International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Syntactic Complexity as a Compensatory Heuristic: Towards a Model of Evaluative Judgment in EFL Writing Assessment

Huda Husain Al-Asfoor[1], Dr. Shadiya Al-Hashmi[2], Prof. Sayed Sadek[3]
[1]MA Student, Department of English Language and Literature, College of Arts, University of Bahrain, Sakhir, Kingdom of Bahrain
[2]Assistant Professor, Department of English Language and Literature, College of Arts, University of Bahrain, Sakhir, Kingdom of Bahrain
[3]Professor, Department of English Language and Literature, College of Arts, University of Bahrain, Sakhir, Kingdom of Bahrain
Corresponding Author: Huda Husain, **E-mail**: 20198516@stu.uob.edu.bh, **ORCID:** https://orcid.org/0009-0000-8271-4709

| ABSTRACT
Existing accounts of second language processing explain how EFL learners extract meaning from text, manage limited working-memory capacity, and register salient linguistic features. However, none specifies what happens when a reader must convert that processing into a single evaluative judgment of writing quality. This paper addresses that gap by proposing the Compensatory Model of Evaluative Judgment (CMEJ), which integrates Input Processing Theory, Cognitive Load Theory, and the Noticing Hypothesis into a single, testable mechanism. CMEJ predicts that evaluative judgment follows one of two pathways depending on a reader's prior topic knowledge. Under sufficient familiarity, judgment rests on content-based criteria such as coherence and evidential strength. Under limited familiarity, judgment defaults to structurally salient heuristics, principally syntactic complexity, as a proxy for quality. The relationship between topic knowledge and reliance on syntactic complexity is modelled as continuous and directional rather than threshold-based and is specified simultaneously as a statistical moderation and a cognitive mediation. The paper positions CMEJ against the three frameworks it builds on, examines the model's own opportunities and boundary conditions, and outlines theoretical and pedagogical implications for rater training, rubric design, and assessment validity in EFL writing assessment. CMEJ is offered here as a conceptual, falsifiable contribution; an empirical test of its central prediction is identified as the direct next step.

| KEYWORDS
Evaluative Judgment, Syntactic Complexity, Cognitive Load, Compensatory Model

| ARTICLE INFORMATION
ACCEPTED: 02 September 2026 **PUBLISHED:** 27 October 2026 **DOI:** 10.32996/ijllt.2026.9.11.2

1. The Acquisition-Evaluation Gap

Comprehension and evaluation are often treated as though they were the same cognitive act performed twice, when in fact they answer entirely different questions. To comprehend a piece of writing is to ask *what does this mean*. To evaluate it is to ask *should I trust this, act on it, or defer to it.* A reader can achieve the first without ever attempting the second, and in an information environment where volume routinely outpaces verification, that gap is not a minor cognitive detail but the primary site of risk. Misinformation rarely succeeds by being incomprehensible. It succeeds by being fluently understood and then insufficiently judged. This is precisely why evaluative judgment deserves theoretical treatment in its own right rather than being assumed to follow automatically once comprehension is complete. A learner, or any reader, who has fully absorbed a text's content has not yet done the harder and more consequential work of deciding what that content is worth.

Research on second language processing has long treated the mechanisms underlying comprehension and those underlying evaluation as separate concerns. Input Processing Theory explains how L2 learners map syntactic structures onto meaning as they process input. Cognitive Load Theory accounts for how limited working memory capacity constrains what learners can process during complex tasks. Schmidt's Noticing Hypothesis adds a further constraint, which claims that conscious awareness of linguistic form, at the level of noticing, is treated as a necessary condition for subsequent acquisition. Each of these frameworks, however,

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

was developed to explain how learners *acquire* language, not how they *judge* the quality of language once acquisition-related processing has already occurred. When an EFL learner is asked to evaluate the quality of a piece of academic writing, they are not simply processing input for acquisition. They are weighing competing cues (syntactic complexity against topic familiarity, structural sophistication against content accessibility) to arrive at a single evaluative judgment. None of the three frameworks, individually or in combination as currently formulated, specifies the mechanism by which such competing cues are weighed and resolved. This paper addresses that gap by proposing the Compensatory Model of Evaluative Judgment (CMEJ), which reframes learner judgment as a compensatory process in which strength on one dimension can offset weakness on another.

This shift, from theories of acquisition to a theory of evaluation, is not something the paper is entitled to assume. It requires its own justification, and two considerations support it. The first concerns what these frameworks are actually claims about. VanPatten, Sweller, and Schmidt did not set out to describe a single event called acquisition. They set out to describe the real-time cognitive architecture, limited attention, limited working memory, a gate of conscious registration, through which a second language learner processes linguistic input at all. That architecture does not switch off once a form has been acquired, nor hand control to some separate faculty for evaluation. The same limited-capacity system a learner uses to acquire language is the system that learner is still using when later asked to judge the quality of someone else's writing. Acquisition and evaluation are different tasks performed on the same machinery, not two machineries running independently.

