

WHEN FLUENCY BECOMES FREE: A PRODUCTIVE FRICTION FRAMEWORK FOR LANGUAGE EDUCATION AND LINGUISTIC RESEARCH IN THE AGE OF ARTIFICIAL INTELLIGENCE

Hassan Raza

(Corresponding Author)

M.Phil Scholar, Department of Applied Linguistics, Government College University Faisalabad, Punjab, Pakistan.

Email: hassanmatyana8@gmail.com

Aymen Rehmani

M.Phil Scholar, Department of Applied Linguistics, Government College University Faisalabad, Punjab, Pakistan.

Email: aymenrehmani135@gmail.com

Dr. Muhammad Asim Mahmood

Professor, Department of Applied Linguistics, Government College University Faisalabad, Punjab, Pakistan.

Email: masimrai@gmail.com

ABSTRACT

Over the centuries, the ability to write fluently became a costly indicator of underlying competence: it was assumed that a fluent well-organized text demonstrated a sustained period of exposure, effort, and skill. Large language models (LLMs) have broken that connection as they have made fluent and register appropriate text available in seconds. The central issue this article addresses is that the most significant issue that AI poses to language education is not plagiarism or factual error but the breakdown of fluency as an indicator of linguistic knowledge, cognitive effort, and communicative interest, as summarized in the Fluency Decoupling Thesis (FDT). The article is based on second language acquisition, cognitive psychology, language assessment, sociolinguistics and natural language processing, and builds the Productive Friction Framework (PFF) consisting of four constructs (linguistic offloading, fluency inflation, register flattening, and productive friction), a five-part typology of the design of the friction (in terms of source, target, type, and magnitude), and three pillars (provenance over product, friction by design, and linguistic justice). A three-study, mixed-methods research programme is outlined, with a Fluency Inflation Index, a qualitative taxonomy of offloading actions, and an analysis of the performance of LLM across linguistic varieties in Pakistan specified, as well as five falsifiable hypotheses and a power analysis. There are five classroom protocols provided for immediate use. The article ends with the idea that language education should "tolerate costless fluency" and not police machines, but rather re-think tasks to make human judgment, voice and intent visible and assessable. No empirical data are provided, the article is a conceptual contribution with a research design provided and pre-registered.

Keywords: fluency decoupling, productive friction, linguistic offloading, generative AI, language assessment, multilingualism, Pakistani English

Introduction

Think about two argumentative essays that were sent to the same university writing course. They are both coherent, grammatically correct, appropriately hedged, and have a clear thesis. The first took more than six hours to write, draft, misjudge, reread and revise by a learner. The second was created in 40 seconds by a learner who cannot tell you why any of the sentences in the text are formed the way they are. There is no way for a reader with only the two texts to determine which is the other. Most important of all, the reader cannot any longer deduct from texts, without further information, what either learner knows.

This little scenario depicts a change in the ecology of language use. Expensive fluency has always been the way to get informative text; that's what made it informative. Well-formed language was treated as a proxy for competence because for the most part it was: rarely, a person who produced accurate academic prose did not have a lot of what it takes. That's an assumption made by large language models (LLMs) trained on web-scale text (Brown et al. 2020) using the transformer architecture (Vaswani et al. 2017). Producing fluent, grammatically correct and register-appropriate text is now possible, in many languages, at low, if any, cost.

The institutional response has been mostly defensive. Universities are using detection software, limiting tool use or going back to paper exams. These responses are based on a diagnosis in which the issue is the inappropriate use of a sound system of assessment. There are three reasons that this article does not agree with that diagnosis. Second, detection is not reliable; indeed, for many learners of English, who represent a large portion of the world's language education population, detection is biased: Liang et al. (2023) found that detection exceeded 50% of a sample of non-native essays that were considered machine-generated. Third, detection is inaccurate and easy to get around: independent studies have shown that detection tools are inaccurate and can be defeated (Sadasivan et al., 2023; Weber-Wulff et al., 2023). Second, prohibition is pedagogically unfeasible because of what today's students will encounter in the professional world where AI-generated writing is commonplace. Third, and most importantly, the problem is not diagnosed correctly on defense. The problem belongs to the epistemological category. Assessments and classroom procedures were loosely aligned to a world where production was a measure of competence, and that world is over.

The question of the present article is therefore different from the one that is being debated in the institutional sphere. Instead of wondering how to block the use of AI, it's asking the question of what linguistic competence is, and how it can be taught and demonstrated, when it is not costing anything. To answer these questions, the analysis is guided by three research questions:

RQ1. In what precise senses does fluent output cease to be valid evidence of a language user's knowledge, effort, and intent when generative AI is available?

RQ2. What pedagogical and assessment designs preserve the learning mechanisms that AI assistance risks bypassing, while still allowing learners to benefit from AI?

RQ3. How do these questions change in multilingual ecologies, such as Pakistan's, where the languages and varieties learners actually use are unevenly represented in AI systems?

The present article makes four contributions. First, it presents the Fluency Decoupling Thesis (FDT), which spells out the type of inferential connections that AI cuts. Secondly, it proposes four constructs with operationalization that take the researchers from anecdote to measurement. Third, it suggests the Productive Friction Framework (PFF) and a typology of friction design consisting of five parts that teachers can use directly. Fourth, it outlines a three-study research programme, with hypotheses and a power analysis, with special focus on Pakistan's multilingual ecology.

Some clarification on scope is required. This is an idea article. It lacks new empirical data and all its statements about learner behaviour are in the form of hypotheses. The design aims to provide a vocabulary, measurable indices, and an implementable design that can be used by researchers and teachers in resource limited multilingual environments. The article further reviews the literatures, establishes the framework, outlines the research programme, provides classroom protocols and addresses implications and limitations.

