitive schemas, resilience of working memory, or prospective self-monitoring needed for simultaneous delivery. A lack of pedagogical

approaches is responsible for the 72% rejection rate by commercial agencies and the need for event organizers to hire foreign interpreters (NESFircroft, 2025).

To overcome this gap, training systems need to be grounded in empirical training designs incorporating high-intensity booth simulations, specific institutional corpora, and structured cognitive coping strategies

.2.3 Synthesis and Conceptual Research Model

The theoretical models and literature review confirm that the performance of simultaneous interpreters depends on the interplay between professional experience, working memory capacity, domain-specific lexical knowledge, and the duration of the cognitive load.

In Table 2, the key variables of the conceptual model of this investigation are listed, along with the underlying theoretical mechanisms and the empirical gap that the experimental design aimed to fill.

Table 2. Conceptual Integration of Theoretical Frameworks, Experimental Variables, and Research Gaps

Dimension	Theoretical Underpinning	Operational Variables	Novice Performance Profile	Expert Performance Profile	Empirical Gap Addressed
Cognitive Resource Allocation	Gile's Effort Models ($SI = L + M + P + C$) (Gile, 2016)	Attentional distribution across listening, memory, production, and coordination	Non-automated parsing; early cognitive saturation; coordination breakdowns under speed stress	Automated syntax; protected cognitive reserve; adaptive allocation under elevated loads	Lack of empirical data measuring cognitive saturation during Arabic-to-English conference speeds
Working Memory Management	Baddeley's Multicomponent WM; Sweller's Cognitive Load Theory (Uçar, 2025)	Central executive control; phonological loop capacity; extraneous load management	High intrinsic load from VSO/SVO divergence; ear-voice span overextensio	Efficient syntactic chunking; calibrated ear-voice span (2.5–4.0s); high	Scarcity of controlled experiments testing WM decay trajectories across 25-

			n; phonological buffer decay	executive attentional control	minute intervals
Acoustic Delivery & Fluency	Phonation acoustics and temporal speech models (Wenda & Nur Salam, 2025)	Silent pauses (>3s vs. <3s); filled hesitations; phonation-time ratio; speech tempo stability	Temporal degradation; hesitation chains (>15s); severe fluency drops in second half of sessions	Prosodic continuity; self-repairs under 3s; consistent phonation ratio across working blocks	Absence of objective acoustic metrics during authentic LEAP/FII conference simulations
Institutional Lexical Retrieval	Schema Theory & Deliberate Practice (Ericsson, 2006; Aal-Hajiahmed, 2023)	Percentage fidelity on 30 validated Saudi Vision 2030 institutional/financial terms	Literal calques; phonetic distortion; omission of specialized acronyms (e.g., PIF, NTP, Nitaqat)	Direct automated retrieval of pre-compiled terminology equivalents; 92%+ fidelity retention	Lack of validated terminology assessments reflecting Saudi socio-economic discourse
Technological & Pedagogical Interface	AI-Augmented Interpreting; AIIC Professional Standards (Alharbi, 2025)	Screen-based CAI usage; booth simulation exposure (28h domestic vs. 400h international standard)	Secondary cognitive overload from CAI interfaces; poor error recovery on AI caption failures	Selective CAI consultation; rapid detection and correction of automated engine inaccuracies	Absence of experimental designs connecting domestic training shortfalls to mega-event failure rates

Note. Theoretical synthesis and empirical variables adapted from: Gile (2016); Timarova et al. (2014); Al Zahran (2021); Aal-Hajiahmed (2023); Kööp (2019); Setton and Dawrant (2016); Alharbi (2025); Dong and Chen (2025); Uçar (2025); and Wenda and Nur Salam (2025).

By evaluating experienced and novice practitioners within a controlled simulation of LEAP and FII plenaries, this research examines the empirical associations among accumulated deliberate practice, coping strategy deployment, fluency stability, and terminology precision in conference interpreting.

3. Methodology

3.1 Research Design

This study employs a 2 × 2 mixed-design quasi-experimental comparative framework to examine the relationship between professional experience, sustained temporal workload, and simultaneous conference interpreting performance (Mellinger & Hanson, 2017). Specifically, because professional experience is a pre-existing participant characteristic rather than a randomly assigned treatment, the between-subjects factor (Experienced vs. Novice) is quasi-experimental, whereas the within-subjects factor (Interval T1 vs. Interval T2) is experimentally controlled within a standardized laboratory simulation. The between-subjects independent variable is Professional Experience, stratified into two distinct cohorts: Experienced Interpreters and Novice Interpreters. The within-subjects independent variable is the Time/Fatigue Interval, divided into two consecutive observation blocks across a 25-minute continuous interpreting task: Interval 1 (T1: 0.00 to 12.50 minutes) representing initial baseline cognitive engagement, and Interval 2 (T2: 12.51 to 25.00 minutes) capturing the onset of cognitive fatigue and attentional strain. Although professional experience is an organismic (quasi-independent) grouping attribute rather than a randomly assigned treatment, this investigation is framed as an experimental study examining the "impact" of experience because it actively manipulates sustained cognitive workload across controlled temporal intervals (T1 vs. T2) under laboratory-standardized acoustic conditions (140 wpm delivery and 45 dB SPL calibrated background noise), utilising discourse and terminology directly drawn from flagship Saudi Vision 2030 events (LEAP and FII) (Mellinger & Hanson, 2017).

The primary dependent variables comprise four performance dimensions evaluated using a standardized 100-point rubric: Fluency Maintenance (40 points), Terminology Fidelity (30 points), Comprehension Equivalence (20 points), and Delivery Naturalness and Register (10 points). Objective acoustic metrics were extracted concurrently to triangulate subjective evaluator assessments, isolating pause density, hesitation chains, and phonation-time ratios across both temporal blocks.

3.2 Participants and Sampling Matrix

A total of $N = 60$ professional Arabic-English conference interpreters participated in the study, split equally into an Experienced group ($n = 30$) and a Novice group ($n = 30$).

Experienced practitioners were recruited through stratified sampling via the Saudi Conventions and Exhibitions General Authority official rosters, the Saudi Interpreters Association registry, and commercial event agency networks operating in Riyadh. Eligibility criteria for the experienced cohort required:

- A minimum of three years of continuous, verifiable conference interpreting practice;
- Documented completion of at least 1,500 active booth hours supporting flagship Vision 2030 gatherings (e.g., LEAP, Future Investment Initiative, Hajj coordination summits, or Global Cybersecurity Forum);
- Native proficiency in Arabic (Language A) and certified professional working competence in English (Language B).

Novice participants were recruited from recent translation and interpreting graduates across three major Saudi universities: King Saud University, Princess Nourah bint Abdulrahman University, and Imam Mohammad Ibn Saud Islamic University. Inclusion criteria for the novice cohort required:

- Completion of an undergraduate or postgraduate degree in translation/linguistics within the preceding 12 months;
- Less than one calendar year of post-graduate professional work experience;
- Accumulation of fewer than 50 total hours of professional simultaneous conference booth exposure;
- The same linguistic profile: native Arabic (Language A) and advanced academic English (Language B).

All prospective participants underwent pre-screening involving self-reported language background questionnaires, formal qualification checks, and booth-log verifications to confirm adherence to the selection criteria. Detailed demographic and background characteristics of both cohorts are summarized in Table 3. Age, along with gender and educational attainment, was systematically tracked and evaluated as a baseline demographic variable (as detailed in Table 3) and subsequently entered as a control variable in Step 1 of the hierarchical regression analysis (Table 8) to isolate the specific effect of professional experience from natural chronological differences (Ericsson, 2006).