The second consideration concerns why evaluation deserves attention within an acquisition-oriented literature specifically, rather than being left to assessment research alone. This is not a claim that evaluation gates acquisition in the way noticing does. A learner can acquire a form through repeated exposure without ever judging its quality, and nothing in Schmidt's account requires otherwise. The claim is narrower: once a learner has the capacity to notice and weigh competing forms, evaluative judgment determines which of those forms the learner treats as reliable enough to consciously adopt, cite, or reproduce in their own output. Evaluation, on this narrower reading, does not control what a learner unconsciously absorbs, but it does control what a learner deliberately relies on and passes on to others. A learner who cannot evaluate has no principled way to distinguish a form worth trusting from one merely encountered. That distinction, not acquisition itself, is what evaluative judgment governs. Read this way, evaluative judgment is not a topic borrowed from outside these three frameworks for convenience. It is the downstream act that gives their upstream claims about processing consequence, and it draws on the identical cognitive resources these frameworks were built to describe. Thus, this paper addresses the following questions:

1. Why do Input Processing Theory, Cognitive Load Theory, and the Noticing Hypothesis fail to account specifically for evaluative judgment, despite explaining related stages of language processing?

2. What cognitive mechanism governs the relationship between syntactic complexity and perceived writing quality in EFL readers?
    2.1. How does that mechanism vary as a function of the reader's prior topic knowledge?

3. What theoretical and pedagogical consequences follow from modelling this relationship as compensatory, rather than fixed, and what does this imply for how academic writing is rated in EFL contexts?

In methodological terms, this paper follows an integrative theory-synthesis and model-building design (cf. Jaakkola, 2020; MacInnis, 2011). Three frameworks, each authoritative within its own strand of second language processing research but individually silent on evaluative judgment, are critically reviewed, and their point of mutual silence is used to motivate a new, integrative model specified with falsifiable propositions. The contribution offered is accordingly one of integration in MacInnis's (2011) sense: the paper does not introduce evaluative judgment as a wholly new construct, nor does it report new data. It demonstrates that a phenomenon three established frameworks each partially bear on, without any naming it directly, can be given a single, coherent, testable account once those frameworks are read against each other.

1.1 VanPatten: Input Processing Theory

Bill VanPatten developed Input Processing Theory across a series of publications beginning in 1996, situating the model within a broader tradition of psycholinguistically motivated approaches to second language acquisition (VanPatten, 2015). The theory centres on how learners convert linguistic input into intake, the subset of input that is actually available for further cognitive processing. Two mechanisms drive this conversion: the first is form-meaning connections, which is the process by which learners map a given linguistic form onto its corresponding meaning. The second is processing strategies, the default parsing preferences learners rely on when interpreting sentence structure (such as assigning subject status to the first noun encountered). Input Processing Theory has been particularly influential in explaining why certain grammatical forms resist acquisition even under conditions of abundant exposure, and it underlies the pedagogical approach known as processing instruction. The theory's explanatory scope, however, stops at intake. It accounts for how a learner extracts form-meaning mappings from a sentence, not for what a learner does with that sentence once extraction is complete. A learner may successfully process a syntactically complex

sentence, may correctly map its form onto its meaning, and still face a separate task: deciding whether that same complexity signals academic sophistication or undermines clarity.

1.2 Sweller: Cognitive Load Theory

John Sweller, an Australian cognitive and educational psychologist based at the University of New South Wales, introduced Cognitive Load Theory in a 1988 study examining why conventional problem-solving methods often fail to produce durable learning (Sweller, 1988). The theory rests on a working-memory constraint: because working memory can hold and manipulate only a limited number of elements at once, any task that exceeds this capacity leaves insufficient cognitive resources for schema construction. Sweller and later collaborators distinguished three components of load: intrinsic load, arising from the inherent complexity of the material itself; extraneous load, arising from how information is presented and therefore avoidable through better design; and germane load, the cognitive effort directly devoted to building new schemas. The theory's central prescription is that instructional design should minimise extraneous load so that working-memory capacity remains available for germane processing. Applied to EFL learners evaluating academic writing, Cognitive Load Theory offers a partial but incomplete account. It explains why a syntactically dense sentence might overwhelm a learner's working memory and impede accurate evaluation; intrinsic and extraneous load both plausibly rise with structural complexity. What the theory does not explain is the direction of the resulting judgment. Overload could just as easily be read by a learner as evidence of sophistication as it could be read as evidence of poor clarity. Cognitive Load Theory predicts that processing becomes harder, not how a learner converts that difficulty into a qualitative evaluation. It describes a constraint on processing capacity, not the evaluative logic learners apply once that constraint is felt.