Literature Review

Fluency as a Signal: Competence, Performance and the Assessment Inference

Applied linguistics has long been concerned with the differences between what a language user knows and what a language user does. Hymes (1972) took Chomsky's idea of competence one step further and introduced communicative competence which incorporates knowledge of when, where and with whom the language is appropriate. Canale and Swain (1980) made these into grammatical, sociolinguistic, discourse, and strategic components, which are continued in the construct definitions of the language assessment today (Bachman & Palmer, 2010). In these accounts, the observable behaviors serve as evidence for the unobservable competence, and the quality of the assessment is based on the support of that evidence.

Validity theory elucidates the inferential structure. Messick (1989) redefined what he called validity as an integrated judgment on the adequacy of inferences made from scores, while Kane (2013) redefined validation as “the examination of an argument from observed performance to meaning and use.” The indices of complexity, accuracy, and fluency (CAF) have similar functions in research on second language development, as they are performance measures that are interpreted as an underlying state of development (Norris & Ortega, 2009). All of these traditions take for granted that only the person who creates the language is the person whose ability is being extrapolated, and that the process of creating the language is a process that is reflected in the score. Generative AI is taking aim at that leap of faith. It is not just a new form of cheating, it is a change in the demonstrable relationship between text and producer.

Form and Function: What Large Language Models Do and Do Not Do

To understand why fluency has decoupled, it's important to understand what LLMs do. Bender and Koller (2020) pointed out that a system trained only on linguistic form does not have access to meaning (i.e. the relation between expressions and communicative intent), and Bender et al. (2021) warned of the readers' ability to draw meaning from fluent output. An important improvement was added by Mahowald et al. (2024). They were able to differentiate between formal linguistic competence (knowledge of rules and patterns) and functional linguistic competence (the use of language in the world, including reasoning, pragmatics, and social cognition), and demonstrated that LLMs today do well on the former but struggle on the latter.

In the case of human learners, there is no isolation of these two competences, and both of them develop together. When a learner can write a well-formed concessive clause, he or she has typically satisfied the situations in which concession can be helpful in argumentation. The two are dissociated by LLM. They provide training in skill competence which does not require the learner to possess either competence. With post-training techniques that align models to human

style, it is easy to identify the output as having a characteristic house style: cooperative, balanced, hedged, and generally polished. As a result, the words on the screen might be grammatically perfect and contain very little of the person that asked for them.

Struggle as Mechanism: Output, Noticing, and Desirable Difficulty

A second body of work is provided that provides an explanation as to why the decoupling is important for learning and not just for assessment. While Swain's (1985) output hypothesis suggested that language production, especially when done in ways that extend learners' comfort zone, encourages the processing of syntax rather than just meaning. These were later elaborated by Swain (1995) who listed several functions of output, such as noticing gaps between what one wants to say and what one can say, testing hypotheses about the language, reflecting on language explicitly. This account is complementary to Schmidt's (1990) noticing hypothesis which holds that difficulty will strike while someone is communicating will stimulate acquisition, and Long's (1996) interaction hypothesis which indicates that difficulty in communication will bring about acquisition.

Cognitive psychology talks about the same thing. Craik and Lockhart (1972) demonstrated that the more deeply a memory is processed, the more enduring the memory trace will be. Bjork (1994), and Bjork and Bjork (2011) suggested that spacing, interleaving, and retrieval practice are conditions that can help learning seem harder, while they simultaneously help learning feel more permanent. Roediger & Karpicke (2006) showed that retrieval practice can lead to improved long-term retention compared to restudy. According to the cognitive-process tradition of writing research, writing is a process of problem solving, where planning, translating, and revising are trying and creative (Flower & Hayes, 1981; Kellogg, 2008). In other words, effort isn't just a part of the side effects of learning, it is part of the learning process.

The socio-cultural theory offers a counterpoint. The zone of proximal development (ZPD) as developed by Vygotsky (1978), and scaffolding (Wood et al., 1976), is support that is contingent and temporary; that is, it is faded as competence increases. AI support can serve as scaffolding when designed to support the learner and then gradually dropped. When it does it, it is used as substitution and does not withdraw. The difference between scaffold and substitute is the fulcrum on which the structure below pivots.

Cognitive Offloading and Its Costs

Cognitive offloading is defined by Risko and Gilbert (2016) as “the use of physical action to change the information-processing demands of a task in such a way as to lower cognitive demand”. Offloading is often rational, and the literature on cognitive load (Sweller, 1988) reminds us that cognitive free space for learning can be gained when extraneous load is reduced. Offloading has proven to have some costs too. Sparrow et al. (2011) discovered that individuals recalled fewer items, when they could be retrieved outside of their own memory. As for generative AI, the early data suggests the same. In a field experiment with secondary-school mathematics students, Bastani and colleagues (2025) observed that, while allowing students to freely use a GPT-based assistant during practice led to better performance, using it during a subsequent unassisted examination resulted in worse performance, and restricting students' use of it with pedagogical safeguards reduced that performance loss. Kosmyna et al. (2025) have

found weaker neural connectivity and poorer recall of their own text in a small, prepublication EEG study of essay writing, with participants writing with an LLM showing poorer performance than participants who wrote without LLM assistance. They are small, limited studies and the latter is non-peer reviewed; they are meant to be suggestive, not definitive.

In the field of second language writing, both affordances and contradictions of second language writing have been recorded. Warschauer et al. (2023) proposed that the AI text generators are capable of assisting multi-lingual writers and can introduce unanswered issues about authorship, learning, and equity. Godwin-Jones (2022) explained that AI can be used as a tool or as a partner with the learner and emphasized the importance of learner agency; Kasneci et al. (2023) discussed both the opportunities and risks of using LLM in the educational context. The literature has not yet provided a systematic description of which parts of linguistic labor learners outsource, their relationship to retention, and what types of tasks can be constructed to ensure that productive aspects of effort remain in the hands of the learner. The following constructs and studies are suggested to solve this.

Linguistic Diversity, Standardization, and AI

The fourth literature is related to the language of whose AI systems are recreated. The majority of the world's languages is severely under-resourced in language technology and a small number of languages dominate, Joshi et al. (2020) demonstrated. Bias in NLP has often been studied without a clear description of the social power dynamics, and Bender and Friedman (2018) called for documentation of the populations and varieties included in data sets. Past studies of bias in NLP often lack a description of the social power dynamics, and Bender and Friedman (2018) encourage the documentation of the populations and varieties present in data sets. Ortega (2019) urged L2 research to adopt the perspective of equitable multilingualism instead of relying on norms of monolingual native speakers.