Table 3. Demographic and Professional Profile of Participants ($N = 60$)

Attribute / Parameter	Experienced Cohort (n=30)	Novice Cohort (n=30)	Test Statistic (t / $\chi 2$)	p-value
Age (Years; Mean ± SD)	35.4 ± 4.2	23.8 ± 1.4	t = 14.18	< .001

Gender Distribution (Male / Female)	16 / 14	13 / 17	$\chi 2 = 0.60$	.438
Mean Professional Experience (Years)	5.8 ± 2.1	0.4 ± 0.3	$t = 13.92$	< .001
Accumulated Booth Hours (Mean ± SD)	2,140 ± 480	24 ± 12	$t = 24.15$	< .001
Language Direction Profile	Arabic A / English B (100%)	Arabic A / English B (100%)	,	,
Highest Degree: Bachelor's Degree	18 (60.0%)	26 (86.7%)	$\chi 2 = 5.45$	.020
Highest Degree: Master's Degree	12 (40.0%)	4 (13.3%)	$\chi 2 = 5.45$	.020
Specialized Booth Training (>100 Hours)	30 (100%)	0 (0%)	$\chi 2 = 60.00$	< .001

Note. Data compiled from pre-experiment screening inventories and verified against agency logs.

As reflected in Table 3, because the two cohorts were recruited from intact professional and graduate populations rather than randomly assigned, they differed significantly not only in years of experience and accumulated booth hours, but also in chronological age ($p < .001$), Master's-level education ($p = .020$), and completion of >100 hours of specialized booth training (100% vs. 0%, $p < .001$). Additionally, while no formal pre-test of economic or technological domain knowledge was administered prior to the simulation, the experienced cohort's mandatory recruitment criterion of prior service in Vision 2030 conferences inherently conferred higher baseline familiarity with sovereign finance and technological discourse. Consequently, while chronological age and educational attainment are statistically controlled in Step 1 of the hierarchical regression model (Table 8), "professional experience" in this study should be interpreted as a composite ecological expertise profile, encompassing accumulated booth practice, specialized training, and prior domain exposure, rather than a single isolated variable.

Ethical Considerations and IRB Compliance

This experimental protocol was reviewed and institutionally approved as ethical for the College of Languages and Translation (IRB Protocol Approval No. CLT-IRB-2025-084). All participants were provided with a study information sheet explaining the purpose of the study, the conditions of the simulation, anonymisation safeguards, and data storage procedures. Informed consent was obtained in writing before testing. The subjects were told that they could always quit the study without having to deal with the consequences in the institution or their jobs. Every audio recording, the evaluation forms, and the transcript of the interviews were provided with an alphanumeric identifier (e.g., EXP-01, EXP-30, NOV-01, NOV-30) and kept on encrypted cloud servers, which were only accessible to the research team.

3.3 Experimental Setting, Stimuli, and Acoustic Simulation

Physical Setup and Booth Specifications

The Advanced Simultaneous Interpreting Laboratory at King Saud University served as the venue for all the experimental sessions. There are 3 soundproof booths that meet ISO 2603 (Fixed booths) and ISO 4043 (Mobile/Built-in acoustical insulation standards). A digital interpreting console, a broad-spectrum condenser microphone, and dual cup binaural closed-back headphones with passive noise attenuation were set up in each booth. Dual high-fidelity digital multi-track audio recorders recorded the participants' output in lossless WAV format at 48 kHz, 24-bit, to be later acoustically parsed and blind rater-reviewed.

Source Speech Development and Lexical Calibration

The experimental source material consisted of a continuous, scripted 25-minute Arabic address comprising 3,500 words, calibrated to maintain a steady presentation rate of 140 words per minute (wpm). The content replicated the register, syntactic structures, and institutional density of live proceedings at LEAP and the Future Investment Initiative.

The speech was divided symmetrically into two thematic modules:

- **Module 1 (0.00 to 12.50 minutes; LEAP Simulation):** Centered on technological transformation, generative AI governance, NEOM digital infrastructure, smart logistics networks, and enterprise cloud cybersecurity protocols.
- **Module 2 (12.51 to 25.00 minutes; FII Simulation):** Focused on Public Investment Fund (PIF) sovereign asset deployments, foreign direct investment regulations, Vision Realization Programs (VRPs), and national labor localization frameworks (*Nitaqat*).

To standardize delivery cadence and acoustic uniformity, the speech was recorded in a studio environment by a native Modern Standard Arabic news broadcaster. The recording maintained consistent

prosodic stress, clear articulation, and zero unscripted pauses. To simulate live conference acoustics, calibrated ambient background noise (subtle audience presence, distant murmurs, and room reverberation) was mixed into the headphone feed at 45 dB SPL, providing high ecological validity without obscuring speech intelligibility.

Pilot Testing and Stimulus Calibration

Prior to full experimental deployment, a pilot trial was conducted with **N = 10** conference interpreters (5 experienced, 5 novice) who were not included in the main experimental sample. The pilot study evaluated whether:

- The 140 wpm delivery cadence induced measurable cognitive load without causing immediate task abandonment;
- The 30 pre-selected institutional terms were recognizable yet technically challenging;
- The acoustic monitoring equipment operated reliably without latency artifacts.

Data from the pilot confirmed that the 25-minute duration provoked noticeable cognitive fatigue around the 12-to-15-minute mark while preserving task completion rates. Minor adjustments to three terminology items were implemented based on pilot rater feedback to ensure uniform difficulty. Furthermore, to minimize potential thematic confounding between Interval 1 (Module 1: LEAP) and Interval 2 (Module 2: FII), both modules were strictly calibrated during the pilot phase for objective psycholinguistic equivalence, maintaining an identical word count (1,750 words per module), speech rate (140 wpm), mean clause length, Verb-Subject-Object (VSO) syntactic density, and specialized terminology density (15 target items per 12.5-minute block)

3.4 Evaluation Rubric and Operationalization of Variables

Performance was assessed using a 100-point rubric adapted from the European Masters in Conference Interpreting (EMCI) and the International Association of Conference Interpreters (AIIC) standards, customized to reflect the institutional context of Saudi mega-events. Table 4 details the criteria, point distributions, and operational indicators across each scoring dimension.

Table 4. Evaluation Rubric for Arabic-to-English Simultaneous Interpreting Performance

Evaluation Domain	Maximum Points	Operational Indicators and Scoring Thresholds
Simultaneous Fluency Maintenance	40 Points	**Unfilled Pause Ratio (15 pts):** Zero silent pauses >3. seconds; minimal silent pauses between 1.0—2.9 seconds. Severe penalties for pauses >4.0 seconds. **Hesitation Density (10 pts):** Low occurrence of vocalized fillers ("uh", "um", clicking), false starts, and trailing utterances per minute.

		Phonation-Time Ratio (10 pts): Maintenance of continuous speech flow (>65% active voicing relative to silence). **Speech Tempo Stability (5 pts):** Consistent ear-voice span (EVS) without abrupt bursts or sudden decelerations.
Terminology Fidelity Reproduction	30 Points	**Accuracy on 30 Target Terms (30 pts; 1 pt each):** Binary scoring (1 = accurate contextual rendering; 0 = omission, phonological distortion, or misleading calque) across pre-selected Vision 2030 institutional items (e.g., PIF sovereign deployments, NTP milestones, Saudization quotas, Hijri fiscal alignments).
Comprehension & Semantic Equivalence	20 Points	**Propositional Integrity (10 pts):** Accurate translation of core thematic arguments, causal linkages, and statistical metrics. **Information Completeness (10 pts):** Absence of unjustified message truncation, arbitrary omissions, or hallucinated content.
Delivery Quality, Register & Prosody	10 Points	**Prosodic Contour & Intonation (5 pts):** Natural English declarative and interrogative cadences; appropriate sentence boundary management. **Conference Register (5 pts):** Confident, professional vocal timbre; absence of audible panic, hyperventilation, or microphone distress.

Note. Adapted from validated EMCI scoring guidelines and calibrated for the Saudi conference landscape.

Adjudication and Inter-Rater Reliability

Audio recordings were randomized, assigned blind codes, and evaluated independently by a panel of five certified senior conference interpreters, each possessing over ten years of international booth experience and active AIIC membership. Evaluators received a two-hour calibration briefing along with an evaluation key containing verified definitions and acceptable target equivalents for the 30 target terminology items. Evaluators were blind to participant identities, years of experience, and study hypotheses. Inter-rater reliability was calculated across all 60 evaluations using Cronbach's alpha (α) for internal consistency and Kendall's coefficient of concordance (W) for scoring concordance across the five raters. Reliability analyses across all 60 evaluations confirmed high inter-rater concordance across all dimensions (as formally detailed in Section 4.1):

- Total Rubric Score: $\alpha = .938$, Kendall's $W = .896$ ($p < .001$);
- Fluency Maintenance: $\alpha = .924$, Kendall's $W = .882$ ($p < .001$);

- Terminology Fidelity: $\alpha = .958$, Kendall's $W = .929$ ($p < .001$);
- Comprehension Equivalence: $\alpha = .915$, Kendall's $W = .874$ ($p < .001$);
- Delivery Quality: $\alpha = .900$, Kendall's $W = .868$ ($p < .001$).