1.3 Schmidt: Noticing Hypothesis

Richard Schmidt set out the Noticing Hypothesis most fully in a chapter drawing in part on his own diary-based observations of learning Portuguese as an adult (Schmidt, 2001). The hypothesis distinguishes several levels of consciousness relevant to language learning (perception, noticing, and understanding), and argues that noticing, the conscious registration of a linguistic form in the input, is a necessary condition for that form to become intake available for acquisition. On this account, learners cannot acquire a form they have not, at some level, consciously attended to; unattended input, however frequent, does not accumulate into learning. Noticing is triggered by factors such as perceptual salience, frequency of occurrence, and instructional emphasis, and it operates in short-term memory as the gateway between exposure and learning. The Noticing Hypothesis, like the two frameworks above, is oriented towards the front end of the acquisition process. It specifies a condition for a form to enter the system, not a mechanism for what happens once several noticed forms must be weighed against one another in a single evaluative act. An EFL learner evaluating academic writing typically notices multiple features simultaneously (syntactic complexity, topic-related vocabulary, structural markers of cohesion). Schmidt's model offers no account of how these separately noticed features are integrated, prioritised, or traded off against one another to produce one overall judgment of quality. Noticing explains what enters conscious awareness, but it does not explain how competing noticed elements are reconciled.

1.4 Syntactic Complexity as a Construct

The fourth construct central to this paper, syntactic complexity itself, warrants the same treatment given to the three frameworks above: a definition, before a critique. Ortega (2003) established syntactic complexity as a measurable index of L2 development in a synthesis of twenty-five studies, defining it as the range and sophistication of grammatical forms a writer deploys, and showing that its relationship to proficiency varies systematically depending on learning context and how proficiency itself is defined. Lu (2010) operationalised the construct computationally, decomposing it into measurable dimensions, clause density, phrasal elaboration, subordination, and length-based indices, that could be extracted automatically from a text rather than coded by hand. Kyle, Crossley, and Verspoor (2021) extended this line of work by tracking these indices longitudinally, showing that syntactic development is neither linear nor uniform across dimensions: clausal and phrasal complexity develop along different trajectories, and length-based measures alone conflate distinctions that finer-grained indices can separate.

Three implications follow from this literature for the argument developed here. First, syntactic complexity is not a single variable but a multidimensional construct (clausal subordination, phrasal elaboration, and sentence length are related but empirically separable aspects of a text), and treatments that collapse complexity into one homogeneous measure risk conflating dimensions that behave differently. Second, this literature has been developed almost entirely to describe production, what a learner writes, and how that writing changes as proficiency develops, rather than reception: *how a reader evaluating a text a learner has produced perceives and weighs that same complexity.* Third, and following from the second point, the literature offers no account of why structural complexity, a property with a well-documented and often weak relationship to writing quality as such (Ortega, 2003), should nonetheless function as a perceptual proxy for quality under some evaluative conditions and not others. This is precisely the question Sections 2 and 3 take up: not what syntactic complexity is, which this literature has answered, but what a reader does with it once meaning-extraction, capacity, and attention, the three concerns of Sections 1.1 through 1.3, place that reader in a position to notice it.

One further question must be addressed before the argument proceeds: *Why syntactic complexity specifically, rather than some other textual property, and rather than treating complexity itself as an irreducible bundle of separate features with no single governing variable.* The literature reviewed above is not settled on this point. Lu (2010) and Kyle, Crossley, and Verspoor (2021) decompose complexity into distinct clausal, phrasal, and length-based dimensions precisely because these dimensions develop at different rates and correlate differently with proficiency, treating unity as an empirical question rather than a starting assumption. Yasuda (2026) pushes this further, arguing that quantification alone, whether unified or multidimensional, cannot adequately capture syntactic complexity without reference to how appropriately a structure is deployed for its communicative purpose. Yasuda (2024) adds a closely related finding: form-based complexity and meaning-based complexity in EFL argumentative writing do not move together, and more of the former does not reliably predict higher perceived quality. Read together, this literature suggests two live positions: syntactic complexity as a decomposed set of independent elements, each carrying its own developmental and evaluative signal, or syntactic complexity as a unified construct, useful precisely because it captures how a reader experiences a text rather than how a writer produced it.

The stakes of this disagreement are easiest to see with a concrete illustration. Consider two texts assigned the same aggregate complexity score, five units, say, by any standard index. The first reaches that score through generic length and clausal coordination, while the second reaches the identical score through nominalisation and passive constructions specifically. A unified, holistic measure treats the two texts as equivalent: same total, same construct value, nothing further to say. A feature-specific measure treats them as different in a way the aggregate obscures, since nominalisation and passive voice carry independent associations with academic register that generic length does not, so two texts of identical aggregate complexity could plausibly earn different quality judgments precisely because of *which structures*, not *how much structure*, produced that total. The disagreement, in other words, is not about whether these texts are equally complex by some measure. It is about whether equal aggregate complexity implies equal evaluative consequence, which is exactly the question a purely quantitative index is not built to answer (Yasuda, 2026).