World Englishes scholars have long claimed that “standard” is a sociopolitical assessment and not a linguistic reality (Bhatt, 2001; Kachru, 1985, 1992). The study of translanguaging and translingual research also views code-mixing not as a problem, but as a resource (Canagarajah, 2013; García & Li Wei, 2014). Demonstrating the symbolic capital of valued forms of language, and the reproduction of inequality through institutions' differential reward for different forms of language, Bourdieu (1991) reminds us. A system which has consistently created the globally dominant register may become a tool for the reproduction of that register.

Pakistan is an example of the consequences. The mother tongues are Urdu (national language), English (official language) and Punjabi, Pashto, Sindhi, Saraiki, Balochi and many others are also widely spoken (Rahman, 1996). English maintains a high status in higher education and elite jobs (Mansoor, 2005; Rahman, 2002) as well as it has its own documented phonological and lexicogrammatical features in Pakistani context (Baumgardner, 1993; Mahboob & Ahmar, 2004). The use of Roman Urdu, and of fluid Urdu-English code-mixing is also significant in everyday digital communication. Urdu shows a 3-way honorific distinction in its pronoun system and it has a highly differentiated kinship vocabulary which is pragmatically consequential (Brown & Levinson, 1987). The capabilities of LLMs in supporting these features

are empirical and have implications for learners, and there is limited research on how well they perform with the Pakistani varieties.

There are no reason-invoking research gaps; indeed, multilingual benchmarking studies have revealed lack of consistency across languages in general (Ahuja et al., 2023). When these literatures are combined, there is an opening. Assessment theory explains why the inference of fluency is not robust; second language acquisition and cognitive psychology explain why effort is a part of learning; the AI literature explains what models do and do not provide; and sociolinguistics explains why the amplification of whose language matters. There is no current system that incorporates all four. The following section will try to do that integration.

The Productive Friction Framework

The Fluency Decoupling Thesis

The framework starts with a single proposition, which is given in full here.

Fluency Decoupling Thesis (FDT). In AI-mediated environments, the linguistic competence of the output no longer yields insight into the linguistic competence, thinking effort, or communicative purpose of the producer.

The thesis identifies three different inferential connections, each of which in the past provided evidence for reading a text. The knowledge link relates fluent form to the producer's knowledge of the language. The effort link is a mechanism that relates the quality of a product to the length of time and the concentration required in order to create the product, as was described above. The intent link is the link between choices in text and the intended meaning of the producer, as Bender and Koller (2020) see it as the essence of meaning. Each link can fail independently. A learner who edits a paragraph he or she wrote with LLM keeps the knowledge links and diminishes the effort link. A learner who asks for a full essay on a topic that they have not considered breaks all three.

Two corollaries follow. Firstly, the validity of product only assessment diminishes with the increase in availability of AI, as the likelihood that a given fluent product is a reflection of the producer's own ability decreases with the greater ease of AI. Second, the loss is not even. It is highest when there is the largest gap between the unaided and aided output, which is likely to be among lower-proficiency learners, where accurate diagnosis is most important. This corollary is testable and is elaborated in the following hypothesis: Hypothesis 1: The more intense the training, the more intense the learning.

Four Core Constructs

The only reason that the FDT is useful is if its effects are measurable. Four constructs are thus postulated and summarized in Table 1. Each is defined in such a way as to be operationalizable for the studies described later.

Table 1

Core Constructs of the Productive Friction Framework

Construct	Definition	Proposed operationalization
Linguistic offloading	Delegation of lexical, syntactic, discourse, argumentative, or stylistic decisions to an AI system	Coded proportion of decisions delegated, by level of language (Study 2 taxonomy)
Fluency inflation	The gap between the quality of a learner's AI-assisted output and their unassisted output	Fluency Inflation Index (FII and FII _z)
Register flattening	Convergence of AI-mediated text on a generic, globally neutral register that erases local voice, pragmatics, and code-mixing	Register Flattening Index (RFI); learner-level voice-retention rating
Productive friction	Designed points at which the learner must decide, justify, revise, diverge from, or delay engagement with AI output	Friction types (Table 2); Editorial Agency Index (EAI)

Linguistic offloading is the delegation of lexical, syntactic, discourse, argumentative, or stylistic decisions to an AI system. The construct builds on Risko and Gilbert (2016)'s cognitive offloading to the specific components of language work. The benefit of using it is that they are very granular. The question "Did the learner use AI?" leaves room for only a binary answer and doesn't provide much clarity. The question "Which decisions did the learner delegate and at which level of language?" produces a profile. The working hypothesis is that offloading is not always detrimental: delegating spelling correction and delegating argument construction are two very different things when it comes to learning.

Fluency inflation is the difference between the quality of a learner's AI-assisted writing and the quality of his or her unassisted writing. It is measured by the Fluency Inflation Index (FII) as described in the research design below. The FDT is talking about the threat to validity—the amount of fluency inflation—that is not the thing that assessors see when they do not observe the learner unassisted in the same task.

Register flattening is the tendency of AI-generated language to converge on a globally neutral, standardized register, erasing local voice, pragmatic conventions, and code-mixing. It has been suggested as a building block of AI evaluation and sociolinguistics. It can be measured at the level of the output (Register Flattening Index, RFI, indicating the percentage of outputs where locally expected features are missing or removed), as well as at the level of the learner (the percentage of the learner's submitted text that contains features typical of his/her own unassisted writing).