Individual evaluator scores were averaged to generate the final composite metrics for statistical testing.

To ensure full methodological transparency and explicitly distinguish instrument validity (whether the rubric accurately measures the intended interpreting constructs) from reliability (inter-rater scoring consistency), the following operational and validation protocols were enforced:

1. Separation of Subjective Rubric Scores and Objective Praat Metrics (Preventing Double-Counting): Evaluators' fluency scores (maximum 40 points) and Praat acoustic measurements (silent pauses >3.0s, filled hesitations, and phonation-time ratios) were analyzed as parallel, independent datasets for methodological triangulation rather than summed into a single metric. In Table 5, the repeated-measures ANOVA on the 40-point rubric fluency score was conducted separately from the independent *t*-tests on the Praat acoustic parameters, ensuring that disfluency features were never counted twice in any composite score.

2. Psycholinguistic Justification of Acoustic Thresholds: The > 3.0-second penalty threshold for unfilled pauses and the > 65% phonation-time benchmark were derived from established conference interpreting acoustics (Tissi, 2000; Pradas Macías, 2009; Kööp, 2019), which demonstrate that silent pauses exceeding 3.0 seconds exceed natural prosodic phrasing and signal cognitive overload to listeners, while continuous simultaneous speech requires active voicing across approximately two-thirds (65%–75%) of total turn duration.

3. Terminology Scoring Key and Contextual Equivalence Rules: For the 30 target terms (10 items each across Sovereign Finance, Saudization Governance, and Deep-Tech, as categorized in Table 6), raters applied a standardized scoring rubric during the calibration briefing. A score of 1 point was awarded not only for verbatim matches to the primary reference term (e.g., Public Investment Fund, Nitaqat labor localization tiers, digital twin architecture), but also for functionally equivalent institutional paraphrases that preserved complete semantic, legal, and economic precision in context (e.g., rendering Nitaqat as "Saudi national employment quota bands"). A score of 0 was assigned for omissions, phonetic distortions, or misleading literal calques (e.g., translating Nitaqat literally as "zones" or "borders").

4. Instrument Content Validity vs. Inter-Rater Reliability: Prior to reliability testing, content and construct validity were established through a two-round modified Delphi review by three independent university interpreting professors and two senior AIIC practitioners, achieving a scale-level Content Validity Index (S-CVI/Ave) of 0.94 after refining three terminology items during the N = 10 pilot trial. Regarding evaluator qualifications, all five adjudicators held postgraduate degrees in Conference Interpreting, held active AIIC

accreditation, had over 10 years of Arabic–English booth experience, and had prior experience serving on institutional accreditation and university examination panels. During the two-hour training workshop, adjudicators independently scored three pilot recordings and resolved scoring discrepancies against the anchored rubric descriptors before evaluating the main cohort.

5. Score Aggregation and Interval Reliability Structure: For each participant, raw scores across the five adjudicators were combined by computing the arithmetic mean ($\sum_{r=1}^{5} X_{ir} / 5$) for each sub-scale at T1 and T2. Cronbach's α and Kendall's W were calculated across the 60 x 5 participant-by-rater matrices for the pooled task and verified separately across both temporal intervals (T1 composite α = .941, W = .908; T2 composite α = .935, W = .884; pooled composite α = .938, W = .896, p < .001), confirming consistent rater stability across both phases of the simulation.

3.5 Data Extraction and Statistical Framework

Acoustic Extraction Procedures

Objective acoustic parameters were extracted using Praat (Version 6.4.12) phonetics software. Audio tracks were batch-processed through an automated silence-detection script calibrated to detect silent pauses using an intensity floor of -25 dB relative to peak intensity, with a minimum duration threshold of 250 milliseconds. The software extracted:

- Total number and mean duration of unfilled pauses under 3.0 seconds;
- Frequency of disruptive silent pauses exceeding 3.0 seconds;
- Phonation-time ratio, defined as the cumulative voicing duration divided by total block duration (12.5 minutes} = 750 seconds);
- Mean length of uninterrupted runs between pause markers.

Qualitative Strategy Coding

After the simulation experiment, a second experiment involved conducting semi-structured retrospective debriefing interviews with a purposefully selected sub-sample of 15 participants (8 experienced, 7 novice). Interviews included: self-reported cognitive load, ear-voice span regulation, coping strategies faced in the lexical challenges and conscious recovery strategies. NVivo (Release 14) was used to import transcripts and code them using thematic analysis. Themes found emerged under three main cognitive coping strategies: Prospective Lexical Anticipation, Strategic Syntactic Chunking/Condensation and Controlled Omission. To operationalize these three qualitative themes as continuous quantitative predictors across the entire experimental cohort (N = 60) for hierarchical regression modeling, a two-phase mixed-methods coding protocol was implemented in NVivo (Release 14). In Phase 1, the retrospective debriefing transcripts from the purposeful sub-sample (n = 15; 8 experienced, 7 novice) were analyzed inductively to establish the qualitative

taxonomy and confirm deliberate metacognitive intent. In Phase 2, these validated coding rules were applied deductively to the verbatim time-aligned target speech transcripts of all $N = 60$ participants (aligned against the Arabic source text) to extract objective behavioral frequency indices for every participant: (1) Prospective Monitoring Index, quantified as the frequency of anticipatory SVO syntactic restructurings initiated prior to the completion of source Arabic clause-final predicates; (2) Syntactic Chunking Frequency, measured as the count of complex coordinated Arabic clauses decoupled into independent English propositions per interval; and (3) Tactical Omission Ratio, calculated as the proportion of non-essential rhetorical redundancies omitted without loss of core propositional meaning relative to total omissions.

The semi-structured debriefing protocol comprised four standardized open-ended prompts investigating: (1) perceived moments of peak mental strain across T1 and T2; (2) deliberate techniques used to manage Arabic verb-initial (VSO) and paratactic sentence structures; (3) decision-making rules when encountering unfamiliar Vision 2030 institutional terms; and (4) conscious recovery mechanisms following hesitation chains. To ensure strict methodological independence and prevent criterion contamination, the qualitative and corpus coding was conducted by two independent bilingual applied linguists who were completely separate from the five AIIC performance evaluators and entirely blinded to the participants' group membership, demographic profiles, and 100-point rubric scores. Both coders independently coded 25% of the datasets ($n = 4$ interview transcripts and n = 15 full-task output transcripts), achieving high inter-coder agreement (Cohen's $\kappa = .87$ for thematic interview coding; Intraclass Correlation Coefficient ICC = .91 for continuous transcript frequency counts), with minor boundary disagreements resolved through consensus adjudication prior to coding the remaining corpus. Finally, to transform the coded transcript behaviors into standardized continuous predictors for Model 3 of the regression ($N = 60$), raw occurrence counts across the 25-minute task were normalized into continuous indices: Prospective Monitoring Index (successful anticipatory SVO initiations per 100 source clauses), Syntactic Chunking Frequency (number of segmented paratactic structures per 10-minute block), and Tactical Omission Ratio (proportion of rhetorical condensations preserving core propositions relative to total omissions, scaled from 0.00 to 1.00).

Statistical Pipeline

Quantitative analyses were performed using SPSS (Version 29.0) alongside R (Version 4.3.2). The analytical pipeline proceeded through four stages:

1. **Preliminary Diagnostics:** Verification of normality via Shapiro-Wilk tests, examination of kurtosis and skewness, and assessment of homogeneity of variance via Levene's test.

