INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY FUTURISTIC DEVELOPMENT

Beyond One Size Fits All Language Instruction: A Diagnostic Framework for Identifying Learner Specific Barriers to Second Language Proficiency Development

Raissa Enyonam Amele Amessouwoe [1*], Kpoka Messanh Isaac [2], Julius Aderogba Odewabi [3]
[1] Wayne State University, Detroit, USA
[2] Independent Researcher, USA
[3] University of Pittsburgh, USA

* Corresponding Author: **Raissa Enyonam Amele Amessouwoe**

Article Info

P-ISSN: 3051-3618
E-ISSN: 3051-3626
Volume: 05
Issue: 02
July - December 2024
Received: 18-09-2024
Accepted: 20-10-2024
Published: 22-11-2024
Page No: 110-123

Abstract
A single proficiency score routinely conceals which component of second language competence is actually limiting a learner's progress. Two learners placed at the same overall band can be held back for entirely different reasons: one carries a stable, rule-governed error pattern into otherwise fluent speech, the other has intact grammar but a narrow phonological repertoire that blocks intelligibility. Instruction pitched at the aggregate score treats both learners identically and helps neither. This paper develops a diagnostic framework for identifying which specific listening, speaking, reading, writing, vocabulary, grammar, pronunciation, fluency, and pragmatic barriers are driving a stalled proficiency trajectory, and for distinguishing a genuine component deficit from a measurement artifact, a motivational barrier, or a contextual constraint. Building on interlanguage theory, error analysis, CEFR-referenced assessment, and individual-differences research on aptitude, working memory, motivation, and anxiety, the framework organizes thirty-six diagnosable constructs into nine domains nested within a five-layer evidence architecture spanning outcome, process, context, implementation, and governance evidence. It specifies falsifiable propositions linking domain-specific diagnostic signals to proficiency growth, a measurement matrix defining how each domain should be operationalized and reported, and a staged pathway for moving diagnostic practice from a single classroom pilot toward validated, transferable use. The framework does not claim an empirically tested intervention: no dataset is analyzed here, and every mechanism is stated as a testable proposition rather than a finding. Its contribution is a shared, falsifiable vocabulary for locating the specific barrier behind a stalled score, so that remediation can be targeted rather than generic.

DOI: https://doi.org/10.54660/IJMFD.2024.5.2.110-123

Keywords: component-specific diagnosis, interlanguage analysis, error analysis, individual differences in language learning, CEFR-referenced assessment, formative assessment, second language proficiency, sconstruct validity

1. Introduction

Second language proficiency is not one thing. A learner's listening comprehension, reading decoding, spoken fluency, pronunciation, grammatical control, vocabulary range, and pragmatic competence can develop at markedly different rates, shaped by first-language background, literacy history, cognitive capacity, motivation, anxiety, and the quantity and quality of interaction the learner actually receives (Dornyei, 2005; Nation, 2001; Ellis, 2008). A composite proficiency score, whether a placement test band or a course grade, compresses this multidimensional profile into a single number. The Common European Framework of Reference separates reception, production, interaction, and mediation precisely because it recognizes that a learner's profile across these modes is rarely uniform (Council of Europe, 2020), yet classroom placement and progress decisions

routinely collapse that profile back into one aggregate figure. Consider two hypothetical learners placed at the same intermediate band on a general proficiency test, offered here for illustration rather than as reported data. The first has a stable, rule-governed interlanguage grammar that produces predictable, low-frequency errors in tense and article use, but near-ceiling listening comprehension and vocabulary breadth; her stalled score reflects a narrow grammatical bottleneck. The second has accurate grammar and wide vocabulary but a restricted phonological inventory that native listeners repeatedly fail to parse, so classroom interaction opportunities dry up and fluency growth stalls for reasons that are only indirectly linguistic. A single score treats these two learners as interchangeable and directs the same generic instruction at both. Neither improves, and the resulting non-improvement is then, circularly, taken as evidence that the learners are simply slow rather than that the diagnosis was too coarse to locate the actual barrier.

The applied linguistics literature is not short of component-specific instruments: diagnostic pronunciation tests (Isbell, 2021), reading diagnostics for beginning learners of Chinese (Yang, 2021), cognitive diagnostic models for grammar (이연숙, 2014), and dynamic assessment protocols that probe the gap between assisted and unassisted performance (Sun, Xu, & Wang, 2022) all exist. What is largely absent is an integrated architecture that specifies, across the full span of component skills and individual-difference variables, what evidence counts as a genuine barrier signal, how that signal should be distinguished from measurement noise or motivational variation, and how a diagnosis should be connected to a decision. This paper addresses that gap directly. It is the first paper in a research program on diagnostic, learner-specific second language instruction, and it is deliberately scoped to the identification problem alone, leaving the design of an adaptive instructional response and the question of institutional scale-up to later stages of that program.

Four questions organize the paper. First, which components of second language proficiency need to be diagnosed separately, rather than folded into a single score, for instruction to be individualized in a meaningful way? Second, what minimum evidence, for each component, distinguishes a genuine barrier from noise, a transient dip, or a measurement artifact? Third, which contextual and individual-difference variables change how a given component signal should be interpreted? Fourth, what evidence would justify treating a diagnostic signal as decision-relevant rather than merely descriptive? These questions structure the critical review, the proposed framework, the construct taxonomy, and the research agenda that follow.

The taxonomy developed in Section 5 groups the thirty-six constructs into nine domains rather than leaving them as an unstructured list, for a reason that matters for how the framework should be read. A list of thirty-six independent diagnostic checks would be impractical for any single classroom to administer and would understate how strongly, for example, working memory demand and processing speed co-vary, or how language anxiety and willingness to communicate jointly shape the same classroom behavior. Grouping related constructs into a domain is itself a testable claim, not a formatting convenience: if two constructs assigned to the same domain turn out to predict proficiency growth through unrelated mechanisms, the grouping should be revised. The propositions in Section 6 are written at the domain level partly so that this grouping claim, and not only the underlying constructs, remains open to empirical challenge.

The review synthesizes peer-reviewed applied linguistics, language testing, and educational psychology scholarship available through 31 December 2024. Because the contribution is conceptual and methodological rather than the report of a completed field study, the paper states hypotheses, sampling and measurement rules, and validation criteria rather than reporting effect sizes, model accuracy, or causal estimates that no dataset here supports; any number that appears below is a specification of what a future study would need to report, not a finding.

Three validity standards discipline the framework throughout. Construct validity requires that a diagnostic measure of, for example, pronunciation intelligibility actually represents intelligibility rather than a proxy such as accentedness that native listeners can rate reliably even when comprehension is unaffected (Isbell & Lee, 2022). Internal validity requires that a proposed barrier, such as limited working memory capacity, be distinguished from competing explanations for the same stalled score, such as low exposure quantity or high language anxiety, before it is treated as the operative mechanism. External validity requires that a diagnostic pattern established in one linguistic, institutional, or first-language context not be assumed to transfer unchanged to another; register and terminology variation across Francophone West African administrative and academic contexts, for instance, can itself act as a confound when diagnostic instruments built for one national curriculum are applied in another (Isaac, 2023, 2024).

The paper's contribution is a construct taxonomy and measurement matrix, not a claim of proven effectiveness. Section 2 synthesizes the theoretical grounds for componential diagnosis in interlanguage theory, error analysis, and individual-differences research. Section 3 proposes a five-layer evidence architecture and situates prior diagnostic instrumentation within it. Section 4 specifies the review and evaluation methodology. Section 5 develops the construct taxonomy across nine diagnostic domains and a measurement matrix for each. Section 6 states falsifiable propositions linking domain signals to proficiency growth. Section 7 lays out a staged pathway from single-classroom pilot to validated, transferable diagnostic practice, and Sections 8 and 9 discuss implications and limitations.

2. Critical Review and Theoretical Foundations

Language proficiency frameworks have long recognized that competence is multidimensional. The CEFR's separation of reception, production, interaction, and mediation gives assessment designers a vocabulary for reporting a learner's profile rather than a single rank (Council of Europe, 2020), and Bachman and Palmer's (2010) treatment of language assessment in practice makes construct definition, not scoring convenience, the starting point for any measure that claims diagnostic value. Alderson's (2005) foundational case for diagnosing foreign language proficiency argues that a genuinely diagnostic test differs from a proficiency test not in its items but in its interpretive claim: a diagnostic instrument must license a specific remedial action, not merely a rank order. Fulcher (2010) reinforces this distinction from a practical testing standpoint, noting that many instruments

marketed as diagnostic in fact report only a finer-grained proficiency score, without the componential interpretation a genuine diagnosis requires. Isbell's (2021) validation of a pronunciation diagnostic is instructive precisely because it treats validation as an argument that specific score patterns support specific pedagogical decisions, a standard the present framework adopts throughout.