Productive friction is designing moments in a task where the learner must make choices, explain why to do, make changes to, go against, or pause an interaction with a piece of AI-generated content. It is vital that the term “productive” be included. Friction does not equal obstruction and friction doesn't equal hard. In what Sweller (1988) calls cognitive load, the idea is to eliminate conflicting load (e.g., the effort of retyping, or looking for a synonym), and either maintain or increase germane load (the effort of processing that adds schemas). To be productive, friction serves to focus effort on the processes that underlie language acquisition – noticing, hypothesis testing, retrieval, and reflection (Schmidt, 1990; Swain, 1995).

The fifth measure, Editorial Agency Index (EAI), implements the notion that a learner's choices of AI output is an actual evidence of competency. When a learner maintains a record of the AI suggestions and his or her reactions to them, the EAI is the percentage of suggestions that the learner rejected or modified and stated a linguistic or rhetorical reason. A high EAI does not indicate the learner's unwillingness to accept help, rather, the learner's judgement regarding the help. The pivot of the first pillar (described below) is the EAI.

A Typology of Friction Design

Teachers should understand what productive friction means in practice if it is a pedagogical remedy. Based on the mechanisms listed above the following 5 types of friction are identified (Table 2). The typology should be considered as a generator: for every AI-supported task, what are the five types that it is missing?

Table 2

A Five-Part Typology of Friction Design

Friction type	Learner action	Primary mechanism	Example task
Decision	Accept, reject, or modify each suggestion	Noticing; evaluation	Three-column log
Justification	State the linguistic or rhetorical reason for each decision	Metalinguistic reflection	“Because...” statements in the log
Divergence	Locate and analyze a flaw in AI output	Critical analysis; hypothesis testing	Blind spot hunting
Localization	Adapt language to a specific audience, variety, or voice	Sociolinguistic awareness; voice	Translate back; voice preservation task

Friction type	Learner action	Primary mechanism	Example task
Delay	Reproduce or explain one's own text after an interval	Retrieval; spacing	Delayed defence

The five types are linked to certain learning mechanisms. The noticing and evaluation (Schmidt, 1990) processes are engaged in decision friction. Justification friction leads to metalinguistic reflection, a function of output, according to Swain (1995). The goal of the divergence friction is to take the AI-generated content and provide it as an objective to analyze and critique, not accepting it as the solution. Localization friction is the learner's need to bring back or modify the language to suit a particular target audience, variety or voice in a particular context. Delay friction leverages the spacing and retrieval effects shown by Bjork and Bjork (2011) and Roediger and Karpicke (2006) and it also offers the added benefit of observing ownership of a text.

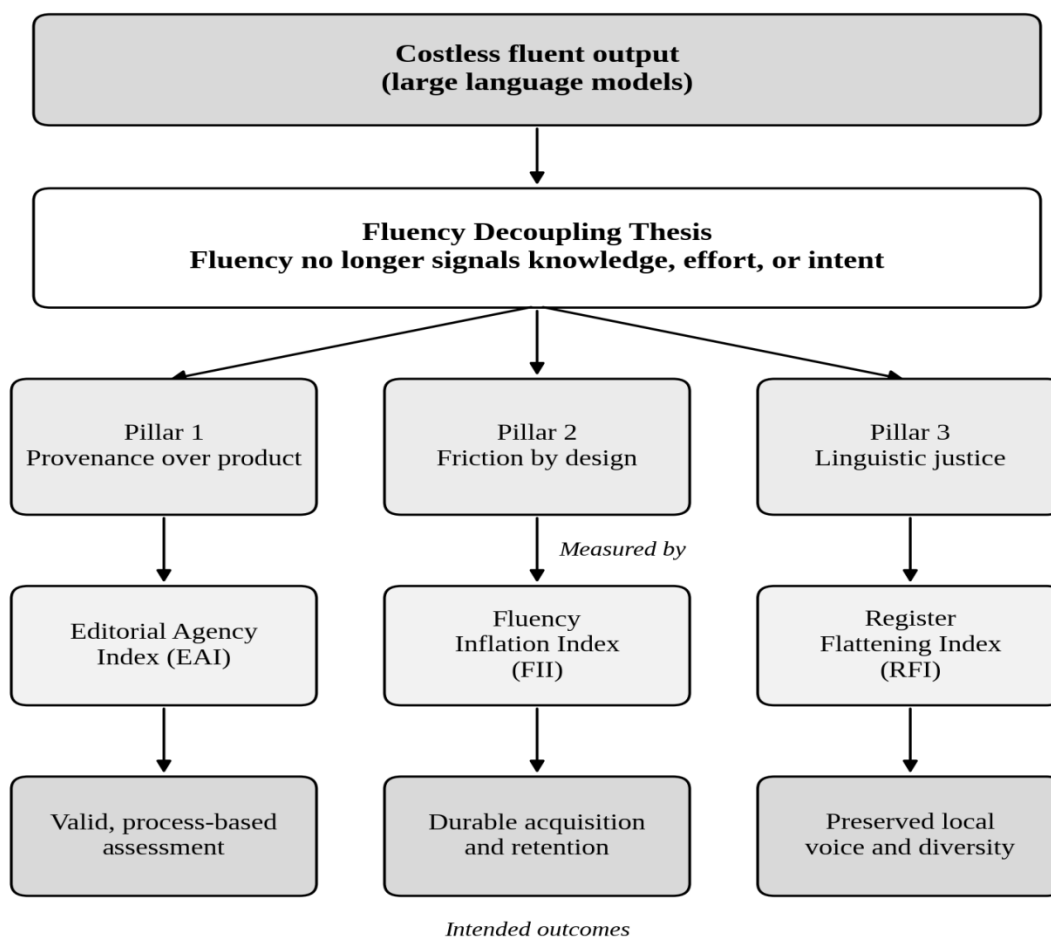
Friction should not be constant, it should be calibrated. The scaffolding principle (Wood et al., 1976) suggests that the amount of AI support is greater for beginners and the amount of friction (e.g., decision friction on a limited number of suggestions) is lighter for them, whereas tasks for advanced learners are more likely to have more divergence and localization frictions. Friction should not diminish but change as the unassisted competence increases: the skills of evaluating the output of machines, resisting its output, and adapting to it are advanced competence skills that are worthy of teaching.