2. **Two-Way Repeated Measures ANOVA (2 X 2):** Conducted on composite rubric scores and individual sub-scales to assess the main effect of Professional Experience (Between-Subjects: Experienced vs. Novice), the main effect of Time Interval (Within-Subjects: T1 vs. T2), and the *Experience X Time* interaction effect. Effect sizes were evaluated via partial eta squared ($\eta p2$).
3. **Independent and Paired t-Tests:** Deployed to identify between-group differences across specific terminology domains and within-group degradation profiles from T1 to T2. Effect sizes were measured using Cohen's *d*.
4. **Hierarchical Multiple Regression:** Modeled performance sustainability, examining whether accumulated booth hours, working memory proxies, and NVivo-coded cognitive strategy deployment predicted total interpreting scores across high-fatigue intervals. Statistical significance across all operations was established at an alpha level of $\alpha = .05$.

Reproducibility and Data Availability Statement: In accordance with open-science standards and IRB confidentiality stipulations (Protocol No. CLT-IRB-2025-084), the anonymized participant-level scoring dataset (N = 60), the Praat silence-detection script (configured at -25 dB silence threshold, 250 ms minimum pause floor, and 100 ms minimum sounding interval), the R (afex and lm packages) and SPSS (Version 29.0) syntax files, and the 30-item terminology scoring key are archived and available from the corresponding author upon reasonable academic request.

4. Results and Data Analysis

4.1 Preliminary Diagnostics and Inter-Rater Reliability

Prior to conducting inferential statistical tests, all scoring datasets and acoustic parameters were screened for statistical assumptions. Normality was evaluated using the Shapiro-Wilk test across both participant cohorts and all rubric sub-scales. For the experienced group, composite scores exhibited normal distribution in both Interval 1 ($W = 0.968$, $p = .482$) and Interval 2 ($W = 0.961$, $p = .325$). Similarly, the novice cohort met normality criteria in Interval 1 ($W = 0.954$, $p = .216$) and Interval 2 ($W = 0.948$, $p = .151$).

Homogeneity of variance was evaluated using Levene's test of equality of error variances. Rubric composite scores satisfied variance homogeneity across groups in Interval 1 ($F(1, 58) = 1.84$, $p = .180$) and Interval 2 ($F(1, 58) = 2.47$, $p = .121$). Sphericity assumptions were inherently satisfied because the within-subjects factor (Time) comprised exactly two measurement intervals (T1 and T2).

Evaluation ratings across the five certified senior conference interpreters revealed high inter-rater agreement across all 60 cases. The overall composite scoring showed an internal consistency coefficient of Cronbach's Alpha = .938 and an overall concordance coefficient of Kendall's $W = .896$ ($p < .001$). Sub-dimension reliabilities were equally strong:

- Fluency Maintenance: $\alpha = .924$, Kendall's $W = .882$ ($p < .001$)
- Terminology Fidelity: $\alpha = .958$, Kendall's $W = .929$($p < .001$)
- Comprehension Equivalence: $\alpha = .915$, Kendall's $W = .874$ ($p < .001$)
- Delivery Quality: $\alpha = .90$, Kendall's $W = .868$ ($p < .001$)

Averaged evaluator scores were compiled for subsequent quantitative modeling.

4.2 Primary Quantitative Findings

4.2.1 Simultaneous Interpreting Fluency and Temporal Decay

To address Research Question 1, delivery fluency was assessed across both subjective rubric criteria (maximum 40 points) and objective acoustic measurements extracted through Praat phonetics software. Performance was tracked across Interval 1 (T1: 0.00–12.50 minutes) and Interval 2 (T2: 12.51–25.00 minutes) to evaluate the impact of sustained cognitive pressure. Table 5 presents the descriptive statistics and group comparisons for the fluency parameters.

Table 5. Descriptive Statistics and Acoustic Delivery Metrics for Experienced vs. Novice Interpreters Across Time Intervals ($N = 60$)

Metric / Dependent Measure	Interval	Experienced Cohort (n=30)	Novice Cohort (n=30)	Mean Difference (Δ)	Independent t (df=58)	p-value
Fluency Score (Max: 40)	T1 (0–12.5 min)	36.42 ± 2.15	26.85 ± 3.42	+9.57	12.98	< .001
	T2 (12.5–25 min)	35.18 ± 2.48	15.68 ± 3.86	+19.50	23.21	< .001
Silent Pauses >3.0s (per min)	T1 (0–12.5 min)	0.28 ± 0.18	1.42 ± 0.52	-1.14	11.24	< .001
	T2 (12.5–25 min)	0.41 ± 0.22	3.86 ± 0.84	-3.45	21.56	< .001
Filled Hesitations (per min)	T1 (0–12.5 min)	1.15 ± 0.44	3.64 ± 0.92	-2.49	13.25	< .001
	T2 (12.5–25 min)	1.38 ± 0.51	6.92 ± 1.45	-5.54	19.78	< .001

Phonation-Time Ratio (%)	T1 (0–12.5 min)	71.4% ± 4.2%	58.2% ± 5.8%	+13.2%	10.08	< .001
	T2 (12.5–25 min)	69.8% ± 4.6%	42.5% ± 6.4%	+27.3%	19.01	< .001

Note. T1 represents the initial module (LEAP simulation); T2 represents the sustained fatigue module (FII simulation). Values denote Mean ± Standard Deviation. Across all comparisons (df = 58, two-tailed), standardized effect sizes (Cohen's *d*) and 95% confidence intervals of the mean difference (CI**Δ**) are: Fluency T1 (*d* = 3.35, 95% CI [8.09, 11.05]), Fluency T2 (*d* = 6.01, 95% CI [17.82, 21.18]); Silent Pauses >3.0s T1 (*d* = -2.93, 95% CI [-1.34, -0.94]), T2 (*d* = -5.62, 95% CI [-3.77, -3.13]); Filled Hesitations T1 (*d* = -3.45, 95% CI [-2.87, -2.11]), T2 (*d* = -5.09, 95% CI [-6.10, -4.98]); Phonation Ratio T1 (*d* = 2.61, 95% CI [10.58%, 15.82%]), T2 (*d* = 4.90, 95% CI [24.42%, 30.18%]).

A two-way repeated-measures ANOVA was performed on the fluency rubric scores, testing the between-subjects effect of Professional Experience (Experienced vs. Novice), the within-subjects effect of Time Interval (T1 vs. T2), and their interaction:

- **Main Effect of Experience:** Statistically significant and large, $F(1, 58) = 492.14$, $p < .001$, $\eta^2_p = .894$. Experienced interpreters achieved systematically higher fluency scores across both intervals ($M = 35.80$, $SE = 0.42$) compared to novice interpreters ($M = 21.27$, $SE = 0.42$).
- **Main Effect of Time Interval:** Statistically significant, $F(1, 58) = 196.42$, $p < .001$, **ηp2** = .772, reflecting general performance degradation from T1 ($M = 31.64$, $SE = 0.37$) to T2 ($M = 25.43$, SE = 0.42).
- **Interaction Effect (Experience X Time):** Highly significant, $F(1, 58) = 118.76$, $p < .001$, **ηp2** = .672.

Post-hoc pairwise comparisons with Bonferroni adjustments indicated that while experienced interpreters exhibited a slight, non-critical decline in fluency scores of 1.24 points (3.4%) from T1 to T2 ($p = .021$), novice interpreters experienced a drop of 11.17 points (41.6%) between intervals ($p < .001$).

Acoustic data mirrored this interaction. For novice interpreters, the frequency of disruptive silent pauses exceeding 3.0 seconds rose from 1.42 pauses/min in T1 to 3.86 pauses/mi in T2, accompanied by a drop in active phonation-time ratio from 58.2% down to 42.5%. In contrast, experienced interpreters maintained active voicing above 69% throughout the task, limiting silent intervals to brief syntactic transitions.

4.2.2 Terminology Fidelity Across Lexical Domains

To evaluate Research Question 2, accuracy across 30 pre-selected Saudi Vision 2030 institutional and technical terminology items was assessed. The test items were divided into three 10-item domains:

1. **Sovereign Finance & PIF Architecture:** E.g., Public Investment Fund Tier-1 capital deployments, sovereign wealth allocations, debt-to-GDP fiscal alignments, Vision Realization Program targets.
2. **Saudization & Labor Market Governance:** E.g., *Nitaqat* premium quota tiers, *Ajeer* temporary contract portal, human capital localization coefficients, balanced Saudization indices.
3. **Deep-Tech & Infrastructure Development:** E.g., NEOM digital twin architectures, THE LINE zero-carbon modular logistics, OXAGON cognitive industrial port protocols, giga-project supply chain transparency.