The theoretical warrant for componential diagnosis comes from interlanguage theory. Selinker (1972) proposed that a learner's non-target output is not random noise but evidence of a systematic, internally consistent linguistic system in its own right, one that can fossilize in some components while continuing to develop in others. Corder's (1967) earlier argument that learner errors are significant, rather than performance failures to be minimized, is the methodological ancestor of every diagnostic instrument reviewed here: an error pattern is data about the learner's current system, not simply an indicator of how far a score falls below a target. This stance has an important corollary for the componential architecture proposed below: first-language transfer is one plausible source of a given error pattern (Jakobovits, 1969), but Gass's (1989) argument for language universals in acquisition shows that the same surface error can arise from developmental sequences common across learners with different first languages. A defensible diagnostic model therefore treats first-language transfer as one candidate explanation among several, to be tested against alternatives such as developmental universals or instructional artifact (Sang, 1983), rather than as a default attribution.

Diagnostic work on receptive skills illustrates both the promise and the difficulty of componential measurement. Harding, Alderson, and Brunfaut (2015) set out diagnostic principles for reading and listening specifically, arguing that a diagnostic reading or listening test must decompose the skill into subcomponents, such as decoding versus inferencing, that respond to different instruction. Yang (2021) operationalized this logic for beginning Chinese-as-a-second-language readers, constructing an instrument that separates character-decoding difficulty from comprehension difficulty rather than reporting a single reading score. Mislevy and Yin (2009) complicate this decompositional approach by framing language itself as a complex adaptive system, in which component skills interact dynamically rather than summing independently, a caution the present framework addresses by treating component measures as correlated indicators within a layer rather than as fully independent causes. Gaillard and Tremblay's (2016) elicited imitation task offers a convergent cross-check: because it taxes real-time processing across several components simultaneously, agreement between elicited-imitation performance and a componential profile strengthens confidence that a diagnosed barrier is real rather than an artifact of one measurement method.

Productive-skill diagnosis raises a further methodological question: whether self-assessment can substitute for, or must be checked against, external rating. Isbell and Lee (2022) found that Korean-as-a-second-language learners' self-ratings of comprehensibility and accentedness diverged from trained raters in systematic ways, and Teló, Kivistö de Souza, O'Brien, and Carlet (2024) report a similar pattern for self- versus other-assessment of sentence stress, with learners under-detecting their own prosodic errors more than trained listeners did. These findings argue against relying on self-report alone as a diagnostic signal for pronunciation and fluency, even though self-report remains valuable for constructs such as willingness to communicate where the learner is the only accessible source of evidence. On the written side, Lee and Coniam (2013) demonstrate that assessment-for-learning practices in Hong Kong EFL writing classrooms depend on feedback specific enough to be acted on, echoing Johnson's (1992) earlier finding that grammatical competence assessed through written versus auditory materials can yield different pictures of the same learner, a reminder that modality itself is a diagnostic variable, not a neutral delivery channel. Figueiredo, Martins, and Fernandes da Silva (2013) extend this concern to diverse learner populations in Portuguese schools, showing that a diagnostic writing measure calibrated on one learner population can misclassify learners from a different literacy background.

Vocabulary and grammar diagnosis pose a related measurement problem: breadth and depth are not interchangeable. Nation (2001) distinguishes how many words a learner knows from how well they know each one, and In'nami (2013) reviews vocabulary assessment research showing that instruments optimized for breadth routinely under-diagnose depth-related barriers such as collocational restriction. Koizumi (2015) synthesizes validity evidence for vocabulary assessment and argues that a genuinely diagnostic vocabulary instrument must report both dimensions separately rather than a single lexical score. On the grammar side, 이연숙 (2014) applies cognitive diagnostic modeling to isolate which specific grammatical subskills a learner has and has not mastered, an approach that converges with Spinner's (2011) finding that morphosyntactic development does not proceed uniformly across grammatical subsystems, so a single grammar score can average over subskills moving in opposite directions. Neumann (2014) shows that teachers' own assessments of learners' grammatical ability in academic writing are prone to systematic bias, which argues for anchoring grammatical diagnosis in structured instruments rather than impressionistic teacher judgment alone. Pragmatic competence adds a further layer that a purely grammatical or lexical profile cannot capture; Roever (2007) demonstrates that differential item functioning in pragmatics assessment can make a pragmatic barrier look like a proficiency-level effect unless the instrument is checked for fairness across learner subgroups, a concern taken up directly in the equity domain of the taxonomy below.

A purely linguistic account of component barriers is incomplete without individual-differences research. Dornyei's (2005) synthesis of the psychology of the language learner treats aptitude, motivation, and anxiety as systematically related to, rather than incidental to, proficiency development, a position with roots in Gardner's (1985) socio-educational model linking integrative motivation and attitudes to second language attainment. Nida's (1956) and Spolsky's (1969) earlier work on motivation and attitudinal variables in language learning established that affective variables are not simply noise around a cognitive signal but can themselves constitute the operative barrier, a point the taxonomy below preserves by treating affective constructs as diagnosable in their own right rather than as covariates to be adjusted away. DeKeyser (2019) frames aptitude not as a fixed trait but as something that interacts with instructional treatment, so the same aptitude profile can be a barrier under one instructional design and an asset under another, which is the aptitude-

treatment interaction logic that motivates individualized instruction. Processing-level evidence reinforces this picture: Wakabayashi (2009) and Pulido (2003) both show that lexical learning and vocabulary acquisition through reading are shaped by processing capacity and topic familiarity respectively, and Shin and Christianson (2011) demonstrate that structural priming effects in second language learners depend on working memory in ways that differ from first language processing, supporting the inclusion of working memory demand and processing speed as separate diagnosable constructs rather than a single generic cognitive-load indicator.

Taken together, this literature establishes that the individual components of second language proficiency, and the individual-difference variables that moderate their development, are each diagnosable with existing instruments and theory. What the literature does not provide is an integrated architecture that specifies, across all of these components and moderators at once, what evidence is required before a signal counts as a genuine barrier, how competing explanations are ruled out, and how a diagnosis connects to an instructional decision. Ellis's (2008) broad synthesis of second language acquisition research and Oxford's (2017) treatment of learning strategies both gesture toward this kind of integration without formalizing it, and formative-assessment scholarship (Black & Wiliam, 1998; Hattie & Timperley, 2007; Ketabi & Ketabi, 2014) establishes the general principle that evidence must change teaching to count as diagnostic, without specifying which second-language-specific components that evidence should cover. The framework proposed in the next section closes this gap.

A further theoretical consideration shapes how the taxonomy should be interpreted once it is applied: componential diagnosis is descriptive of where a barrier currently sits, not a judgment about a learner's capacity to overcome it. Bachman and Palmer's (2010) construct-first approach to assessment and Fulcher's (2010) practical testing framework both warn against letting a diagnostic label drift from a statement about current performance into a claim about fixed ability, a drift that individual-differences research on aptitude makes especially easy given how readily aptitude is popularly, if inaccurately, treated as immutable (DeKeyser, 2019). The taxonomy below is written throughout in the language of current barriers rather than learner deficits for this reason, and the governance layer specified in Section 3 exists partly to keep that distinction operative once a diagnosis leaves the page and enters a classroom.

3. Proposed Framework

The framework organizes learner-specific barrier diagnosis across five linked evidence layers, each answering a different question about a stalled proficiency trajectory: what changed (outcome), how it changed (process), under what conditions (context), how faithfully the diagnostic protocol was actually delivered (implementation), and who is accountable for acting on the result (governance). The five layers are ordered as a diagnostic sequence rather than a checklist: a credible barrier diagnosis moves from outcome evidence to process evidence before it is allowed to generate an implementation-layer or governance-layer decision.

The outcome layer records proficiency growth, or its absence, on a primary measure with a defined denominator, observation window, and direction, plus a limited set of secondary measures that reveal tradeoffs a single score would hide, such as gains in reading comprehension purchased at the cost of stalled pronunciation practice time. Composite scores that combine components should publish their weights and show how sensitive the overall picture is to reasonable alternative weightings, since an unweighted average can mask the very componential pattern the framework exists to surface.

The process layer is where componential diagnosis actually happens. It records, for each of the nine diagnostic domains developed in Section 5, a time-stamped measurement with a stated unit, source, and observation window, so a barrier can be traced to a specific construct rather than inferred from movement in the aggregate score alone. A process indicator earns a place in the framework only when it has a plausible mechanism connecting it to proficiency growth and when it supports a specific instructional action; indicators merely correlated with the outcome but suggesting no action are recorded for research purposes but excluded from the operational diagnostic protocol.