Three Pillars

The framework has three pillars (see Figure 1) and the constructs are operationalizations of these pillars.

Figure 1

Conceptual Model of the Productive Friction Framework



Note. EAI = Editorial Agency Index; FII = Fluency Inflation Index; RFI = Register Flattening Index.

Pillar 1: Provenance over product. When the product no longer does the talking, then the assessment needs to focus on the path to the product. Examples of evidentiary evidence of student learning in the form of prompts, drafts, rejected suggestions, and learner justifications. The learner's editorial choices are the main evidence of competence and the EAI gives a first quantification. This pillar reflects a tradition of assessment that is formative, values the use of process evidence (Black & Wiliam, 1998; Hattie & Timperley, 2007) and, importantly, places the burden of proof on the learner rather than the assessor as the assessor is required to now see evidence of judgement instead of evidence of machine text.

Pillar 2: Friction by design. At the very least, each task performed with the aid of the AI must include a single task at which the machine is unable to complete what the learner must perform. This doesn't mean that AI itself needs to be banned. It involves identifying where effort

is needed, safeguarding that effort, and delegating it to AI for other tasks. According to self-determination theory, there is a further reason for caring: tasks that are meaningful and autonomy supportive maintain motivation (Deci & Ryan, 2000), while friction that is perceived as being arbitrary punishment will demotivate. Friction should then be explained to learners as a communicative goal, be tied to communicative goals, and be calibrated as outlined above.

Pillar 3: Linguistic justice. The framework becomes a way of thinking about local languages and varieties as a research and protection object rather than noise. Specifically, it calls for the development of tasks that will require voice preservation as a learning goal, for assessing the impact of AI systems on the varieties used by learners, and for identifying voice flattening as a harmful effect instead of a stylistic curiosity. This pillar is inspired by the traditions of the World Englishes and translanguaging traditions (Canagarajah, 2013; Kachru, 1985; García & Li Wei, 2014) and by Ortega's (2019) call for equitable multilingualism. Furthermore, it includes a warning about the equity aspect of Pillar 1: if a detector fails to recognize non-native writing, then any system that assigns a writing source must be designed so that multilingual learners do not face an unfair disadvantage in having their writing suspected.

The pillars together address the first two research questions. The shift in the evidence from product to provenance (RQ1) restores the validity of assessment. Learning mechanisms are safeguarded by creating friction where it is needed (RQ2). The third pillar, as well as the audit described below, reflects the multilingual dimension (RQ3).

A Three-Study Research Programme

Overview and Rationale

The framework yields a list of empirical questions which can't be answered by any individual study. The following programme is therefore proposed, combining a quantitative experiment (Study 1), to determine the size and distribution of fluency inflation, with a qualitative study (Study 2), to explain what is delegated to the AI systems and why, and with an audit (Study 3), to assess how the AI systems are operating on the Pakistani varieties central to the third pillar. The studies are interconnected by sharing the results in a common display (Creswell & Plano Clark, 2018): qualitative results are used to explain the heterogeneity in the quantitative results, and audit results are used to detail the content of the localization tasks. As a matter of fact, this programme is a design. No data has been collected yet and the hypotheses below are presented in falsifiable form, to allow for preregistration and to be falsified, if they are false.

Study 1: The Fluency Inflation Index

Aims and Hypotheses

The present Study 1 estimates the level of fluency inflation, investigates whether it depends on the level of fluency, and assesses whether the "friction" in the task shortens the gap between the assisted output and later unassisted performance. Three hypotheses are tested as well as a rater-discrimination hypothesis.

H1. Fluency inflation will be larger for lower-proficiency learners than for higher-proficiency learners.

H2. Compared with unrestricted AI assistance, friction-designed tasks will produce a smaller gap between assisted output quality and delayed unassisted performance.

H3. Learners who offload argument and structure will show weaker delayed retention of their own text than learners who offload only surface form (lexical and grammatical editing).

H4. Trained raters will not reliably distinguish AI-assisted from unassisted texts on the basis of the texts alone, but will do so significantly better when given process data (H5 is tested in Study 3).

Participants

The proposed sample consisted of 120 undergraduate and postgraduate students of English as a foreign or additional language (L2) studying at a public university in Punjab, Pakistan, divided into three English language proficiency bands (lower, intermediate, upper; $n = 40$ each) on the basis of a placement test. Based on a within-subjects design ($N = 120$), a sensitivity analysis suggests that the design has 80% power to detect a standardized paired difference of $d_z = 0.26$ (two-tailed, $\alpha = .05$; Cohen, 1988; Faul et al., 2007). This is sufficient for major differences in conditions. It is not sufficient to detect small interactions, and interaction effects (including comparison of proficiency in H1) will therefore be reported with confidence intervals and interpreted as estimates, not confirmatory tests.

Design and Materials

The study is a 3×3 mixed method design with condition as the within subjects factor. Unassisted: Learners write an argumentative text without consulting AI. When they write in the assisted condition, they can access a researcher-setup LLM interface where all the prompts and responses are recorded. In the friction designed condition, they write with the same interface, but need to take 1 step to write a log (AI suggestion, response), 1 step of divergence (identify a problem in an AI suggestion), and 1 step of localization (adapt a passage for local audience). Three parallel argumentative prompts that have been pretested to be of similar difficulty are counterbalanced across conditions using a Latin square design. To minimize carry-over, sessions are spaced at least 1 week apart.

Procedure and Measures

Each writing period (60 minutes) is followed 7 days later by a delayed unassisted task that requires the learner to reconstruct the central argument of the text in 15 minutes without consulting the text, and to answer recall questions about their choices and construction of lexical and structural elements. This gives unassisted performance score for each condition with a delay.