Table 6 details the accuracy rates and error distributions across both cohorts.

Table 6. Terminology Fidelity Scores and Error Patterns Across Specialized Lexical Domains ($N = 60$)

Specialized Lexical Domain (10 Terms Each)	Experienced Accuracy (%)	Novice Accuracy (%)	Independent t (df=58)	p-value	Cohen's d	Dominant Novice Error Pattern
1. Sovereign Finance & PIF (10 Pts)	94.3% ± 5.2%	44.7% ± 9.8%	24.42	< .001	6.30	Calques; omission of financial acronyms
2. Saudization & Labor Quotas (10 Pts)	91.7% ± 6.4%	38.3% ± 11.2%	22.71	< .001	5.86	Literal word-for-word translation; distortion
3. Deep-Tech & NEOM Projects (10 Pts)	89.3% ± 7.1%	49.0% ± 10.4%	17.48	< .001	4.51	Message truncation; semantic vagueness
Total Terminology Fidelity (30 Pts)	**27.53 ± 1.48**	**13.20 ± 2.45**	**27.32**	**< .001**	**7.05**	**Pervasive lexical abandonment**

Note. Correct renderings earned 1 point; omitted, phonetically garbled, or semantically distorted terms received 0 points.

As detailed in Table 6, experienced interpreters demonstrated high terminology reproduction across all domains, averaging an overall fidelity rate of 91.8% (27.53 out of 30 points). Novices, conversely, achieved an overall accuracy of only 44.0% (13.20 out of 30 points).

Error classification revealed distinct failure mechanisms. In Domain 2 (Saudization & Labor Governance), novices struggled with culturally and institutionally embedded jargon. For example, the term *Nitaqat* was frequently translated as "scopes" or "borders" rather than "Saudization employment tiers" or "labor localization quota system." In Domain 1 (Sovereign Finance), novice interpreters omitted acronyms such as PIF or parsed them as literal descriptive phrases (e.g., "the public money group"), leading to semantic erosion. In Domain 3 (Deep-Tech), novices frequently engaged in strategic message abandonment when complex technological terms co-occurred with rapid source delivery.

4.2.3 Overall Composite Performance

Composite scores were aggregated across all four rubric dimensions (Fluency: 40 points; Terminology: 30 points; Comprehension: 20 points; Delivery: 10 points). Table 7 presents the cumulative performance comparison.

Table 7. Composite Performance Comparison on the Standardized 100-Point Rubric ($N = 60$)

Rubric Performance Dimension	Maximum Points	Experienced Cohort (n=30)	Novice Cohort (n=30)	Difference (Δ)	Independent t (df=58)
Simultaneous Fluency Maintenance	40 Points	35.80 ± 2.18	21.27 ± 3.12	+14.53	20.91***
Terminology Fidelity Reproduction	30 Points	27.53 ± 1.48	13.20 ± 2.45	+14.33	27.32***
Comprehension & Semantic Integrity	20 Points	17.65 ± 1.22	11.42 ± 1.84	+6.23	15.42***

Delivery Quality, Register & Prosody	10 Points	8.72 ± 0.74	5.35 ± 1.15	+3.37	13.52***
Total Composite Score	**100 Points**	**89.70 ± 4.12**	**51.24 ± 6.84**	**+38.46**	**26.34*** **

Note. ***p < .001. Standard deviations shown alongside means.

For all independent t-tests (df = 58, two-tailed, p < .001), standardized effect sizes (Cohen's d) and 95% confidence intervals of the mean difference (CI**Δ**) are as follows: Fluency (d = 5.40, 95% CI [13.14, 15.92]); Terminology (d = 7.05, 95% CI [13.28, 15.38]); Comprehension (d = 3.98, 95% CI [5.42, 7.04]); Delivery (d = 3.49, 95% CI [2.87, 3.87]); and Total Composite Score (d = 6.81, 95% CI [35.54, 41.38]).

Experienced interpreters maintained high composite scores (M = 89.70, SD = 4.12), reflecting standard international conference expectations. Novice practitioners exhibited lower overall scores (M = 51.24, SD = 6.84), driven by compounded breakdowns across fluency continuity and terminology accuracy under sustained cognitive load. From a methodological perspective, the exceptionally large between-group effect sizes observed across Tables 5, 6, and 7 (e.g., **ηp2** = .894 for fluency and Cohen's d = 7.05 for terminology fidelity) are a direct mathematical consequence of three structural factors inherent to the experimental design: (1) an extreme-groups sampling contrast comparing active AIIC-standard summit veterans (2,140 ± 480 booth hours) against recent university graduates (24 \pm 12 booth hours) who had never undergone intensive simultaneous training; (2) high criterion-referenced ecological alignment, where experienced participants had previously worked inside live LEAP and FII booths and already possessed consolidated mental schemas for the exact 30 institutional terms tested, creating a near-ceiling clustering with very small within-group standard deviations (SD = 1.48); and (3) score aggregation across five calibrated raters (α = .938), which attenuates random measurement error and compresses within-group error variance (σ_{pooled}), thereby amplifying standardized mean differences (d and **ηp2**) relative to naturalistic field studies with heterogeneous intermediate samples.

4.3 Regression Modeling and Qualitative Strategy Integration

4.3.1 Hierarchical Multiple Regression Analysis

A three-stage hierarchical multiple regression was conducted to determine the primary predictors of sustained performance during the high-fatigue interval (T2 composite scores). In Step 1, chronological age and

formal educational attainment were entered as control variables. In Step 2, Accumulated Professional Booth Hours were added. In Step 3, Cognitive Strategy Deployment, quantified in NVivo across all N= 60 time-aligned target interpretation transcripts using the coding taxonomy established during the n = 15 retrospective debriefings (comprising the Prospective Monitoring Index, Syntactic Chunking Frequency, and Tactical Omission Ratio), was entered. Table 8 summarizes the model parameters.

Table 8. Hierarchical Multiple Regression Models Predicting Sustained Interpreting Performance at Interval T2

Model Stage / Predictor Variable	**B**	**SE B**	**β**	**t-statistic**	**p-value**	**R^2**	**ΔR^2**	**ΔF**
Model 1: Baseline Controls						.082	.082	2.55
Age (Years)	0.34	0.28	.162	1.21	.231			
Formal Education Level	2.14	1.85	.154	1.16	.252			
Model 2: Experience Metric						.684	.602	106.72***
Accumulated Booth Hours	0.016	0.002	.785	10.33	< .001			
Model 3: Cognitive Strategies						**.812**	**.128**	**12.03*** **
Accumulated Booth Hours	0.009	0.002	.441	4.52	< .001			
Prospective Monitoring Index	4.18	0.82	.382	5.10	< .001			

Syntactic Chunking Frequency	3.22	0.74	.295	4.35	< .001
Tactical Omission Ratio	2.84	0.91	.214	3.12	.003

Note. Dependent variable = Composite Rubric Score at T2 (12.5–25.0 minutes). ***p < .001.

The hierarchical regression revealed that baseline demographic variables in Model 1 accounted for only 8.2% of the variance in sustained performance (p = .087). Introducing accumulated booth hours in Model 2 yielded a significant increase in explanatory power (ΔR^2 = .602, p < .001), confirming that deliberate practice is a key driver of performance stability.

The inclusion of cognitive processing strategies in Model 3 accounted for an additional 12.8% of unique variance ($\Delta F(3, 53) = 12.03$, p < .001), bringing the total explained variance to 81.2% (R^2 = .812, Adjusted R^2 = .791). Within the final model, Accumulated Booth Hours (β = .441, p < .001), Prospective Monitoring (β = .382, p < .001), and Syntactic Chunking (β = .295, p < .001) emerged as the strongest individual predictors of an interpreter's ability to withstand cognitive fatigue.