The context layer prevents a diagnostic pattern established in one setting from being applied unchanged in another. Relevant context includes the learner's first language and literacy background, the institutional and curricular setting, the linguistic and register variation of the target language across regions and registers, such as the terminology differences documented across Francophone West African administrative and educational registers (Isaac, 2023, 2024), and the instructional resources actually available. Interaction terms or stratified reporting should test whether a given component barrier predicts proficiency growth similarly across these contexts before the pattern is treated as general.

The implementation layer distinguishes a genuine absence of a barrier from a diagnostic protocol that was never properly delivered. It records whether the intended diagnostic instruments were actually administered (reach), whether teachers used the resulting profile in planning (adoption), whether the assessment was administered under the conditions its validation assumed (fidelity), what local adaptations were made and why, and what time and resource burden the protocol placed on teachers and learners. A null diagnostic finding recorded under low fidelity is evidence about implementation, not evidence that the underlying barrier does not exist.

The governance layer assigns who owns diagnostic data, who reviews a borderline or contested classification, how a learner or teacher can contest an inaccurate diagnostic label, and when the diagnostic protocol itself should be paused for revision. Because a diagnostic label such as a working-memory-related barrier can affect how a learner is grouped or advised, governance must protect against a diagnosis hardening into a fixed trait attribution; version control should link every classification to the evidence and protocol version in force when it was made, so a revised understanding of a learner's profile can supersede an earlier one without erasing the record.

The five layers form a loop rather than a one-time report. Baseline process evidence proposes an initial diagnostic

profile; further process evidence over time either confirms or revises it; outcome evidence tests whether the profile actually predicted where growth would and would not occur; implementation evidence explains discrepancies between the predicted and observed pattern; and governance decides whether to retain, revise, or retire a given diagnostic construct. Cases in which a diagnosed barrier failed to predict subsequent growth must be preserved in this loop, not discarded, since a taxonomy validated only on its successes will look more accurate than it is.

Table 1: Five-layer evidence architecture for barrier diagnosis

Layer	Diagnostic question it answers	Minimum evidence	Decision it supports
Outcome	What changed in the learner's proficiency profile	Primary and secondary component-level growth measures with denominator and window	Judge whether a diagnosed barrier is associated with a proficiency gap
Process	Which component produced the change	Time-stamped, domain-specific diagnostic signals (Section 5)	Trace a stalled score to a specific construct
Context	Under what conditions the pattern holds	First-language, literacy, institutional, and register variables	Decide whether a diagnostic pattern generalizes beyond the setting where it was found
Implementation	Was the diagnostic protocol actually delivered as intended	Reach, fidelity, adaptation, and burden records	Distinguish a true null finding from a delivery failure
Governance	Who is accountable for the diagnostic decision	Ownership, contestation, and version records	Authorize, revise, or retire a diagnostic classification

3.1. Evidence Synthesis by Diagnostic Instrumentation

The instruments already used to measure individual components fall into five method families, each with different strengths for barrier diagnosis specifically, as distinct from proficiency ranking generally.

CEFR-linked instruments (Council of Europe, 2020) and practical guidance for using them diagnostically (Xu & Li, 2024) give component diagnosis a common proficiency-level vocabulary, but Thomas (1994) cautions that band descriptors built for research on second language acquisition were never designed to license fine-grained diagnostic claims, and applying them as if they were invites over-interpretation. Poole (1992) shows that classroom language socialization patterns interact with how learners are positioned by level descriptors, meaning a CEFR band can partly reflect classroom social dynamics rather than underlying competence alone. Norm-referenced instruments therefore anchor the taxonomy's terminology without being treated as diagnostic in themselves.

Cognitive diagnostic modeling (이연숙, 2014) and dynamic assessment, which probes the gap between a learner's unassisted and mediated performance (Sun, Xu, & Wang, 2022), are the two method families closest to the present framework's logic, because both are built to license a specific instructional inference rather than a rank. Ghanbarpour's (2017) meta-synthesis of dynamic assessment research in second and foreign language learning finds consistent evidence that mediated-performance gaps predict subsequent independent gains better than static test scores alone, though Mehri and Amerian (2015) catalogue practical obstacles, including the time cost of one-on-one mediation and inconsistent mediator training, that limit dynamic assessment's use at scale. This tension between diagnostic precision and implementation cost recurs throughout the taxonomy below and is addressed directly in the implementation layer.

Computer-adaptive testing promises efficient, component-specific measurement (Dunkel, 1999), and adaptive e-learning models for English as a second or foreign language operationalize this promise for classroom use (Kostolanyova & Simonova, 2020a, 2020b). Hashim, Salam, Nurul, and Syafiatun (2018) extend adaptive self-assessment into massive open online course environments, and Menaka and Sankar (2019) report broader technology-mediated assessment practice for second language learners, but Norris (2001) raises a durable methodological concern that remains relevant to newer systems: computerized oral proficiency scoring can penalize legitimate variation in speaking style that has nothing to do with proficiency, so automated diagnostic signals require the same validity scrutiny as any other instrument. Li's (2024) more recent examination of large language model feedback for second language writing extends this concern to generative systems, finding that automated feedback quality varies by error type in ways a diagnostic protocol must anticipate rather than assume away.

Self-assessment contributes usefully to constructs, such as willingness to communicate and self-efficacy, where the learner is the most direct source of evidence (Butler, 2023, 2024), but its reliability for externally observable constructs such as pronunciation and writing accuracy is more limited, as the divergence between self- and other-ratings documented earlier in this review (Isbell & Lee, 2022; Telós., 2024) illustrates. Rafi and Pourdana (2023) combine collaborative and individual e-diagnostic oral tasks within a differentiated instruction framework and find that peer-mediated tasks surface barriers that neither pure self-assessment nor a single individual performance sample reveals alone, supporting the taxonomy's decision to treat interactional competence and willingness to communicate as related but distinct constructs rather than a single social-affective composite.

Writing diagnosis has its own instrumentation history. Safari and Ahmadi (2023) develop and validate a diagnostic checklist for integrated second language writing that separates accuracy, organization, and source-use errors rather than reporting a holistic score, and Zhang (2013) argues on conceptual grounds that writing functions simultaneously as a target skill and as a vehicle for learning, which is why writing diagnostics must distinguish barriers specific to writing from barriers, such as limited grammatical control, that also show up elsewhere in the taxonomy. Oh (2019) shows that the writing resources learners draw on, including translation and first-language planning, shape what a writing sample can validly reveal about underlying proficiency, and

Lanteigne (2017) proposes sentence-reordering tasks as an authentic alternative that reduces the confound between writing-specific mechanics and broader grammatical knowledge. Macaro (2003) and Brown (2013) both argue, from different angles, that classroom teachers are well placed to generate and interpret this kind of fine-grained evidence themselves, provided the instruments they use are built for diagnosis rather than improvised from ranking tools, a design requirement the measurement matrix in Section 5 makes explicit.

A sixth family addresses equity directly rather than as an afterthought to the other five. Badrun Kholid, Rahman, and Irawan (2024) implement diagnostic assessment for differentiated instruction at the junior secondary level and report that the same instrument produced markedly different classification patterns across classes with different resourcing, a finding that argues for local calibration rather than assuming an instrument validated in one setting transfers unchanged to another. Heriansyah, Darni, Fajri, and Sahardin (2021) reach a related conclusion from competency-based curriculum research, and Mohamadi (2017) shows that task engagement itself, not only linguistic knowledge, is a legitimate object of assessment when the goal is diagnosis rather than ranking. Taken together, this family argues that instrumentation choices are never purely technical; they carry assumptions about which learners an instrument was built to describe accurately, assumptions the taxonomy's fairness and accommodation domain in Section 5.9 makes explicit rather than leaving implicit.

4. Research Design and Evaluation

The review searched Linguistics and Language Behavior Abstracts, ERIC, Scopus, and Google Scholar for English- and French-language peer-reviewed work on second language diagnostic assessment, component-specific proficiency measurement, and individual differences in language learning, published through 31 December 2024. Search terms combined each of the nine diagnostic domains developed in Section 5 with terms for diagnosis, assessment, and individual differences; citation chaining from key methodological sources (Alderson, 2005; Bachman & Palmer, 2010; Harding, Alderson, & Brunfaut, 2015) identified additional foundational and regional work.

Appraisal distinguished instrument-validation studies, which report evidence for a specific diagnostic claim, from broader proficiency-assessment studies that report ranking evidence without a diagnostic interpretation; only the former were treated as direct evidence for the taxonomy's measurement matrix, while the latter informed the theoretical synthesis in Section 2. Authority and citation count were not treated as substitutes for this distinction, since several of the most influential broad-proficiency instruments in the field make no diagnostic claim at all.

Construct definitions in Section 5 were developed by extracting, from each validated instrument identified in the review, the specific claim it licenses about a learner's profile, then grouping constructs that share a diagnostic mechanism into a common domain. Each domain retains an operational definition, a stated unit and observation window, and at least one representative instrument or protocol from the reviewed literature, so the taxonomy can be audited against its sources rather than treated as a free-standing list.