This paper does not resolve which specific features carry disproportionate evaluative weight, that is an empirical question beyond its scope. It takes a position on a related, prior question instead: *how much of that feature-level detail, whatever it turns out to be, a reader can actually access under real-time cognitive constraint*. Rather than treating the two positions above as mutually exclusive accounts of the construct, this paper argues that CMEJ reconciles them within a single, dynamic account of reception. CMEJ is a model of how a text is read, not of how it was written. It asks what a reader under cognitive constraint notices and weighs, not how a writer's interlanguage develops. On this view, whether a reader engages with syntactic complexity as a decomposed set of features, the Feature-Specific View, or as a single undifferentiated impression, the Unitary, or Aggregate, View, is not a fixed property of readers or of the construct itself. It is a function of the cognitive load the reader is under "at the moment" of evaluation, which is in turn a function of their prior topic knowledge. When a reader holds low prior topic knowledge, working memory is already heavily committed to basic meaning-extraction. Under this load, the reader lacks the spare cognitive resources fine-grained syntactic parsing would require, separating clausal subordination from phrasal elaboration, and so falls back on the coarser option: registering structural density as a single, undifferentiated impression. This is the Unitary View, and it is the processing mode CMEJ assumes under Pathway Two. When a reader holds high prior topic knowledge, by contrast, semantic processing is comparatively effortless, leaving working memory free. Syntactic complexity still contributes minimally to the judgment under these conditions, content continues to dominate, but on the rare occasion the reader does attend to it, that spare capacity permits the finer-grained discrimination the Feature-Specific View describes: distinguishing, for instance, a passive construction from a subordinate clause rather than registering both as generic complexity. Multidimensional decomposition and unified processing are, on this account, not two competing theories of what syntactic complexity is. They are two processing modes available to the same reader, and which mode is active is precisely what Section 3's moderation mechanism formalises.

The remainder of this paper proceeds as follows. Section 2 synthesises the three acquisition-oriented frameworks reviewed above to motivate the Compensatory Model of Evaluative Judgment, formally presented in Section 3 together with its moderation and mediation structure. Section 4 positions CMEJ against the theoretical frameworks it integrates, the empirical rater cognition literature, and the methodological grounding of the instrument a companion study will use to test it. Section 5 examines what the model makes newly possible and where its own explanatory reach runs out, before Section 6 relates this argument back to the three research questions directly. Sections 7 through 9 then turn to implications for researchers, practitioners, and assessment validity, limitations of the present account, and directions for empirical validation, before Section 10 concludes.

2. Synthesis

The three frameworks reviewed above share a common orientation. Each was developed to explain a stage of language processing (intake, capacity, or conscious registration) rather than the evaluative judgment that follows once processing has occurred. Input Processing Theory specifies which linguistic elements a learner is likely to extract from a sentence but offers no account of how extracted elements are subsequently appraised for quality. Cognitive Load Theory specifies the conditions under which processing

becomes effortful but does not specify whether that effort is interpreted by the reader as a marker of difficulty or of sophistication. The Noticing Hypothesis specifies which features are salient enough to enter conscious awareness but offers no mechanism for reconciling multiple simultaneously noticed features into a single judgment. Individually, then, each theory addresses a necessary precondition for evaluation (attention, capacity, awareness) without addressing evaluation itself.

Read together, however, the three frameworks converge on a shared implication that none states explicitly. Judgment is not primarily a response to a text's objective properties, but to the interaction between those properties and the finite cognitive resources available to the reader at the moment of evaluation. Topic familiarity's broader consequences for L2 performance are increasingly well documented (Wu & Qin, 2025), and its specific consequence for evaluative judgment follows the same underlying logic: when topic familiarity is low, meaning is harder to anchor (VanPatten, 2015); when meaning is harder to anchor, processing becomes effortful (Sweller et al., 2011); and under effortful processing, readers default to whichever features are perceptually salient enough to have been noticed at all (Schmidt, 2001). Syntactic complexity, by virtue of its structural visibility, is disproportionately likely to be among those noticed features, and so disproportionately likely to stand in for a quality judgment the reader lacks the resources to make more substantively. This convergence, implicit across all three frameworks but formalised in none of them, is the mechanism the Compensatory Model of Evaluative Judgment (CMEJ) makes explicit.