To verify that the high explained variance in Model 3 (R^2 = .812, Adjusted R^2 = .791) was not an artifact of multicollinearity, overfitting, or predictor-group collinearity, comprehensive regression diagnostics were conducted. Variance Inflation Factors (VIF) across all six predictors ranged between 1.14 and 2.86 (with corresponding Tolerance values between 0.35 and 0.88), remaining well below the conservative multicollinearity threshold of VIF < 5.0. Residual independence was confirmed via the Durbin-Watson statistic (DW = 1.94), and Cook's distance values (D_i < 0.18) indicated no undue influential outliers. Furthermore, to guard against small-sample overfitting (N = 60, k = 6, yielding an observation-to-predictor ratio of 10:1), non-parametric bootstrap resampling (5,000 iterations, bias-corrected and accelerated 95% BCa CIs) and leave-one-out cross-validation (LOOCV R^2_{pred} = .768) confirmed high parameter stability. Finally, entering Accumulated Booth Hours as a continuous metric alongside continuous behavioral strategy counts mitigated dichotomous group-membership redundancy while preserving within-cohort variance.

4.3.2 Qualitative Cognitive Strategy Profiles

Thematic analysis of the 15 retrospective debrief interviews conducted via NVivo identified three recurring cognitive coping strategies that systematically differentiated experienced practitioners from their novice counterparts:

1. Prospective Monitoring and Syntactic Anticipation: Experienced interpreters reported actively anticipating the end of standard Arabic verb-noun collocations and institutional formulations. Rather than waiting for complete syntactic closures, experts used contextual probability to forecast administrative arguments, allowing them to initiate English SVO structures earlier and preserve working memory capacity. Novices, by contrast, relied on linear word-by-word tracking, remaining passive until the speaker completed full syntactic clauses. As one experienced practitioner explained: "In FII speeches, once I hear an opening verb like 'tastaḥwidhu' [acquires/allocates] followed by 'Ṣundūq al-Istithmārāt' [PIF], I don't wait for the object at the end of the sentence; I immediately anchor 'The Public Investment Fund' as my English subject and anticipate the allocation metric" (EXP-04). Conversely, a novice noted: "I kept waiting for the speaker to finish the long Arabic phrase so I wouldn't guess wrong, but by then I had forgotten the beginning" (NOV-12).
2. Syntactic Decoupling and Tactical Chunking: When faced with complex Arabic sentence parataxis (e.g., coordinated clauses linked by *wa-*, *fa-*), experienced practitioners deliberately decoupled source units into shorter, independent English propositions. This syntactic segmentation prevented cognitive saturation in the phonological loop. Novices attempted to mirror the long, compound clausal structures of the source speech, resulting in cognitive overload, grammatical entrapment, and mid-sentence stalls. An experienced interpreter illustrated this segmentation strategy: "Arabic speakers chain three or four ideas with 'wa-kadhālika' [and furthermore]; I deliberately slice them into short, standalone English declarative sentences before my short-term memory overflows" (EXP-19). In contrast, a novice reported: "I tried to connect everything with 'and' and 'which' just like the speaker, and I got trapped inside a sentence I couldn't grammatically finish" (NOV-07).
3. Tactical Condensation versus Involuntary Omission: Experienced practitioners deployed deliberate, tactical condensation when speech rates exceeded comfortable processing capacity. They stripped away rhetorical redundancies and honorific phrases while preserving core financial figures and institutional terms. In contrast, novices exhibited involuntary omissions, dropping crucial data points, numerical metrics, and key policy names once their working memory became overwhelmed. Reflecting this distinction, an expert stated: "When the pace hits 140 words per minute, I consciously compress ceremonial titles and repetitive adjectives so I have breathing room to hit the exact Nitaqat percentages and PIF figures" (EXP-22), whereas a novice admitted: "When the financial acronyms piled up in the second half, my mind went blank and I accidentally skipped whole sentences containing the numbers" (NOV-18).

4.4 Synthesis of Key Empirical Findings

In summarizing the primary statistical and qualitative outcomes across the experimental protocol, the study establishes three core empirical benchmarks:

1. *Temporal Fluency Degradation:* Repeated-measures ANOVA confirmed significant main effects for Professional Experience and Elapsed Time, alongside a robust interaction effect ($F(1,58) = 118.76, p < .001, \eta_p^2 = .672$). While experienced practitioners maintained high delivery continuity and phonation-time ratios ($> 69\%$) across both intervals, novice interpreters experienced an acute 41.6% decline in fluency scores during T2, coupled with a nearly threefold increase in disruptive silent pauses exceeding 3.0 seconds (1.42 to 3.86 pauses/min) and an active voicing drop to 42.5%.
2. *Domain-Specific Terminology Reproduction:* Experienced interpreters exhibited high lexical accuracy across specialized Vision 2030 discourse ($M = 91.8\%$, $27.53/30$ points), whereas novices achieved only 44.0% accuracy ($13.20/30$ points), yielding a substantial between-group gap ($t(58) = 27.32, p < .001, d = 7.05$). Novices displayed systematic vulnerability to culturally and administratively embedded labor and financial concepts (scoring 38.3% in Saudization governance and 44.7% in sovereign finance) predominantly due to literal theological calques and acronym omission.
3. *Predictive Strategy Modeling:* Hierarchical multiple regression analysis ($R^2 = .812$,Adjusted $R^2 = .791, p < .001$) established that accumulated booth hours ($\beta = .441, p < .001$) in synergy with observable cognitive restructuring strategies—namely prospective syntactic anticipation ($\beta = .382, p < .001$), syntactic decoupling/chunking ($\beta = .295, p < .001$), and tactical condensation ($\beta = .214, p = .003$)—accounted for over 81% of the variance in sustained interpreting performance under elevated fatigue conditions.

5. Discussion

5.1 Cognitive Resource Allocation and Effort Modeling Dynamics

The results show support for the central tenets of Daniel Gile's Effort Models which describe the differential cognitive allocations that enable simultaneous interpreting in an authentic conference setting. More experienced practitioners maintained balance throughout the entire 25-minute simulation. Senior interpreters spent less mental energy on Listening and Analysis (*L*) and Speech Production (*P*) tasks, allowing them to focus on higher level morphological parsing, syntactic reordering, and routine lexical conversions. This structural automation enabled the maintenance of the necessary reserve capacity in the Coordination Effort (*C*) so that

the interpreter could redirect attention to the source language when the speaker spoke faster or used more acronyms.

In comparison, novice interpreters had textbook symptoms of cognitive saturation. Since recent graduates were not familiar with the structural changes that should be made between Arabic Verb-Subject-Object (VSO) and English Subject-Verb-Object (SVO) structures, they had to devote more resources to Listening and Short-Term Memory (*M*) tasks. The total processing capacity required (TR) was greater than the available mental resources (*TA*) when complex Vision 2030 terminology was introduced in incoming speech.The condition ($TR \leq TA$) was violated when complex Vision 2030 terminology was introduced in incoming speech, with $TR > TA$. As described in Gile's tightrope hypothesis, cognitive saturation caused a series of failure factors: ear-voice span became unsustainable, prosodic delivery became dislocated with too many silent pauses, and high-density propositions were dropped. In interpreting these findings, it is methodologically essential to distinguish between three analytical levels: (1) what was directly measured (objective acoustic parameters via Praat, blind rubric evaluations of target output, and corpus frequency counts of structural shifts); (2) what participants reported (retrospective subjective perceptions of mental strain and deliberate coping choices during the post-task debriefings); and (3) what is theoretically inferred via Gile's Effort Models, Baddeley's working memory architecture, and Cognitive Load Theory. Because this study did not administer independent psychometric working-memory span tests or direct neurophysiological biometric measures (such as pupillometry, EEG, or fNIRS), constructs such as cognitive saturation, working-memory depletion, and processing automaticity are presented as theoretical explanatory mechanisms that account for the observed behavioral degradation rather than directly measured neural states.

5.2 The Terminology Fidelity Gap and Schema Theory

Lexical retrieval in real time is influenced by domain-specific cognitive schemas, as evidenced by the fact that expert interpreters reported 91.8% fidelity in terms of the terminology used, while novice interpreters reported 44.0%. Schema theory assumes that domain-specific knowledge is stored in the Long Term Memory as a network of interconnected schemas so that when an expert is presented with an appropriate context, the schema can be retrieved automatically. In this research, the trained practitioners showed consolidated mental schemas for the specialized lexicon of Saudi governance, such as Nitaqat labor quota system, Vision Realization Programs (VRPs), and Public Investment Fund (PIFs) sovereign portfolios. The retrieval was limited in cognitive effort when they appeared in the source stream, when speech production and monitoring channels were not affected.