Future validation of the taxonomy requires preregistered, multi-site studies that hold out at least one context, a cohort, an institution, or a language pairing, not used in taxonomy development, report calibration and subgroup performance rather than an aggregate accuracy figure alone, and retain null findings where a hypothesized domain fails to predict subsequent growth. A domain that does not survive this kind of external check should be revised or removed from the taxonomy rather than retained on the strength of face validity alone.

Because diagnostic classification can influence how a learner is grouped, advised, or perceived by teachers, ethical review of any future empirical application of this taxonomy must examine the risk that a diagnostic label, once assigned, becomes a self-fulfilling constraint on the opportunities a learner is offered, independent of whether the original diagnosis remains accurate. Institutional review processes designed for intervention studies do not always anticipate a conceptual and taxonomy-development paper of this kind; where a future study collects even pilot data to test the propositions in Section 6, that study's own protocol, not this paper, is the appropriate object of ethics review, and this paper's contribution is to specify in advance what such a protocol would need to include.

A methodological limitation of the review itself should be stated directly. Restricting the search to English- and French-language, peer-reviewed sources through 31 December 2024 likely underrepresents diagnostic assessment scholarship published in other languages or in practitioner-oriented outlets that do not index in the databases searched, and the taxonomy's nine domains should be read as a synthesis of what this literature currently supports rather than as an exhaustive account of every diagnosable construct a future review might surface.

Reporting standards apply once a domain is measured. Every reported domain signal should state its unit, observation window, source instrument, and missingness rate; composite domain scores should disclose the weight given to each constituent construct; and any classification derived from a domain signal should be accompanied by the confidence or uncertainty associated with it rather than presented as a categorical fact. These requirements follow directly from the construct-validity and calibration criteria in Table 2 and are intended to make a future study's report auditable against the taxonomy it claims to test, rather than merely consistent with it in spirit.

Table 2: Evaluation criteria for the diagnostic taxonomy

Criterion	What it requires of a domain measure	Diagnostic interpretation
Construct validity	Content, convergent, and discriminant evidence that the measure represents the named construct and not a neighboring one	Does a low score on this domain reflect the construct it claims to, rather than a different barrier
Reliability	Interrater, internal consistency, or test-retest evidence appropriate to the measure	Is the signal stable enough to support a classroom decision
Convergent and discriminant evidence	Agreement with independent measures of the same construct and divergence from measures of different constructs	Does the domain add diagnostic information beyond neighboring domains
Calibration	Agreement between predicted and observed growth for learners classified into a given domain barrier	Do classifications correspond to real growth patterns, not just internally consistent scores
Transfer	Performance of the taxonomy in a held-out cohort, institution, or first-language group	Does a domain diagnosis generalize beyond its development setting
Fairness	Subgroup comparison of classification rates and error patterns, including differential item functioning checks (Roever, 2007)	Does the instrument classify learners from different backgrounds consistently
Decision value	Evidence that a domain diagnosis changed an instructional decision and that decision improved an outcome	Did the diagnosis do more than describe

5. Diagnostic Construct Taxonomy and Measurement Matrix

The taxonomy groups the thirty-six diagnosable constructs identified in this review into nine domains. Each domain below states what distinguishes its constituent constructs diagnostically, summarizes representative instrumentation, and specifies the minimum evidence, in unit, observation window, and threshold logic, that the measurement matrix in Table 3 requires before a domain-level signal is treated as a genuine barrier rather than normal variation.

5.1. Receptive Processing

Receptive processing separates the ability to perceive and decode linguistic input from the ability to construct meaning from it. Listening discrimination and reading decoding are lower-level constructs, measurable through minimal-pair identification tasks and timed word- or character-recognition tasks respectively (Yang, 2021), while listening comprehension and reading comprehension require integrating decoded input with prior knowledge and context (Harding, Alderson, & Brunfaut, 2015). Diagnostically, this separation matters because a learner can fail a comprehension task for a decoding-level reason, a comprehension-level reason, or both, and only a two-tier measurement design can tell these apart; a single comprehension score cannot. The measurement matrix therefore requires a decoding-level or discrimination-level measure alongside any comprehension measure in this domain, with both administered at baseline and re-administered when instructional exposure plausibly changes one but not the other.

5.2. Oral Production

Oral production constructs are diagnostically distinct because a learner can be strong on one and weak on another in ways that a single speaking score averages away. Spoken fluency, typically measured through speech-rate and pausing analysis, and pronunciation intelligibility, measured through listener transcription or comprehensibility rating rather than accentedness alone (Isbell & Lee, 2022; Waniek-Klimczak, 2018), are related but separable; a learner can speak fluently with low intelligibility or haltingly with high intelligibility. Interactional competence, the ability to manage turn-taking, repair, and co-construction of meaning in real exchange, and pragmatic competence, the ability to use language appropriately for context and social relationship, add a further layer that individual-task speaking measures cannot capture; Roever's (2007) fairness concerns are especially relevant here because pragmatic norms vary by culture and register in ways that can be mistaken for a proficiency deficit. The matrix requires that pronunciation be measured through intelligibility rather than accent rating, and that interactional and pragmatic competence be assessed through interactive rather than monologic tasks.

5.3. Written Production

Written production constructs separate the sentence-level accuracy of a text from its discourse-level organization, and both from the learner's capacity to act on feedback and the literacy background the learner brings to the task. Safari and Ahmadi's (2023) diagnostic writing checklist operationalizes the accuracy-organization distinction directly, while feedback uptake, whether a learner incorporates specific written or oral feedback into a subsequent draft, requires longitudinal, draft-to-draft comparison rather than a single-timepoint score (Lee & Coniam, 2013; Hattie & Timperley, 2007). Literacy background, including prior schooling in the first language and exposure to the target script, is included as a diagnosable moderator because Figueiredo, Martins, and Fernandes da Silva (2013) show that writing instruments calibrated on learners with strong first-language literacy misclassify learners whose literacy gap is not primarily linguistic. The matrix requires that writing accuracy and organization be scored separately using a shared rubric across at least two drafts, so feedback uptake can be computed rather than assumed.

5.4. Lexicogrammatical Systems

Vocabulary breadth and depth are treated as separate constructs because an instrument optimized for one under-diagnoses the other (Nation, 2001; In'nami, 2013; Koizumi, 2015), and grammatical control is treated as multidimensional rather than unitary because morphosyntactic subskills develop at different rates within the same learner (Spinner, 2011). First language transfer is included in this domain, rather than treated as a background variable, because lexical and grammatical error patterns are where transfer effects are most directly observable and most easily confused with developmental error (Jakobovits, 1969; Gass, 1989); the matrix requires that a transfer hypothesis be tested against a developmental-sequence alternative before it is recorded as the operative explanation for a given error pattern, following the interlanguage logic set out in Section

2. Ströbel, Kerz, and Wiechmann's (2020) finding that first-language writing complexity predicts second-language writing complexity in advanced learners illustrates why transfer should be measured directly rather than inferred solely from error type.

5.5. Cognitive Processing Capacity

This domain covers the cognitive resources a learner brings to a language task, as distinct from the linguistic knowledge measured elsewhere in the taxonomy. Strategic competence, the ability to compensate for a linguistic gap using circumlocution or other strategies (Oxford, 2017), working memory demand, and processing speed jointly determine how much of a learner's linguistic knowledge is actually deployable under real-time task conditions; Shin and Christianson (2011) show that structural priming, a marker of real-time processing efficiency, depends on working memory capacity in second language learners in ways that differ from first language processing patterns. Metacognitive control, the learner's ability to monitor and regulate their own language use, is included because Koda (2007) demonstrates that crosslinguistic differences in reading development are partly explained by differences in metacognitive monitoring strategies rather than decoding ability alone. Because these constructs are less directly observable than linguistic performance, the matrix specifies indirect but validated measures, such as dual-task and self-paced reading paradigms, rather than relying on teacher impression.

5.6. Affective and Motivational Profile

Affective and motivational constructs are diagnosable barriers in their own right, not adjustments applied after a cognitive or linguistic diagnosis. Language anxiety and willingness to communicate jointly determine how much of a learner's underlying competence is expressed under classroom conditions, and self-efficacy and motivation quality determine how a learner responds to a diagnostic finding once it is shared, which has direct implications for how diagnostic feedback should be delivered. Nida (1956) and Spolsky (1969) established the foundational case that these variables independently predict language learning outcomes, and Gardner's (1985) socio-educational model formalizes motivation quality as multidimensional rather than a single high-low scale, distinguishing integrative from instrumental orientations that predict different patterns of persistence. Foster-Cohen (2004) cautions, from a relevance-theoretic perspective, against treating affective variables as simply reducible to linguistic behavior, reinforcing the taxonomy's decision to measure this domain with validated affective instruments rather than inferring it from linguistic performance alone.