There are three measures of text quality. Lexical Sophistication is measured using the Tool for the Automatic Analysis of Lexical Sophistication (Kyle & Crossley, 2015). The L2 Syntactic Complexity Analyzer (Lu, 2010) is used to calculate syntactic complexity. Coh-Metrix indices (McNamara et al., 2014) are used to calculate cohesion and discourse features. Three

trained raters, unaware of condition, also give a holistic and an analytic rating, both which are adjusted for rater severity using many-facet Rasch measurement (Linacre, 1989). Intraclass correlation coefficients will be used to measure interrater reliability with an acceptable level of .75 (Koo & Li, 2016). These indices are combined into a composite quality score that is scaled to a 0–100 range.

The Fluency Inflation Index

The primary index is as follows: A is the composite quality score with AI assistance and U is the composite quality score without AI assistance:

$$FII = (A - U) / U$$

A standardized form of ratio index also is calculated, due to the fact that ratio indices are unstable when U is small.

$$FII_z = (A - U) / SD_U$$

SDU is an unassisted scores' standard deviation in the sample. Both variants will be reported. The retention gap is defined analogously as the difference between the assisted quality and the delayed unassisted performance, and is measured in standardized units.

Analysis Plan

Linear mixed-effects models with random intercepts for participants and prompts are used for primary analyses (Bates et al., 2015). The condition $\times$ proficiency interaction is investigated on quality scores for H1. The retention gap between the assisted and friction-designed conditions is tested for H2. The offloading codes generated in Study 2 are applied to the remaining participants' Study 1 logs and delayed retention is modeled as a function of the proportion of argument-level and structure-level offloading. In H4, six trained raters independently rate 60 texts (20 in each condition) twice – with and without process data – and the results of accuracy and agreement of the ratings are compared between the two rating phases. All hypotheses, exclusion rules, and the analysis code will be preregistered before data collection.

Study 2: The Offloading Taxonomy

Study 2 questions what learners outsource to AI, what level of language, and why. Thirty participants from Study 1 (10 per proficiency band) provide screen recorded writing sessions from the assisted condition, the applicable logs, and stimulated-recall interviews (Gass & Mackey, 2000) conducted within 48 hours. During the interviews, participants watch selected sections of their own session, and describe their reasoning. Interviews are done in the chosen language of the participant (Urdu, Punjabi or English); and the excerpts reported are translated by bilingual researchers with back-translation by another bilingual researcher.

Analysis is a mix of deductive and inductive coding. An a priori matrix crosses three dimensions: the level of language delegated (lexis, grammar, discourse structure, argument, stance and tone, language choice), the mode of interaction (requesting, accepting, adapting,

rejecting) and the motive (efficiency, insecurity, uncertainty about content, curiosity and learning). The matrix is then refined using inductive reflexive thematic analysis (Braun & Clarke, 2006), which helps to uncover patterns not expected by the matrix. 20% of the data are coded independently by two different coders, and Cohen's (1960) kappa is used to assess agreement, and a desired level is .70 or above. Reflexive memos, an audit trail, and member checking of the developing taxonomy with a smaller number of participants contribute to the trustworthiness. The resulting taxonomy should be empirically informed and subsequently analyzed in the remaining 90 Study 1 participants' logs in the quantitative evaluation of H3.

Study 3: The Pakistani Linguistic Audit

Aims and Hypothesis

Study 3 compares the performance of LLMs in the linguistic varieties of Pakistani learners' communicative reality. It is meant to enable measureability of flattening of a register and to test the following hypothesis.

H5. Pragmatic errors and register flattening in LLM output in Punjabi and Roman Urdu will be much higher than in English, when using parallel prompts.

Design

Five variants have been included: Urdu (standard urdu script), Roman Urdu, Punjabi (in Shahmukhi script), Urdu–English code-mixed and Pakistani English. Ten prompts are written for each of the six pragmatic domains for each variety (a) honorific address, (b) kinship terms, (c) requests, refusals, and other politeness-sensitive acts, (d) culturally embedded formulae (greetings, condolences, etc.), (e) idioms and proverbs, and (f) register shifting (formal letter to close friend, etc.). This yields 300 prompts. In English, bilingual linguists create parallel prompts to invoke comparable pragmatic demands. For each prompt, at least three popular LLMs are provided with fixed and documented configurations, and two sample outputs are generated per prompt, for a total of ~1,800 outputs. Since models evolve over time, the model identifier, access date and settings are recorded for each output.

Rating and Measurement

Table 3 provides the rubric that is used to rate each output independently by two native or near-native speakers who are linguistically trained in the form of the output. Many-facet Rasch measurement is used for analyzing ratings, and intraclass correlations are used to assess reliability. The Register Flattening Index (RFI) is found by calculating the percentage of outputs, where one or more LFE (required honorific, kinship term, code-mixing, or culturally expected formula) is either omitted or replaced with a generic equivalent or normalized to a different variety without being asked to do so. The RFI is reported both variety and domain.

Table 3

Rubric for Rating LLM Outputs in the Pakistani Linguistic Audit

Dimension	Description	Scale
Grammatical accuracy	Conformity to the grammar of the requested variety	1–5
Pragmatic appropriateness	Politeness, honorifics, and speech-act conventions suited to the situation	1–5
Cultural accuracy	Correct use of culturally embedded formulae, references, and idioms	1–5
Register fidelity	Match to the requested level of formality and audience	1–5
Variety fidelity	Retention of requested script, code-mixing, and local features without unrequested normalization	1–5, plus binary flattening flag

Note. The binary flattening flag feeds the Register Flattening Index (RFI).

Following Bender and Friedman's (2018) documentation guidelines, all prompts, rubrics, and anonymized ratings will be shared to enable the audit to be replicated and expanded. Another goal is to be developmental – to create a core of an evaluation data set for Pakistani varieties, to shift the researchers of the region from the subject of language technology to its contributors.

Hypotheses and Their Mapping

Table 4 below is a summary of the five hypotheses as well as the studies that test them and the results that would contradict each of the hypotheses. Advance plans of any outcomes that are stated as disconfirming is meant to maintain the integrity of the programme.