3. The Compensatory Model of Evaluative Judgment

Building on the theoretical convergence outlined above, this paper proposes the Compensatory Model of Evaluative Judgment (CMEJ). The model formalises a single claim: when a reader's capacity for content-based evaluation is constrained, evaluative attention shifts towards structurally salient features of the text as a substitute basis for judgment. The term compensatory refers to this substitution, perceptually accessible form standing in for conceptual access the reader cannot otherwise achieve. CMEJ specifies two evaluative pathways, activated as a function of the reader's prior topic knowledge rather than as fixed reader types. Under Pathway One (knowledge-informed evaluation), sufficient topic familiarity allows meaning to be extracted with relatively low cognitive demand (VanPatten, 2015), leaving working memory resources available for higher-order appraisal (Sweller, Ayres, & Kalyuga, 2011). Evaluative judgment under this pathway is driven primarily by substantive criteria (argument coherence, thesis clarity, and evidential strength) with syntactic complexity contributing minimally to the overall judgment. Under Pathway Two (knowledge-limited evaluation), insufficient topic familiarity forces the majority of available cognitive resources towards basic meaning construction, leaving little capacity for evaluating argument quality directly. Under these conditions, perceptually salient structural features, such as sentence length, clausal density, and phrasal elaboration (Schmidt, 2001), are recruited as a low-effort heuristic proxy for quality, independent of their actual contribution to the argument's soundness.

Critically, CMEJ does not model this shift as a threshold effect. The relationship between topic knowledge and reliance on syntactic complexity is proposed as graded and proportional. Reliance on structural cues increases continuously as topic knowledge decreases, and diminishes continuously as topic knowledge increases. The model therefore predicts an interaction between topic knowledge and syntactic complexity in shaping evaluative judgment, rather than a direct effect of either variable in isolation. This distinguishes CMEJ's central claim from prior accounts that treat syntactic complexity as a uniformly positive or uniformly negative predictor of perceived writing quality. Under CMEJ, the sign and magnitude of that relationship is conditional on the reader's cognitive access to content.

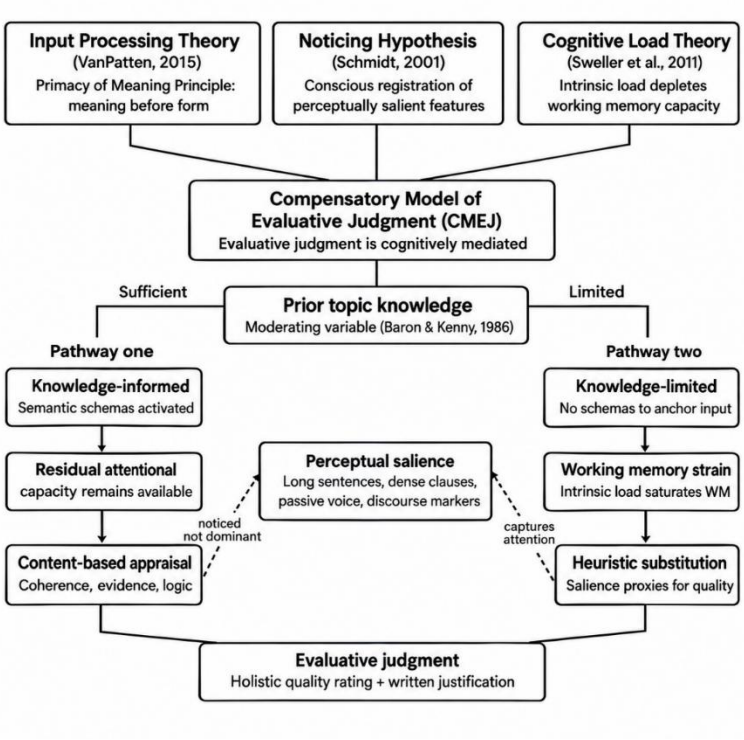


Figure 1. Compensatory Model of Evaluative Judgment (CMEJ)

Figure 1 sets out the mechanism described above. The branch point is topic familiarity. A reader who knows the subject well proceeds down Pathway One, where enough semantic scaffolding is already in place that judgment can rest on content, coherence, evidence, and the logic holding an argument together. A reader without that background proceeds down Pathway Two instead. There, meaning has to be built from scratch, working memory fills fast, and little capacity is left for appraising the argument on its merits. What fills the gap is whatever the eye catches without effort: sentence length, clause density, the visible architecture of the prose. Syntactic complexity, in other words, does not carry a fixed value in this model. Under Pathway One it barely registers. Under Pathway Two it becomes the judgment.

3.1 Topic Knowledge as Moderator and Mediator

By now, it becomes clear that topic knowledge does two different jobs in this model, and the two jobs sit at different levels of explanation entirely. At the level of statistics, it behaves as a moderator. A moderator does not predict the outcome directly. Instead, it changes the strength, and here the direction, of the relationship between two other variables. Run the numbers on CMEJ and this is what a regression would recover: an interaction term between topic knowledge and syntactic complexity, where the coefficient linking complexity to evaluative judgment shifts depending on how much the reader already knows about the subject. Under high topic familiarity that coefficient is close to flat. Under low familiarity it steepens. This is the pattern: it is what shows up in a table of results, and nothing more than that.