5.7. Exposure and Opportunity

This domain records how much genuine opportunity a learner has had to develop each of the other domains, because a component barrier that is actually an exposure barrier requires a different response than a barrier rooted in aptitude or instruction quality. Exposure quantity and interaction quality are distinguished because de Bot (2014) shows that early and extensive but low-quality exposure does not reliably outperform later but higher-quality, interactionally rich exposure, and Navracsics (2014) similarly finds that the intimacy and communicative demand of input, not raw quantity alone, predicts outcomes in bilingual input research. Attendance and digital access are included as structural preconditions for exposure that are often available from institutional records and therefore inexpensive to measure relative to their diagnostic value; Forgor's (2021) study of assessment-for-learning and adaptive learning systems found that unequal digital access alone explained a meaningful share of the proficiency gap the study otherwise attributed to instructional design. The matrix treats this domain as a moderator layer: a component deficit co-occurring with low exposure should be interpreted as an opportunity gap before it is treated as an aptitude or instructional gap.

5.8. Social and Instructional Context

This domain captures how the immediate instructional environment shapes whether a learner's underlying competence is visible in assessment at all. Teacher expectancy effects, task familiarity, the relevance of instructional goals to a learner's own purposes, and learner agency, the degree to which a learner can influence their own learning pathway, jointly determine engagement independent of ability. Poole (1992) shows that classroom language socialization patterns can systematically position some learners as more or less competent regardless of measured performance, and Heriansyah, Darni, Fajri, and Sahardin (2021) argue that a competency-based curriculum's own assessment design shapes which learners appear to struggle, a caution against treating this domain's constructs as purely learner-internal. Caplan's (2019) argument that language is learned through language, not only about language, supports treating goal relevance and learner agency as diagnosable rather than as background classroom-management variables, since a learner disengaged because instructional content is irrelevant to their goals can present with the same surface profile as a learner facing a genuine cognitive or linguistic barrier.

5.9. Fairness and Accommodation

The final domain audits the taxonomy's own instruments rather than the learner. Assessment fairness and differential item functioning checks (Roever, 2007) test whether a given diagnostic instrument classifies learners from different first-language, cultural, or socioeconomic backgrounds consistently, and disability accommodation records whether accommodations were in place and adequate when a diagnostic measure was administered, since an unaccommodated disability can present as a component-specific barrier that is actually a testing-condition artifact. Retention and transfer, whether a diagnosed gain is durable and generalizes beyond the task on which it was measured, closes the loop back to the outcome layer in Section 3, since Pourdana's (2022) work on computer-assisted diagnostic assessment and collaborative writing found that short-term gains from a targeted intervention did not always predict durable improvement, underscoring why retention must be measured directly rather than assumed from immediate post-test performance. Badrun Kholid, Rahman, and Irawan (2024) and Yeotak Yoon, 김정우, and Min (2008) both report classroom implementations of diagnostic assessment for differentiated instruction at the secondary level that depended on exactly this kind of fairness and retention checking to avoid entrenching an early misclassification.

The nine domains are not independent in practice, and the taxonomy does not treat them as such. Receptive processing

and oral production constructs interact through interactional competence, which draws on both; cognitive processing capacity constrains how much of a learner's lexicogrammatical knowledge becomes available for real-time use; and the exposure, social-context, and fairness domains chiefly condition how reliably the other six domains can be measured at all rather than measuring a further linguistic skill of their own. This layered structure, linguistic-productive domains resting on cognitive and affective capacity, which in turn rests on exposure, context, and fairness, is itself a testable claim, formalized as Proposition 10 in Section 6, and it is the reason the measurement matrix in Table 3 is read as an ordered diagnostic sequence rather than as thirty-six unrelated checkboxes to be completed in any order.

Table 3: Diagnostic construct taxonomy: nine domains, thirty-six constructs

Domain	Constituent constructs	Representative instrument or protocol	Distinguishing diagnostic signal
Receptive processing	Listening discrimination, listening comprehension, reading decoding, reading comprehension	Minimal-pair and timed decoding tasks paired with comprehension tasks (Yang, 2021; Harding *et al*., 2015)	Comprehension failure with intact decoding versus comprehension failure with decoding failure
Oral production	Spoken fluency, pronunciation intelligibility, interactional competence, pragmatic competence	Listener transcription and comprehensibility rating, interactive task rating (Isbell & Lee, 2022)	Fluent but unintelligible speech versus intelligible but disfluent speech
Written production	Writing accuracy, writing organization, feedback uptake, literacy background	Shared-rubric multi-draft scoring (Safari & Ahmadi, 2023)	Sentence-level error versus discourse-level organization failure
Lexicogrammatical systems	Vocabulary breadth, vocabulary depth, grammatical control, first language transfer	Breadth and depth vocabulary instruments, cognitive diagnostic grammar model (이연숙, 2014)	Word-knowledge gap versus grammatical subskill gap versus transfer-pattern error
Cognitive processing capacity	Strategic competence, working memory demand, processing speed, metacognitive control	Dual-task and self-paced processing paradigms (Shin & Christianson, 2011)	Knowledge present but not deployable under real-time conditions
Affective and motivational profile	Language anxiety, willingness to communicate, self-efficacy, motivation quality	Validated affective and motivational inventories (Gardner, 1985)	Competence masked by anxiety or low willingness to communicate
Exposure and opportunity	Exposure quantity, interaction quality, attendance, digital access	Institutional exposure and access records (Forgor, 2021)	Component deficit co-occurring with an opportunity gap
Social and instructional context	Teacher expectancy, task familiarity, goal relevance, learner agency	Classroom observation and engagement measures (Poole, 1992)	Disengagement profile resembling a cognitive or linguistic barrier
Fairness and accommodation	Assessment fairness, differential item functioning, disability accommodation, retention and transfer	Subgroup DIF analysis, durability follow-up (Roever, 2007; Pourdana, 2022)	Instrument artifact versus genuine, durable barrier

6. Testable Propositions

Proposition 1: A learner whose comprehension score is low but whose decoding or discrimination score is at or near ceiling will show larger proficiency gains from comprehension-focused instruction, such as inferencing and schema-building tasks, than from further decoding practice; a learner with the reverse pattern will show the opposite. A valid test should preregister the decoding-comprehension split, counterbalance which instructional focus each profile receives, hold out at least one cohort, and report the interaction rather than a main effect alone. A null interaction would indicate that the decoding-comprehension split does not carry the instructional relevance this proposition assumes, and the taxonomy's separation of these constructs should be revisited rather than defended.

Proposition 2: Pronunciation intelligibility, measured through listener comprehensibility rating, will predict subsequent classroom interaction opportunity more strongly than accentedness rating does, because intelligibility, not accent, is what limits a listener's willingness to sustain an exchange. A valid test should collect both rating types on the same speech sample, track subsequent interaction opportunity independently, and report both correlations with confidence intervals rather than only the stronger one. If accentedness predicts interaction opportunity as well as or better than intelligibility, the framework's emphasis on intelligibility over accent should be reconsidered rather than retained by default.

Proposition 3: Feedback uptake, measured as the proportion of specific feedback points incorporated in a learner's next draft, will predict subsequent writing accuracy gains better than baseline accuracy score alone. A valid test should code feedback specificity, track uptake across at least three drafts, and report whether uptake adds predictive value beyond baseline score after adjustment for literacy background. A finding that baseline score alone predicts gains as well as uptake would suggest that feedback processes matter less than this proposition assumes, at least for the writing tasks and learner population tested.

Proposition 4: An error pattern attributable to first-language transfer will respond differently to explicit contrastive instruction than a developmentally universal error pattern responds to the same instruction, with the transfer-attributed pattern showing a larger short-term gain. A valid test should classify error patterns using both a transfer hypothesis and a developmental-universals hypothesis before treatment assignment, preregister the predicted interaction, and report subgroup results by classification rather than pooling all grammatical errors together. Failure to find this interaction would argue against treating transfer and developmental-universal errors as instructionally distinct, contrary to the theoretical case made in Section 2.

Proposition 5: Working memory demand, measured independently of linguistic knowledge, will moderate the

effect of processing-speed-intensive tasks, such as timed oral response, on measured fluency, such that learners with lower working memory capacity show smaller fluency gains from time-pressured practice than learners with higher capacity, even when both groups have equivalent underlying grammatical and lexical knowledge. A valid test should measure working memory using a validated span task independent of the target language, preregister the interaction, and report results separately by capacity band. A null interaction would suggest that time-pressured fluency practice can be prescribed without regard to working memory capacity, contrary to the aptitude-treatment interaction logic motivating this domain.