Table 4

Hypotheses, Tests, and Disconfirming Outcomes

	Prediction	Study	Outcome that would count against it
H1	Fluency inflation is larger for lower-proficiency learners	1	No proficiency gradient, or a reversed gradient
H2	Friction-designed tasks yield a smaller retention gap than unrestricted assistance	1	No difference, or a larger gap

	Prediction	Study	Outcome that would count against it
H3	Offloading argument and structure predicts weaker retention than offloading surface form	1, 2	No difference in retention by offloading level
H4	Raters cannot reliably distinguish assisted from unassisted texts without process data	1	Above-chance discrimination from text alone, or no gain when process data are added
H5	Punjabi and Roman Urdu outputs show more pragmatic error and higher RFI than English	3	Equal or lower error and RFI

Note. RFI = Register Flattening Index.

Ethics, Data Governance and Open Science

The programme is participatory, in which the learners enter in a hierarchy with the researchers. Therefore, the following steps should be taken to ensure that ethical safeguards are included in the OLS: Because all the data learners enter into the AI interface is not shown to the AI providers, the researcher provides accounts of the data, and learners are explicitly told what types of data are being collected by the AI interface. Logs are de-identified at source. Preregistration, shared materials, and public analysis code are consistent of open science practice. A deviation from preregistered plan will be reported openly.

Classroom Protocols

A framework is helpful for teachers only if it can be put into practice on Monday morning. The five protocols below do not need any specific software other than access to a general purpose LLM, and are each linked to one or more of the types of friction listed in Table 2. They are suggestions/dispositions whose impact is to be assessed in the program described above and they are provided at this time because the problem to which they refer exists in classrooms.

The Three-Column Log

Learners write all AI suggestions given, their response (accepted, rejected, or modified), and their reason. Grades are based on the quality of the reasoning, which is not the number of accepted texts. The protocol realizes the concept of decision and justification friction and provides the data for the Editorial Agency Index. A teacher could start with a short paragraph task and elicit linguistic reasons for using the word (e.g., a word is too formal for a message to a colleague) and gradually increase the language level desired.

Blind Spot Hunting

Students are presented with an AI-generated text on a topic they are familiar with and asked to identify at least three factual inaccuracies, claims that are not supported, cultural inaccuracies, or tone changes. This is "divergence friction". It flips the learning-machine dynamic from usual, in which the machine is the source of information and the learner is the authority, to one where the output of the AI is an object of analysis and one where the critical reading that the literature on AI in education keeps calling for is cultivated. The problems found can be compared by groups and they can classify them, extending the exercise to corpus style reflection of common machine text problems.

Translate Back and Question

Learners translate an English sentence into Urdu or Punjabi (or back) and note the loss or change in level of politeness, choice of honorific, connotation, idiom or register. This is a localization exercise that brings the differences in pragmatics between languages to the fore, and provides multilingual learners an edge that AI cannot provide. It also provides naturally qualitative evidence related to the flattening of registers that Study 3 aims to quantify systematically, at very little cost.

The Delayed Defence

One week on, learner provides the teacher with a two-minute oral commentary on a text randomly selected from those submitted, explaining both the reasons why this word, this structure or this clause has been used, and what might be different if the text had been read to a different audience. This is delay friction and takes advantage of the spacing and retrieval effects (Bjork & Bjork, 2011; Roediger & Karpicke, 2006). Its purpose is to make ownership obvious, not to question, and it is intentionally brief and non-threatening. For large classes it may be operated with a random subset of students each week or in small groups using peer questioning.

The Voice Preservation Task

Learners receive their own drafts that have been polished by AI, and write their own again with the features they like, such as the rhetorical move, dialect feature, culturally specific metaphor or code-mixing, that they prefer and belong to. They mark-up what they have restored and why. This is localization friction and the real-life manifestation of the third pillar. It challenges students to make distinctions between correctness and voice (Hyland, 2002) and that "more standard" is not synonymous with "better."

Assessing Provenance

Protocols only support evidence when it is rewarded during assessment. A typical marking scheme for a writing task assigns 40% of the marks for the quality of decisions in the log, 30% for the quality of justifications, 20% for a voice or localization element, and 10% for the polished end product, but gives marks in each element for the quality of the product, not the element itself. Weightings should be adapted to context and, ideally, validated (Kane, 2013). Teachers can limit the number of suggestions to be recorded by students and give sentence frames for justifying it for the beginners. Divergent and localization tasks can become

predominant for advanced learners. Throughout, learners should be informed of the purpose of friction, and how it will be evaluated to allow the process to be experienced as support, rather than as surveillance.

Discussion

Implications for Linguistics

The aim of this article is to argue that generative AI does not render linguistics redundant, but rather renders linguistic analysis a pressing need. Each subfield gets a particular task.

The solution of corpus linguistics also is capable of identifying the stylistic profile of machine-produced English (and other languages) and, by means of the update of the models, monitor its changes, looking at the possibility of convergence of human writing. Sociolinguistics may consider the standardization/erasure of who is standardized/erased as a measurable sociolinguistic outcome, and register flattening as a phenomenon. The pragmatic approach, using the audit method, can be used to challenge the models that systematically miss the mark on politeness, honorifics, and implicature, including Brown and Levinson's (1987) theory about politeness. The output and noticing mechanisms can be explored in second language acquisition in the context of offloading patterns and when a machine is used to mediate output. Language assessment can develop validated instruments for a post-fluency world, where the construct of interest is an increasing one of judgement on the output of a machine rather than output unaided. In Pakistan, computational linguistics can develop assessment tools for the regional varieties of the language, which in turn can be utilized as part of the data ecosystem instead of just being a consumer of existing data (Joshi et al., 2020).

Implications for Assessment and Policy

There are three practical implications. The first step in this is to base integrity policies on process rather than detectors. The ethical reason for that change is as compelling as the pedagogical one; documented unreliability and bias of detectors (Liang et al., 2023; Weber-Wulff et al., 2023) make it very clear that the way to do that is not through the detector. Second, there should be support in the institutions for the language rather than just the technical aspects of AI literacy for teachers: teachers should be able to identify and explain register flattening, pragmatic failure and the difference between formal and functional competence. Thirdly, policy and funding bodies should encourage open resources for regional languages, developed by communities. If they are not available, learners in low-resource settings will use tools that will do the least to serve their language(s).