These structured lexical representations were not found among novices. Novices were forced to use slow and effortful controlled processing because there were no automatic translation equivalents of local

socioeconomic constructs. This resulted in literal calques, for instance, translating Nitaqat as "ranges" or "zones" rather than "labor localization quota tiers", as well as pervasive lexical omissions, such as completely dropping specialized institutional acronyms like PIF (Public Investment Fund) and VRPs (Vision Realization Programs) under high cognitive load.

Such human performance shortcomings are similar to recent AI benchmarks. The automation of speech-to-text technology is found to have only 62% terminology fidelity for specialized Saudi business-governmental discourse because of dialectal nuances, regional administrative acronyms, and ever-changing institutional jargon, in the words of Alharbi (2025). The result shown here (44.0%) with the novice was not only below the baseline for the fully trained human annotator, but it was also below the baseline accuracy rate for basic automated captioning, leaving one to wonder what the risk of sending untrained graduates into high stakes conference booths is. Neither new interpreters nor automated systems can ensure communicative fidelity in situations of high-level sovereign negotiations without experienced practitioners who can perform metacognitive monitoring at a high level.

5.3 Acoustic Decay, Cognitive Fatigue, and Language Directionality

Findings of 41.6% decrease in fluency scores and a nearly threefold increase in pauses in the novice group that >3.0 seconds in duration and a drop in active phonation-time to 42.5% of the total in the second half of the task offer empirical evidence for cognitive fatigue dynamics. These patterns are consistent with the international studies conducted by Timarova et al., (2014), Uçar (2025) and Dong and Chen (2025), who suggest that working memory depletion is a main factor causing dysfluency in sustained simultaneous processing.

However, because the experimental protocol presented the technology module (T1) and the sovereign finance/labor governance module (T2) in a fixed sequential order without counterbalancing, this second-half decline should not be attributed to temporal cognitive fatigue in isolation. Even though both 12.5-minute intervals were strictly equated during pilot testing for word count (1,750 words), delivery rate (140 wpm), syntactic complexity, and terminology density (15 target items per block), novice interpreters exhibited lower lexical accuracy on culturally embedded labor and financial terms (38.3%–44.7%) than on technology terms (49.0%). Therefore, the steep performance drop in T2 reflects a cumulative interaction between progressive working memory fatigue and the heightened lexical difficulty of institutional governance terminology.

The results also show how the directionality of language contributes to cognitive fatigue. Structural asymmetries in the Arabic-English processing process will continually strain working memory. The use of clausal parataxis, rhetorical ornamentation and verb-initial structure is a crucial element of Arabic conference

oratory, and the interpreter must retain the source elements in the phonological loop as they wait for the main element, which is the semantic predicate.

The long, coordinate sentences were modified by experienced interpreters by chunking and anticipating the words, whereas novices tried to imitate the speaker's structures. At this time, novices' working memory executive control began to fail as cognitive fatigue increased beyond 12 minutes, resulting in longer pause chains and articulatory stalls. It shows that, in divergent language pairs, 25 minutes of untrained booth simultaneous interpreting causes extensive mental fatigue on the interpreter.

6. Practical, Policy, and Economic Implications

6.1 Securing the Vision 2030 Conference Ecosystem

While the empirical findings of this study are derived from a single 25-minute laboratory simulation (N = 60) and do not directly measure macro-economic return on investment, graduate employment outcomes, or the longitudinal impact of specific curriculum hours, they offer exploratory implications for stakeholder consideration when contextualized alongside independent industry reports (Setton & Dawrant, 2016; Al-Suhaim, 2024; NES Fircroft, 2025).

Simultaneous interpreting directly affects the credibility of institutions and the returns on investments, as Saudi Arabia increases its presence internationally with events such as LEAP and Future Investment Initiative. The use of un-trained novices poses significant risks in operation, leading to session disruption, cancellation and damage to reputation, valued at SAR 2.5 million per mega-event.

On the other hand, the 65% of conference interpreter employees who are imported from international agencies are paid five times as much as their own nationals, resulting in significant economic costs and disrupting national localisation priorities. The need to build up a skilled domestic team of conference interpreters is a need both from an economic and strategic point of view, as it allows the Kingdom to keep the money earned from conferences within the country, while also protecting the diplomatic and commercial communication from foreign interpreters.

6.2 Structural Overhaul of Higher Education Translation Programs

In order to solve this problem of a mismatch in talent, Saudi universities need to modernize and update the curriculum of translation. Academic departments need to get away from the old idea that written translation and non-booth literary exercises are the things that should be taught.

As a potential direction for curriculum reform and future pedagogical investigation, Saudi higher education programs could explore progressively expanding supervised booth practice toward international postgraduate benchmarks, such as the 300 to 400 practice hours recommended by EMCI and AIIC training

frameworks (Setton & Dawrant, 2016), while integrating authentic institutional corpora and targeted training in the coping strategies identified in this study:

- Prospective syntactic anticipation;
- Clausal decoupling and chunking;
- Tactical condensation of rhetorical redundancy.

6.3 Infrastructure Optimization for Public Event Authorities

Over SAR 120 million has been invested in modern conference equipment such as soundproof booths and multi-channel wireless audio systems by the Saudi Conventions and Exhibitions General Authority and related institutions. But much of this equipment is unused because of the lack of qualified simultaneous interpreters in the country.

In doing so, formal university-industry partnerships and credentialing frameworks will serve to close this divide, helping to turn underused booths into active training labs and working conference tools. This assures public capital expenditures will be transformed into usable, area specific conference delivery.

Nevertheless, because a laboratory simulation of financial and technological plenaries cannot fully capture the sociological pressures of live multi-day summits or other specialized interpreting domains (e.g., legal, diplomatic, or medical interpreting), these educational and infrastructure proposals should be treated as tentative hypotheses requiring dedicated longitudinal and econometric validation.

7. Methodological Limitations

This investigation recognizes certain methodological shortcomings: First, the experimental booth setup and the audio equipment used was similar to those found in a professional conference, but laboratory simulation cannot completely reflect the social pressures, audience reactions and physical exertion of a multi-day international summit. Second, the 25-minute testing protocol was developed with the goal of assessing onset to cognitive fatigue and is consistent with standard international booth rotation windows. It does not evaluate cumulative performance deterioration of full-day assignments that involve 6-8 hours of alternating shifts between different booths. Third, the source material was limited to the themes of deep-tech innovation, sovereign finance and labour market governance. The results are not necessarily directly applicable to specialized settings such as medical, judicial, or multilateral treaty contexts, as these comprise approximately 85% of most LEAP and FII content. Fourth, although Module 1 (LEAP technological transformation in T1) and Module 2 (FII sovereign finance and labor localization in T2) were rigorously calibrated for syntactic complexity, delivery rate (140 wpm), and terminological density (15 items per interval), the fixed sequential presentation without AB/BA counterbalancing represents a potential threat to internal validity. Because novice interpreters exhibited slightly lower baseline lexical familiarity in sovereign finance and labor governance compared to

deep-tech domains, a portion of the fluency degradation observed in T2 may reflect an interaction between cumulative cognitive fatigue and domain-specific lexical difficulty. Future experimental replications should employ a counterbalanced crossover design (AB/BA) to fully isolate temporal fatigue from thematic domain effects. Fifth, because professional experience is a non-randomized, organismic grouping variable, between-group differences are naturally intertwined with co-occurring cohort characteristics, specifically specialized booth training (>100 hours in 100% of experts vs. 0% of novices) and pre-existing familiarity with economic and technological subject matter. Although statistical controls in our regression modeling ruled out chronological age and formal degree level as significant predictors ($p > .20$), statistical adjustment cannot entirely disentangle accumulated working hours from formal booth training and prior domain exposure; hence, the observed performance differences reflect a holistic expertise profile rather than years of service in isolation.