Proposition 6: Reducing measured language anxiety through a structured intervention will increase the gap between a learner's classroom-observed performance and their performance on a low-stakes diagnostic task, because anxiety selectively suppresses classroom performance rather than underlying competence. A valid test should measure anxiety, classroom performance, and low-stakes diagnostic performance before and after the intervention, preregister the predicted gap reduction, and report results for learners with high baseline anxiety separately from those with low baseline anxiety, since the proposition predicts no effect for the latter group. Failure to find a differential effect by baseline anxiety would undercut the case for anxiety-focused intervention as a distinct instructional response.

Proposition 7: Among learners with equivalent scores on the domains in Sections 5.1 through 5.6, those with lower exposure quantity and interaction quality will show larger subsequent gains from increased structured interaction opportunity than from further direct instruction in the deficient domain. A valid test should stratify learners by exposure and interaction quality independent of their component profile, counterbalance the instructional response, and report the interaction rather than assuming it. If direct instruction outperforms increased interaction opportunity regardless of baseline exposure, the opportunity-gap interpretation this domain assumes should be revised.

Proposition 8: Task familiarity will interact with domain-specific performance such that unfamiliar task formats produce systematically lower scores than familiar formats for the same underlying competence, and this gap will be larger for learners with shorter institutional tenure. A valid test should administer the same construct through both familiar and unfamiliar task formats to the same learners, preregister the tenure interaction, and report results by tenure band. A null tenure interaction would suggest that task-familiarity effects are uniform across learners and do not require the tenure-sensitive interpretation this domain currently assumes.

Proposition 9: A diagnostic instrument that shows differential item functioning across first-language subgroups will produce classification rates for a given domain that differ across those subgroups even after controlling for independently measured proficiency, and revising the flagged items will reduce that gap without reducing overall classification accuracy for the majority subgroup. A valid test should conduct DIF analysis before deployment, preregister the subgroup comparison, and report classification-rate differences before and after item revision. Failure to reduce the subgroup gap after revision would indicate that the flagged items were not the source of the disparity, and a different fairness mechanism should be investigated.

Proposition 10: The nine domains proposed in Section 5 will predict proficiency growth better jointly than any single domain predicts alone, but the marginal contribution of each additional domain will diminish once receptive processing, oral production, and written production are already included, because affective, cognitive, and contextual domains chiefly operate by moderating how much of a learner's linguistic knowledge is expressed rather than by adding independent linguistic knowledge of their own. A valid test should fit nested models that add domains in a preregistered order, report the marginal predictive gain at each step, and test whether affective and contextual domains show larger marginal gains for learners with low exposure quantity than for learners with high exposure quantity. A finding that every domain contributes an equal, non-diminishing marginal gain would argue against the layered interpretation this proposition assumes and would suggest the nine domains are more independent of one another than the clustering rationale in Section 1 proposes. Because this proposition concerns the taxonomy's own architecture rather than any single construct, it should be tested last, after the domain-specific propositions above have themselves been evaluated on their own evidence.

7. Implementation Pathway

7.1. Definition

The definition phase requires a school or program to select, from the nine domains in Section 5, the subset relevant to its learner population and instructional goals, and to agree in writing on the primary outcome measure, the accountable role for each diagnostic instrument, and the evidence threshold that will count as a genuine barrier rather than normal variation. Teams should record explicitly which domains they are deliberately excluding at this stage and why, since a program that claims to diagnose all thirty-six constructs from the outset is more likely to implement none of them with fidelity. This phase produces a written protocol, not yet any learner data.

7.2. Readiness

Readiness assessment checks whether the instruments selected in the definition phase can actually be administered with the staff time, training, and data infrastructure the program has, and whether the ethical approvals appropriate to the setting are in place before any learner is assessed. A program that lacks a secure way to record and correct diagnostic classifications should address that infrastructure gap before piloting, since a diagnostic label recorded incorrectly and left uncorrected causes more harm than no diagnosis at all.

7.3. Pilot

The pilot phase administers the selected instruments to a single class or cohort, without yet using the results for high-stakes grouping or advising decisions, so instrument reliability and burden can be assessed under real classroom conditions before any decision depends on the results. Pilot data should be used to test, at small scale, whether the domain separations proposed in Section 5, such as decoding versus comprehension or breadth versus depth, actually behave as separable in this specific learner population, since the taxonomy was built from diverse published instruments that may not all transfer cleanly to a new setting.

7.4. Protected Evaluation

Protected evaluation holds out at least one class, cohort, or time period not used to develop the pilot's instructional response, and evaluates whether the diagnosed domain pattern predicts subsequent growth in the held-out group with preregistered analysis. This phase is where the propositions in Section 6 are formally tested, and the evaluation should report calibration and subgroup performance, not only an aggregate accuracy figure, before the program proceeds to advisory use.

7.5. Advisory Use

Advisory use permits teachers to see diagnostic profiles alongside, not instead of, their own classroom judgment, with an explicit statement of what the diagnostic evidence does and does not support, and a documented process for a teacher or learner to contest a classification. At this stage, diagnostic output should inform grouping and instructional emphasis without yet being used for high-stakes decisions such as placement or grade progression, since protected evaluation establishes predictive value but not yet the reliability required for stakes of that kind.

7.6. Scale

Expansion beyond the original pilot site should follow the transfer evidence gathered in protected evaluation, not administrative convenience or software availability alone; a domain pattern validated in one institutional and linguistic context should be re-checked, not assumed, when the diagnostic protocol moves to a new context, particularly one with different first-language backgrounds or different register norms, such as the Francophone administrative and academic contexts discussed in Section 3. The institutional and cross-setting dimensions of scaling are developed in full elsewhere in this research program; this paper's contribution ends at establishing that a domain pattern is worth scaling in the first place.

7.7. Monitoring

Ongoing monitoring tracks whether diagnostic classifications drift over time, whether teachers override diagnostic recommendations and why, whether any subgroup shows a rising rate of a particular classification that warrants a fairness review, and whether the program remains sustainable in the time and resources it demands from teachers and learners. A diagnostic protocol that cannot be monitored on an ongoing basis after its initial validation should not be treated as validated indefinitely. Monitoring should also track a specific failure mode this framework is designed to guard against: domain scores that were diagnostic at initial validation losing their predictive value as teachers begin teaching to the instrument rather than to the underlying construct, a version of measurement reactivity that threatens any repeated classroom assessment. A monitoring plan that periodically re-checks calibration against an independent measure, not only against the instrument's own prior results, is the safeguard against this drift.

8. Discussion

This review's central finding is negative in a useful sense: no single instrument or score, however well validated for ranking purposes, is adequate for locating which specific component of second language proficiency is limiting a given learner. The nine-domain taxonomy developed here does not resolve this by adding another composite score; it resolves it by making the componential structure explicit enough that a stalled proficiency trajectory can be traced to a specific, falsifiable, and actionable cause rather than treated as evidence that the learner is generally weak.

The most direct practical implication is sequencing. Programs that attempt to automate diagnostic classification, through computer-adaptive testing or machine-scored writing feedback, before establishing clear construct definitions, accountable roles, and a contestation process are likely to reproduce, at greater speed and scale, the same misclassification risk this framework is designed to reduce. Kostolanyova and Simonova's (2020a, 2020b) adaptive e-learning systems and Li's (2024) analysis of generative-model writing feedback both illustrate that technology can implement a diagnostic protocol faithfully or unfaithfully, and the technology itself does not settle which.

Methodologically, the framework's insistence on held-out evaluation and preregistered propositions is not a formality. Several of the instruments reviewed here were validated on the same population used to develop them, which is adequate for establishing that a construct can be measured reliably but not for establishing that the resulting diagnosis transfers to a new cohort or institution. The propositions in Section 6 are written so a future study can fail them; a taxonomy that cannot fail a well-designed test is not doing diagnostic work.

Ethically, the framework treats a diagnostic label as provisional and contestable by design, because the individual-differences literature synthesized in Section 2 shows how easily an early classification, whether cognitive, affective, or transfer-based, can harden into a fixed expectation that then shapes the very opportunities a learner is offered, independent of whether the classification remains accurate. Governance-layer requirements, including a documented contestation process and version-linked records, are not administrative overhead; they are the mechanism that keeps a diagnosis revisable.

The framework's main limitation is also its central claim: it is a taxonomy and measurement specification, not a report of validated diagnostic accuracy. Every domain separation proposed in Section 5 rests on evidence that each construct is separately measurable in isolated studies; whether the full nine-domain structure holds together, with manageable overlap and burden, when applied jointly to a single learner population is an empirical question this paper poses rather than answers. The propositions in Section 6 are the paper's attempt to make that question answerable rather than rhetorical.