The framework also redefines the main rule that is put on the table every time someone uses AI in a classroom. The framework also restates the most prevalent rule in the classroom when using AI. There's no pedagogically thin way to implement a blanket ban. A blanket permission eliminates the circumstances in which learning is experienced. The way forward, as proposed by the PFF, is a third course – permission with design. Teachers in this course not only define whether AI can be used in a task but where the learner's efforts are required.

Limitations and Future Directions

There are some disclaimers that should be noted. First and foremost, this is a conceptual article, in which constructs, indices and hypotheses are proposals and some of them might not be confirmed empirically. This, in particular, is based on general learning principles that have been established, but have not been demonstrated in the specific case of language tasks with AI assistance. Second, the evidence reviewed on the cognitive costs of AI use is nascent, in part in the form of preprints, and mostly from other domains than language, and from populations different from the target population. Third, the studies proposed are all based on a single institution, English-language tasks in Study 1, and a restricted number of varieties in Study 3, so generalization to other contexts, languages and educational level would require further research. Fourth, LLM changes quickly, so audit results will outline specific models at specific time, and the usefulness of the audit is in its repeatability, not any one result. Fifth, the framework emphasizes writing; it needs to be extended to speaking and listening and multimodal communication, especially given the increasing capabilities of voice-based systems.

Replication across institutions, across languages, and across semesters is suggested for future work; studies of teacher uptake; and the development of validated rubrics for assessing based on language of origin are also suggested. One of the most interesting developments to be explored would be the possibility of adapting the scaffolding and friction to the individual learner based on process data, which is in line with the idea of learner-centred and postmethod pedagogy (Kumaravadivelu, 2001)

Conclusion

Understanding wasn't the same as fluency. This was the most effective proxy that language teachers had, and it was effective because of its expense. The profession can grieve the loss of the proxy or can take this opportunity to educate and test the profession's long-held values – thought, voice, judgement and intent. In this article, it has been suggested that the first step is to correctly describe the problem. The issue is not that students will cheat, it is that fluent text no longer tells us what we need to know. These constructs for measurement, typology for design, the three pillars for policy and practice, and the research programme with multilingual, Pakistani realities at the heart of the research, and not at the edge, are offered by the Productive Friction Framework from that starting point.

Those who ban the machine or those who just bow down before it are not the ones that will be the winners of the new language education future. It will be theirs who will create the friction that keeps us thinking, and the ones that will safeguard the varieties and voices that would otherwise be smoothed. If the machine speaks perfectly, the human's job is to make a meaning.

REFERENCES

- Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., Bali, K., & Sitaram, S. (2023). MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), Article e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Baumgardner, R. J. (Ed.). (1993). *The English language in Pakistan*. Oxford University Press.
- Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587–604. https://doi.org/10.1162/tacl_a_00041
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Bhatt, R. M. (2001). World Englishes. *Annual Review of Anthropology*, 30, 527–550. <https://doi.org/10.1146/annurev.anthro.30.1.527>
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.
- Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>

- Blodgett, S. L., Barocas, S., Daumé, H., III, & Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.485>
- Bourdieu, P. (1991). *Language and symbolic power* (G. Raymond & M. Adamson, Trans.). Harvard University Press.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Canagarajah, S. (2013). *Translingual practice: Global Englishes and cosmopolitan relations*. Routledge.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47. <https://doi.org/10.1093/applin/1.1.1>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671–684. [https://doi.org/10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X)
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39*(2), 175–191. <https://doi.org/10.3758/BF03193146>

- Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32(4), 365–387. <https://doi.org/10.2307/356600>
- García, O., & Li Wei. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan.
- Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Lawrence Erlbaum Associates.
- Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5–24. <https://doi.org/10125/73474>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics*, 34(8), 1091–1112. [https://doi.org/10.1016/S0378-2166\(02\)00035-8](https://doi.org/10.1016/S0378-2166(02)00035-8)
- Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Penguin.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.560>
- Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the Outer Circle. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge University Press.
- Kachru, B. B. (Ed.). (1992). *The other tongue: English across cultures* (2nd ed.). University of Illinois Press.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, 1(1), 1–26. <https://doi.org/10.17239/jowr-2008.01.01.1>

- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2506.08872>
- Kumaravadivelu, B. (2001). Toward a postmethod pedagogy. *TESOL Quarterly*, 35(4), 537–560. <https://doi.org/10.2307/3588427>
- Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757–786. <https://doi.org/10.1002/tesq.194>
- Liang, W., Yuksekogonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), Article 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. <https://doi.org/10.1075/ijcl.15.4.02lu>
- Mahboob, A., & Ahmar, N. H. (2004). Pakistani English: Phonology. In B. Kortmann & E. W. Schneider (Eds.), *A handbook of varieties of English: Vol. 1. Phonology* (pp. 1003–1015). Mouton de Gruyter.
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6), 517–540. <https://doi.org/10.1016/j.tics.2024.01.011>
- Mansoor, S. (2005). *Language planning in higher education: A case study of Pakistan*. Oxford University Press.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education & Macmillan.

- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578. <https://doi.org/10.1093/applin/amp044>
- Ortega, L. (2019). SLA and the study of equitable multilingualism. *The Modern Language Journal*, 103(S1), 23–38. <https://doi.org/10.1111/modl.12525>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askeel, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Rahman, T. (1996). *Language and politics in Pakistan*. Oxford University Press.
- Rahman, T. (2002). *Language, ideology and power: Language-learning among the Muslims of Pakistan and North India*. Oxford University Press.
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
- Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-generated text be reliably detected?* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2303.11156>
- Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158. <https://doi.org/10.1093/applin/11.2.129>
- Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778. <https://doi.org/10.1126/science.1207745>
- Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.
- Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford University Press.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing*, 62, Article 101071. <https://doi.org/10.1016/j.jslw.2023.101071>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, Article 26. <https://doi.org/10.1007/s40979-023-00146-z>
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>