On the other hand, mediation operates on a different question. Where moderation asks when a relationship holds, mediation asks how it comes to hold at all, through what sequence of intervening steps one variable produces its effect on another. Here, limited topic knowledge does not act on evaluative judgment in one move; it sets off a chain. First, knowledge gaps force the reader to spend working memory on basic meaning construction rather than argument evaluation (Sweller et al., 2011). Second, depleted capacity leaves little room for weighing coherence or evidence directly. Under those conditions, the reader defaults to whichever features are perceptually loud enough to register without effort (sentence length, clause density, the visible architecture of the prose) (Schmidt, 2001). Each step in that chain causes the next. Remove any one link and the chain does not reach its conclusion.

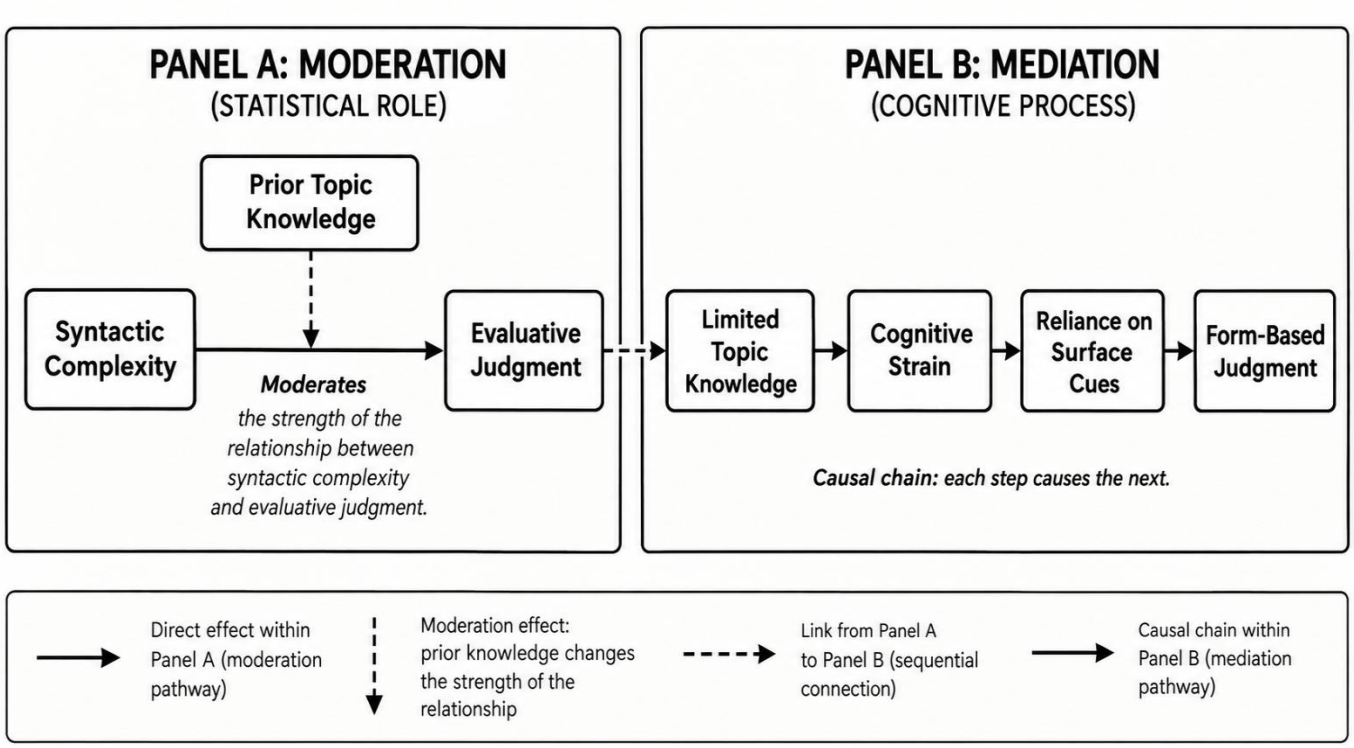


Figure 2. Moderation versus Mediation in the CMEJ

The difference is easiest to see side by side. Take two readers scoring the same paragraph on unfamiliar subject matter. A moderation account stops at: readers with less topic knowledge weight syntactic complexity more heavily. True, but it explains nothing about why that weighting happens. It could, for all the model says, be an arbitrary preference. A mediation account fills that gap: the reader is not choosing to favour complexity. They are compensating for a working-memory bottleneck they may not even be aware of, reaching for the one cue still available to them once content-based judgment becomes too costly to sustain. CMEJ needs the moderation account because it is the part an empirical study can actually test. You can measure an interaction, but you cannot directly measure a chain of unobserved cognitive states. However, CMEJ needs the mediation account because it is the part that makes the model a psychological explanation rather than a statistical curiosity. A model that reports the interaction without the mechanism behind it has described a pattern. CMEJ is built to explain one.