This paper is scoped deliberately to identification. It does not specify how a teacher should respond instructionally once a domain-specific barrier is diagnosed, nor does it address how a validated diagnostic protocol should be adapted, trained, and monitored across the different institutional conditions of secondary and postsecondary settings. Both questions are taken up in later stages of this research program; treating them here would blur the line between what this paper actually establishes, a taxonomy and its evidentiary requirements, and what remains to be tested.

A further limitation concerns the burden a nine-domain, thirty-six-construct taxonomy places on a single teacher or program, even after the clustering justified in Section 1. Several of the instrument families reviewed in Section 3.1, particularly dynamic assessment and structured writing checklists, are time-intensive to administer well (Mehri &

Amerian, 2015; Safari & Ahmadi, 2023), and a program that adopts the full taxonomy without the staffing to administer it faithfully will generate implementation-layer failure rather than diagnostic insight. The implementation pathway in Section 7 addresses this directly by requiring teams to select a relevant subset of domains at the definition stage rather than treating full coverage as a prerequisite for beginning at all.

9. Conclusion

Beyond One Size Fits All Language Instruction sets out a diagnostic framework, not a validated intervention. The paper's contribution is to show that the applied linguistics and language testing literature already contains component-specific instruments for nearly every one of the thirty-six constructs organized here, and that what has been missing is not measurement capacity but an integrated architecture that specifies how those measurements should be combined, checked, and connected to a decision without collapsing back into the single aggregate score the framework exists to move beyond.

Future work executing this framework empirically must use data collected under the protocol specified here, preregistered propositions, appropriate ethics review for the age group and institutional setting involved, held-out validation, and transparent reporting of subgroups and null results; no accuracy, effect, or causal claim should be drawn from this paper alone, because none is offered here. What the paper does offer is a falsifiable answer to a genuine problem: aggregate proficiency scores are too coarse to guide individualized second language instruction, and the nine-domain taxonomy developed here is a testable proposal for the granularity individualization actually requires.

Its central proposition is that a stalled proficiency trajectory is rarely explained by a single cause, and that instruction individualized around a genuine, evidence-based diagnosis of the specific barrier in play will outperform instruction pitched at the aggregate score, provided the diagnosis is held to the same validity standards as any other measurement claim. The framework is ready for the empirical testing this paper deliberately does not attempt.

For readers less interested in the taxonomy's internal architecture than in its practical upshot, the argument reduces to three claims. First, a learner's proficiency score is an average over components that can be moving in different directions, and averaging discards exactly the information a teacher needs to help. Second, the applied linguistics field already has the instruments needed to measure most of those components separately; the missing piece has been an architecture for combining them responsibly. Third, that architecture must be falsifiable, contestable by the learners it classifies, and revisable when it is wrong, or it risks replacing one blunt instrument, the aggregate score, with thirty-six blunt instruments instead of one.

Declarations

Ethics approval and consent to participate: No human or animal data were collected for this conceptual and methodological paper. Any future empirical study applying the taxonomy proposed here must obtain approvals appropriate to its participants, including, where minors are involved, parental consent and institutional review, records, and jurisdiction.

Consent for publication: Not applicable.

Availability of data and materials: No new dataset was generated or analyzed. The construct taxonomy, measurement matrix, and evaluation criteria are reported in full within the manuscript.

Competing interests: The authors declare no competing interests.

Funding: The authors received no specific funding for this work.

Authors' contributions: Raissa Enyonam Amele Amessouwoe led conceptualization of the diagnostic taxonomy, drafted the critical review and theoretical foundations, and coordinated the manuscript. Kpoka Messanh Isaac developed the five-layer evidence architecture and the measurement matrix, and contributed the contextual and register-variation analysis. Julius Aderogba Odewabi developed the testable propositions and the implementation pathway, and conducted reference and citation verification. All authors reviewed and approved the final manuscript.

Acknowledgments: None.

References

1. Alderson JC. Diagnosing Foreign Language Proficiency. Continuum; 2005.
2. Bachman LF, Palmer AS. Language Assessment in Practice. Oxford University Press; 2010.
3. Badrun Kholid A, Rahman A, Irawan LA. Implementing Diagnostic Assessment in Designing Differentiated Learning for English Language Learning at the Junior High Schools. J Lang Lit Stud. 2024;4(2):445-58. doi:10.36312/jolls.v4i2.1934
4. Black P, Wiliam D. Assessment and classroom learning. Assess Educ. 1998;5(1):7-74.
5. Brown A. Multicompetence and Second Language Assessment. Lang Assess Q. 2013;10(2):219-35. doi:10.1080/15434303.2013.769551
6. Butler YG. Self-assessment in second language learning. Lang Teach. 2023;57(1):42-56. doi:10.1017/s0261444822000489
7. Butler YG. Self-assessment in second language learning. Lang Teach. 2024.
8. Caplan NA. Learning through language: A response to Polio, Keeping the language in second language classes. J Second Lang Writ. 2019;46:100677. doi:10.1016/j.jslw.2019.100677
9. Corder SP. The significance of learners' errors. Int Rev Appl Linguist Lang Teach. 1967;5(1-4):161-70.
10. Council of Europe. Common European Framework of Reference for Languages Companion Volume. 2020.
11. de Bot K. The effectiveness of early foreign language learning in the Netherlands. Stud Second Lang Learn Teach. 2014;3:409-18. doi:10.14746/ssllt.2014.4.3.2
12. DeKeyser RM. Aptitude-treatment interaction in second language learning. J Second Lang Stud. 2019;2(2):165-8. doi:10.1075/jsls.00007.int
13. Dörnyei Z. The Psychology of the Language Learner. Lawrence Erlbaum; 2005.
14. Dunkel PA. Considerations in developing or using second/foreign language proficiency computer-adaptive tests. Lang Learn Technol. 1999;2(2):77-93. doi:10.64152/10125/25044
15. Ellis R. The Study of Second Language Acquisition. Oxford University Press; 2008.
16. Figueiredo SA, Martins MA, Fernandes da Silva C. Education and Learning in Second Language Contexts: Language and Diagnostic Assessment of Diverse