8. Conclusion and Future Directions

This quasi-experimental investigation examined the influence of professional experience on simultaneous interpreting fluency, terminology fidelity, and cognitive fatigue during sustained Arabic-to-English conference delivery within the context of Saudi Vision 2030 events. By evaluating performance across a standardized 25-minute simulation replicating high-level plenary discourse, the study provides definitive, empirical answers to its core research questions:

First, professional experience serves as the critical buffer against temporal fluency degradation. While seasoned conference interpreters maintain steady acoustic pacing and phonation continuity throughout prolonged turns, novice practitioners suffer acute communicative breakdown past the midpoint of a standard booth rotation, demonstrating that baseline bilingualism is insufficient to sustain delivery under continuous speed pressure.

Second, professional expertise decisively governs terminology fidelity in specialized public-private discourse. Experienced interpreters successfully rely on consolidated institutional schemas to render complex sovereign finance and labor localization terms, whereas novices frequently resort to literal calques or message abandonment when encountering non-standardized administrative neologisms.

Third, the resilience of expert practitioners is driven by proactive, metacognitive coping behaviors. The systematic deployment of early syntactic anticipation, clausal decoupling, and tactical condensation enables working memory preservation and prevents cognitive saturation during high-density Arabic-to-English structural reordering.

Future Research Directions:

To build upon these baseline empirical insights, future interpreting research in the Arabic–English context should prioritize three avenues: (1) longitudinal tracking of trainee cohorts undergoing intensive, booth-based curriculum interventions to determine the precise threshold of practice hours required to automate syntactic chunking; (2) incorporating non-invasive neuroimaging or biometric instrumentation (such as fNIRS, EEG, or pupillometry) to triangulate behavioral degradation metrics with direct indices of prefrontal cognitive workload; and (3) examining human-machine teaming in the booth, assessing how real-time AI-driven terminology prompts interact with visual attention and cognitive load across varying levels of interpreter expertise.

References

Aal-Hajiahmed, M. J. (2023). *Cognitive processing Arabic-English simultaneous interpreting* (Doctoral dissertation, Universitat Autònoma de Barcelona). https://ddd.uab.cat/pub/tesis/2023/hdl_10803_687927/mjah1de1.pdf

Al-Suhaim, D. S. (2024). Exploring theoretical dimensions in interpreting studies: A comprehensive overview. *Arab World English Journal for Translation & Literary Studies*, *8*(1), 15–43. https://doi.org/10.24093/awejtls/vol8no1.2

Al Zahran, A. (2021). Structural challenges in English>Arabic simultaneous interpreting. *The International Journal of Translation and Interpreting Research*, *13*(1), 51–70. https://doi.org/10.12807/ti.113201.2021.a04

Alharbi, M. A. (2025). Navigating AI-driven translation in Saudi Arabia's media: Challenges and opportunities. *Arabic Journal for Translation Studies*, *4*(12), 10–31. https://democraticac.de/?p=105854

Aluthman, E. S., & Al-Buraidi, H. M. (2024). A psycholinguistic approach to consecutive interpretation: Identifying problems among Saudi interpreters. *World Journal of English Language*, *14*(5), 229–239. https://doi.org/10.5430/wjel.v14n5p229

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Dong, D., & Chen, M.-L. (2025). Metacognitive strategies in translation: A comparative study of student and professional translators. *Humanities and Social Sciences Communications*, *12*(1), Article 5289. https://doi.org/10.1057/s41599-025-05289-7

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–703). Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.038

Fantinuoli, C. (2018). Computer-assisted interpreting: Challenges and future perspectives. In I. Durão (Ed.), *Trends in e-tools and resources for translators and interpreters* (pp. 153–174). Brill. https://doi.org/10.1163/9789004351790_009

Gile, D. (2016). Effort models simultaneous interpreting. In F. Pöchhacker, M. Liu, & H. Liu (Eds.), *Routledge encyclopedia of interpreting studies* (pp. 134–137). Routledge. https://doi.org/10.4324/9781315738268.ch27

Hervais-Adelman, A., Moser-Mercer, B., & Golestani, N. (2015). Brain functional plasticity associated with the emergence of expertise in extreme language control: The case of simultaneous interpreters. *NeuroImage*, *114*, 264–273. https://doi.org/10.1016/j.neuroimage.2015.03.072

Kingdom of Saudi Arabia. (2016). *Saudi Vision 2030*. https://www.vision2030.gov.sa

Kööp, A. M. (2019). Cognitive load and pause patterns in simultaneous interpreting: An empirical study of professional and novice interpreters. *Target*, *31*(3), 405–432. https://doi.org/10.1075/target.18121.koo

Liu, M., Schallert, D. L., & Carroll, P. J. (2004). Working memory and expertise in simultaneous interpreting. *Interpreting*, *6*(1), 19–42. https://doi.org/10.1075/intp.6.1.04liu

Macnamara, B. N. (2012). The role of working memory capacity in simultaneous interpreting. *Translation and Interpreting Studies, 7*(2), 216–233. https://doi.org/10.1075/tis.7.2.06mac

Mellinger, C. D., & Hanson, T. A. (2017). *Quantitative research methods in translation and interpreting studies*. Routledge. https://doi.org/10.4324/9781315647708

Moser-Mercer, B., Künzli, A., & Korac, M. (1998). Prolonged turns in interpreting: Effects on quality, physiological and psychological stress (Pilot study). *Interpreting, 3*(1), 47–64. https://doi.org/10.1075/intp.3.1.03mos

NES Fircroft. (2025). *Saudi Arabia talent trends and workforce localization report*. NES Fircroft Insights.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language, 87*(3), 539–558. https://doi.org/10.1353/lan.2011.0057

Pöchhacker, F. (2004). *Introducing interpreting studies*. Routledge. https://doi.org/10.4324/9780203504802

Pradas Macías, E. M. (2009). Céilleur's pauses in simultaneous interpreting: An empirical study. *Sendebar, 20*, 199–226. https://revistaseug.ugr.es/index.php/sendebar/article/view/178

Rendón, C., & Moser-Mercer, B. (2000). Expert-novice differences in simultaneous interpreting: The acquisition of cognitive skill. *Interpreting, 5*(2), 133–155. https://doi.org/10.1075/intp.5.2.05ren

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing methods and new perspectives. *Interpreting, 13*(2), 176–204. https://doi.org/10.1075/intp.13.2.02see

Seeber, K. G. (2013). Cognitive load in simultaneous interpreting: Measures and methods. *Target, 25*(1), 18–32. https://doi.org/10.1075/target.25.1.03see

Setton, R. (1999). *Simultaneous interpretation: A cognitive-pragmatic analysis*. John Benjamins Publishing Company. https://doi.org/10.1075/btl.28

Setton, R., & Dawrant, A. (2016). *Conference interpreting: A complete course*. John Benjamins Publishing Company. https://doi.org/10.1075/z.199

Shlesinger, M. (2000). Strategic allocation of cognitive resources in simultaneous interpretation. *Interpreting, 5*(2), 119–131. https://doi.org/10.1075/intp.5.2.04shl

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation: Cognition in education* (Vol. 55, pp. 37–76). Academic Press. https://doi.org/10.1016/B978-0-12-387693-5.00002-8

Timarova, Š., Čeňková, I., Meylaerts, R., Hertog, E., Szmalec, A., & Duyck, W. (2014). Simultaneous interpreting and working memory executive control. *Interpreting, 16*(2), 139–168. https://doi.org/10.1075/intp.16.2.01tim

Tissi, B. (2000). Silent pauses and disfluencies in simultaneous interpretation: A descriptive study. *The Interpreters' Newsletter, 10*, 27–42. https://hdl.handle.net/10077/2436

Uçar, N. G. (2025). Working memory in interpreting and translation research. *Adnan Menderes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 12*(2), 124–142. https://doi.org/10.30803/adusobed.1810413

Van Besien, F. (1999). Anticipation in simultaneous interpretation: A psycholinguistic approach. *Babel, 45*(3), 250–259. https://doi.org/10.1075/babel.45.3.06van

Wenda, W., & Nur Salam, A. A. (2025). *Accuracy and fluency in interpreting*. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5774363