Figure 2 separates two claims that are easy to blur into one. Panel A is a statistical statement: topic knowledge does not predict evaluative judgment on its own, it changes how strongly syntactic complexity predicts it. Strip the model down to numbers and this is the part a regression would recover, an interaction term. Panel B says why the interaction exists, not just that it exists, by tracing a chain of cognitive events. Limited topic knowledge drives up processing effort, that effort eats into the capacity a reader would otherwise spend on the argument itself, and what remains is a reader reaching for the nearest available cue (length, density, the shape of the sentence) as a stand-in for a judgment they no longer have the resources to make directly.

4. Positioning CMEJ Within Existing Theoretical Accounts

Three frameworks have shaped the argument developed so far, and it is worth pausing on why none of them, individually, was ever positioned to explain evaluative judgment. This is not a shortcoming to be corrected retroactively. Each framework was built to answer a different question, from within a different research tradition, at a different point in the processing sequence. VanPatten's concern was psycholinguistic: *How does a learner's parser convert linguistic form into meaning under real-time constraints?* Sweller's concern was cognitive-architectural: *What happens to instruction and comprehension when working memory, a resource with well-documented and fixed limits, is exceeded?* Schmidt's concern was attentional: *What conditions must be met for a linguistic feature to cross from unregistered input into a state available for further cognitive work?* None of the three asked what a reader does once comprehension, capacity, and attention have already done their part and a single evaluative verdict is still required. Table 1 sets this out systematically, mapping each framework's explanatory territory against the point at which that territory ends.

Table 1. Explanatory Scope of Prior Frameworks Relative to the CMEJ

Framework	Explains	Stops short of	CMEJ's extension
Input Processing Theory (VanPatten, 2015)	Which elements of a sentence a learner extracts during comprehension, and why some forms resist uptake despite repeated exposure	What the learner does with that extracted meaning once comprehension is complete	Treats successful or constrained extraction as the precondition that determines which of the two evaluative pathways is activated
Cognitive Load Theory (Sweller et al., 2011)	The conditions under which processing becomes effortful once intrinsic, extraneous, and germane load exceed working-memory capacity	Which direction that effort pushes a subsequent quality judgment, towards penalising or rewarding the text	Treats load as the mechanism that reroutes evaluation away from content-based appraisal and towards salience-based heuristics
Noticing Hypothesis (Schmidt, 2001)	Which features of the input are perceptually or instructionally salient enough to cross the threshold into conscious registration	How multiple simultaneously noticed features are weighed, prioritised, or reconciled into a single judgment	Treats syntactic complexity specifically as the noticed feature most likely to be recruited once content-based judgment is no longer affordable

A pattern becomes visible once the three are read against each other rather than in isolation. It is also a pattern none of the original authors had reason to name, since each was working within a separate literature and towards a separate explanatory goal. VanPatten's learner completes an act of extraction, and the theory, by design, has nothing further to say. Sweller's reader reaches the limit of working-memory capacity, and the theory, again by design, does not specify what a reader does once that limit is reached, only that performance degrades. Schmidt's reader registers a salient feature in conscious awareness, and the Noticing Hypothesis stops at registration, offering no account of what happens when several such features compete for the same limited attentional resource. Three independent silences, arising from three independent research programs, converge on the same unanswered question: given a reader who has extracted some meaning, exhausted some capacity, and noticed some feature, what judgment do they actually produce? It is worth being precise about what CMEJ is not doing here, because the temptation with any integrative model is to overstate the novelty of the integration itself. Simply stacking Input Processing, Cognitive Load, and Noticing together, as though evaluative judgment was merely their sum, would not answer the question above. It would only restate it in three vocabularies at once. A reader who has extracted meaning under some load and noticed some feature could, on additive grounds alone, still reach almost any evaluative outcome. Nothing in the three frameworks constrains the direction of that outcome. What CMEJ contributes is not the observation that these three processes co-occur, which is close to self-evident, but a specific, falsifiable claim about how they resolve into judgment. The resolution is governed by a single variable, prior topic knowledge, and its effect is continuous and directional rather than threshold-based. This is the claim an empirical study is positioned to test directly: whether a reader's evaluative output does, in fact, move predictably along the continuum CMEJ predicts, from content-anchored judgment at high topic familiarity to form-anchored judgment at low topic familiarity.