Learners in Portuguese Schools. Int J Divers Educ. 2013;13(1):31-40. doi:10.18848/2327-0020/cgp/v13i01/40078
17. Forgor AT. The impact of adopting of assessment for learning and adaptive learning system on students' foreign language performance. Laplage Rev. 2021;7(1):221-8. doi:10.24115/s2446-6220202171707p.221-228
18. Foster-Cohen SH. Relevance Theory and second language learning/behaviour. Second Lang Res. 2004;20(3):189-92. doi:10.1191/0267658304sr236ed
19. Fulcher G. Practical Language Testing. Hodder Education; 2010.
20. Gaillard S, Tremblay A. Linguistic Proficiency Assessment in Second Language Acquisition Research: The Elicited Imitation Task. Lang Learn. 2016;66(2):419-47. doi:10.1111/lang.12157
21. Gardner RC. Social Psychology and Second Language Learning: The Role of Attitudes and Motivation. Edward Arnold; 1985.
22. Gass SM. Language Universals and Second Language Acquisition. Lang Learn. 1989;39(4):497-534. doi:10.1111/j.1467-1770.1989.tb00901.x
23. Ghanbarpour M. A Qualitative Meta-synthesis of Research on Dynamic Assessment of Second/Foreign Language Learning: Implications for Language Teachers. J Lang Teach Res. 2017;8(4):731. doi:10.17507/jltr.0804.12
24. Harding L, Alderson JC, Brunfaut T. Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. Lang Test. 2015;32(3):317-36. doi:10.1177/0265532214564505
25. Hashim H, Salam S, Nurul S, Syafiatun N. The Designing of Adaptive Self-Assessment Activities in Second Language Learning using Massive Open Online Courses (MOOCs). Int J Adv Comput Sci Appl. 2018;9(9). doi:10.14569/ijacsa.2018.090937
26. Hattie J, Timperley H. The power of feedback. Rev Educ Res. 2007;77(1):81-112.
27. Heriansyah H, Darni P, Fajri R, Sahardin R. Competency-based curriculum, relevant second language learning theories, and its language assessment. Engl Educ J. 2021;12(3):416-27. doi:10.24815/eej.v12i3.21631
28. In'nami Y. Second-language vocabulary assessment research: Issues and challenges. Vocab Learn Instr. 2013;2(1):74-81. doi:10.7820/vli.v02.1.innami
29. Isaac KM. Comparative analysis of administrative French in Francophone African states and its impact on translation practice. J Public Adm Soc Welf Res. 2023;8(2):149-63. doi:10.56201/jpaswr.v8.no2.2023.pg149.163
30. Isaac KM. Terminological challenges in the translation of civil registry and administrative documents. Int J Soc Sci Manag Res. 2024;10(11):552-78. doi:10.56201/ijssmr.v10.no11.2024.pg.552.578
31. Isbell DR. Can the Test Support Student Learning? Validating the Use of a Second Language Pronunciation Diagnostic. Lang Assess Q. 2021;18(4):331-56. doi:10.1080/15434303.2021.1874382
32. Isbell DR, Lee J. Self-Assessment of Comprehensibility and Accentedness in Second Language Korean. Lang Learn. 2022;72(3):806-52. doi:10.1111/lang.12497
33. Jakobovits LA. Second language learning and transfer theory: A theoretical assessment. Lang Learn. 1969;19(1-2):55-6. doi:10.1111/j.1467-1770.1969.tb00933.x
34. Johnson JS. Critical Period Effects in Second Language Acquisition: The Effect of Written versus Auditory Materials on the Assessment of Grammatical Competence. Lang Learn. 1992;42(2):217-48. doi:10.1111/j.1467-1770.1992.tb00708.x
35. Ketabi S, Ketabi S. Classroom and Formative Assessment in Second/Foreign Language Teaching and Learning. Theory Pract Lang Stud. 2014;4(2). doi:10.4304/tpls.4.2.435-440
36. Koda K. Reading and Language Learning: Crosslinguistic Constraints on Second Language Reading Development. Lang Learn. 2007;57(s1):1-44. doi:10.1111/0023-8333.101997010-i1
37. Koizumi R. Second language vocabulary assessment studies: Validity evidence and future directions. Vocab Learn Instr. 2015;4(1):36-46. doi:10.7820/vli.v04.1.koizumi
38. Kostolanyova K, Simonova I. Adaptive e-learning model for learning English as a second/foreign language. Int J Innov Learn. 2020;27(3):274. doi:10.1504/ijil.2020.10028282
39. Kostolanyova K, Simonova I. Adaptive e-learning model for learning English as a second/foreign language. Int J Innov Learn. 2020;27(3):274. doi:10.1504/ijil.2020.106811
40. Lanteigne B. Unscrambling jumbled sentences: An authentic task for English language assessment? Stud Second Lang Learn Teach. 2017;7(2):251-73. doi:10.14746/ssllt.2017.7.2.5
41. Lee I, Coniam D. Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. J Second Lang Writ. 2013;22(1):34-50. doi:10.1016/j.jslw.2012.11.003
42. Li M. Leveraging ChatGPT for Second Language Writing Feedback and Assessment. Int J Comput Assist Lang Learn Teach. 2024;14(1):1-11. doi:10.4018/ijcallt.360382
43. Macaro E. Second language teachers as second language classroom researchers. Lang Learn J. 2003;27(1):43-51. doi:10.1080/09571730385200071
44. Mehri E, Amerian M. Challenges to Dynamic Assessment in Second Language Learning. Theory Pract Lang Stud. 2015;5(7):1458. doi:10.17507/tpls.0507.19
45. Menaka G, Sankar G. The language learning assessment using technology for the second language learners. Int J Linguist Lit Cult. 2019;5(4):1-6. doi:10.21744/ijllc.v5n4.674
46. Mislevy RJ, Yin C. If Language Is a Complex Adaptive System, What Is Language Assessment? Lang Learn. 2009;59(s1):249-67. doi:10.1111/j.1467-9922.2009.00543.x
47. Mohamadi Z. Task engagement: a potential criterion for quality assessment of language learning tasks. Asian-Pac J Second Foreign Lang Educ. 2017;2(1). doi:10.1186/s40862-017-0025-z
48. Nation ISP. Learning Vocabulary in Another Language. Cambridge University Press; 2001.
49. Navracsics J. Input or intimacy. Stud Second Lang Learn Teach. 2014;3:485-506. doi:10.14746/ssllt.2014.4.3.6
50. Neumann H. Teacher assessment of grammatical ability

in second language academic writing: A case study. J Second Lang Writ. 2014;24:83-107. doi:10.1016/j.jslw.2014.04.002
51. Nida EA. Motivation in second language learning. Lang Learn. 1956;7(3-4):11-6. doi:10.1111/j.1467-1770.1956.tb01198.x
52. Norris J. Concerns with Computerized Adaptive Oral Proficiency Assessment. Lang Learn Technol. 2001;5(2):99-105. doi:10.64152/10125/25131
53. Oh S. Second Language Learners' Use of Writing Resources in Writing Assessment. Lang Assess Q. 2019;17(1):60-84. doi:10.1080/15434303.2019.1674854
54. Oxford RL. Teaching and Researching Language Learning Strategies. Routledge; 2017.
55. Poole D. Language Socialization in the Second Language Classroom. Lang Learn. 1992;42(4):593-616. doi:10.1111/j.1467-1770.1992.tb01045.x
56. Pourdana N. Impacts of computer-assisted diagnostic assessment on sustainability of L2 learners' collaborative writing improvement and their engagement modes. Asian-Pac J Second Foreign Lang Educ. 2022;7(1). doi:10.1186/s40862-022-00139-4
57. Pulido D. Modeling the Role of Second Language Proficiency and Topic Familiarity in Second Language Incidental Vocabulary Acquisition Through Reading. Lang Learn. 2003;53(2):233-84. doi:10.1111/1467-9922.00217
58. Rafi F, Pourdana N. E-diagnostic assessment of collaborative and individual oral tiered task performance in differentiated second language instruction framework. Lang Test Asia. 2023;13(1). doi:10.1186/s40468-023-00223-7
59. Roever C. DIF in the Assessment of Second Language Pragmatics. Lang Assess Q. 2007;4(2):165-89. doi:10.1080/15434300701375733
60. Safari F, Ahmadi A. Developing and evaluating an empirically-based diagnostic checklist for assessing second language integrated writing. J Second Lang Writ. 2023;60:101007. doi:10.1016/j.jslw.2023.101007
61. Sang F. Second language acquisition and second language learning. System. 1983;11(1):84-9. doi:10.1016/0346-251x(83)90013-1
62. Selinker L. Interlanguage. Int Rev Appl Linguist Lang Teach. 1972;10(1-4):209-31.
63. Shin JA, Christianson K. Structural Priming and Second Language Learning. Lang Learn. 2011;62(3):931-64. doi:10.1111/j.1467-9922.2011.00657.x
64. Spinner P. Second language assessment and morphosyntactic development. Stud Second Lang Acquis. 2011;33(4):529-61. doi:10.1017/s0272263111000301
65. Spolsky B. Attitudinal aspects of second language learning. Lang Learn. 1969;19(3-4):271-5. doi:10.1111/j.1467-1770.1969.tb00468.x
66. Ströbel M, Kerz E, Wiechmann D. The Relationship Between First and Second Language Writing: Investigating the Effects of First Language Complexity on Second Language Complexity in Advanced Stages of Learning. Lang Learn. 2020;70(3):732-67. doi:10.1111/lang.12394
67. Sun Z, Xu P, Wang J. Dynamic Assessment of the Learning Potential of Chinese as a Second Language. Lang Assess Q. 2022;20(1):127-42. doi:10.1080/15434303.2022.2151911
68. Teló C, Kivistö de Souza H, O'Brien MG, Carlet A. Second Language Sentence Stress Assignment: Self- and Other-Assessment. Lang Learn. 2024;75(3):832-62. doi:10.1111/lang.12682
69. Thomas M. Assessment of L2 Proficiency in Second Language Acquisition Research. Lang Learn. 1994;44(2):307-36. doi:10.1111/j.1467-1770.1994.tb01104.x
70. Wakabayashi S. Lexical learning in second language acquisition: optionality in the numeration. Second Lang Res. 2009;25(2):335-41. doi:10.1177/0267658308100293
71. Waniek-Klimczak E. Review of Second language pronunciation assessment: Interdisciplinary perspectives (T. Isaacs & P. Trofimovich, Eds.). Stud Second Lang Learn Teach. 2018;8(3):701-6. doi:10.14746/ssllt.2018.8.3.9
72. Xu H, Li G. CEFR-informed learning, teaching and assessment: A practical guide. Stud Second Lang Learn Teach. 2024;14(2):365-70. doi:10.14746/ssllt.32669
73. Yang S. Constructing Diagnostic Reading Assessment Instruments for Low-level Chinese as Second Language Learners. Read Foreign Lang. 2021;33(2):212-37. doi:10.64152/10125/67401
74. Yeotak Yoon, 김정우, Min H. A Study of the Diagnostic Assessment of Basic Competency in Middle School Korean Language Learning. Korean Lang Educ Res. 2008;32:305-40. doi:10.20880/kler.2008.32.305
75. Zhang LJ. Second language writing as and for second language learning. J Second Lang Writ. 2013;22(4):446-7. doi:10.1016/j.jslw.2013.08.010
76. 이연숙. Assessing Second Language Grammar Using Cognitive Diagnostic Assessment Method. J Stud Lang. 2014;30(2):395-415. doi:10.18627/jslg.30.2.201408.395