4.1 CMEJ and Empirical Rater Cognition Research

The theoretical account developed above should also be positioned against a separate, empirically oriented literature: research on rater cognition in L2 writing assessment. This literature has documented, largely independently of the acquisition-oriented frameworks discussed above, that raters do not apply scoring criteria uniformly. Vaughan (1991) identified distinct reading styles among raters scoring the same set of essays, including a grammar-oriented style in which judgment anchored heavily on surface-level correctness. Wolfe, Kao, and Ranney (1998) extended this line of inquiry by comparing scorers at different proficiency levels, finding that more proficient scorers tended to process essays holistically and invoke general, rubric-aligned criteria, whereas less proficient scorers processed essays in a more interrupted, bottom-up manner, relying more heavily on self-generated, specific observations not explicitly named in the scoring rubric. Eckes (2012), working within a many-facet Rasch measurement framework, showed that raters can be reliably classified into stable types according to which scoring criteria they weight most heavily, and that this classification predicts patterns of rater bias in live scoring sessions. More recent work extends this picture. Eckes and Jin (2024), tracking halo tendencies across four successive high-stakes writing assessments, found that illusory halo, bias attributable to judgmental distortion rather than genuine overlap between criteria, was present in a meaningful minority of raters and showed evidence of both stability and change over repeated occasions, indicating that the tendency to conflate criteria is neither a fixed trait of some raters nor a uniform feature of the rating context. Trüb, Möller, Lohmann, Jansen, and Keller (2025) add a complementary finding at the level of judgment accuracy rather than bias: across nine assessment criteria in EFL writing, judgment accuracy varied significantly by criterion, with linguistic features assessed more accurately than pragmatic or content-based ones, suggesting that raters' capacity to judge different textual features is itself uneven and criterion-dependent.

This body of research establishes, empirically, that something like the shift CMEJ proposes does occur: raters vary in how they weight textual features, and that variation tracks differences in expertise or interpretive orientation. What this literature has not supplied is a specified cognitive mechanism for why such variation should track topic familiarity specifically, as opposed to general rating experience or stable individual disposition. The distinction matters. Wolfe et al. (1998) and Eckes (2012) both document between-rater differences, that this rater consistently favours general criteria, that rater consistently favours specific ones, treated as comparatively stable properties of the rater. CMEJ, by contrast, predicts a within-rater, topic-contingent shift: the same rater, applying the same criteria, is predicted to weight syntactic complexity differently depending on their familiarity with the specific topic of the essay in front of them, not their general scoring experience. This is a narrower and more specific claim than the rater-type literature has tested, and it is falsifiable in a way the broader construct of rater proficiency is not, since it requires manipulating topic familiarity while holding the rater and the rating criteria constant.

4.2 Instrument Validation and Preliminary Pilot Observations

This paper is part of a larger research project investigating how EFL readers' prior topic knowledge shapes their reliance on syntactic complexity when judging academic writing quality; the Compensatory Model of Evaluative Judgment developed above was formulated within, and is currently being tested through, that project. This subsection reports on the instrument that project uses, and specifically on the pilot phase conducted to establish the instrument's clarity and feasibility before the main data collection.

Participants were five senior (fourth-year) EFL students drawn from the same purposively sampled population as the main study, selected on identical criteria: a minimum GPA of 3.00 out of 4 and a grade of (A) in an advanced grammar course, ensuring sufficient syntactic competence that reliance on surface features could not be attributed to limited proficiency. Each task consisted of two topic phases: Linguistics, focusing on the Sapir-Whorf Hypothesis, and History, focusing on the Industrial Revolution. The Linguistics passages fell within participants' own academic field, while the History passages fell outside it, giving the High- and Low-Topic-Familiarity conditions an independent, curriculum-based justification.

Stimulus paragraphs were experimentally constructed rather than naturally sourced. In the first rating task, participants rated 10 different paragraphs in each phase, assigning each a holistic quality score ranging from 1 (lowest) to 10 (highest) and providing a brief justification for their rating. Each paragraph contained one specific linguistic feature that was used extensively, allowing the influence of individual linguistic cues on quality judgments to be examined. In the second task, three versions of the same content were constructed: a Simplified version, which reduced the content to high-frequency vocabulary and minimal syntactic embedding, an Authentic Baseline, adapted from established academic sources to represent standard scholarly prose, and a Hyper-Complexed version, which manipulated the same content through exaggerated linguistic features designed to simulate what may be perceived as an illusion of academic sophistication. Participants were asked to select the version they considered to be of the highest academic quality and provide written justifications for their choice, as well as for rejecting the other two versions. Across both tasks, informational content was held constant and remained factually accurate. Participants were not informed of this, thereby minimizing potential demand characteristics and allowing any reliance on surface linguistic cues to reflect genuine evaluative behaviour. Written justifications from both tasks were thematically coded (Braun & Clarke, 2006) into content-based and structure-based categories, with a second coder independently verifying a subset of codes. Given the small sample (N = 5), the results below are descriptive and illustrative only. No inferential claims are made.

Table 2. Instrument Pilot: Procedural Notes and Preliminary Observations (N = 5)

Metric	High Topic Familiarity (Pathway One)	Low Topic Familiarity (Pathway Two)	Note
Instruction clarity	100% comprehension	100% comprehension	Procedural: no structural changes to task instructions required
Content-based justifications	80% (4/5)	0% (0/5)	Preliminary observation: shift away from content-based justification
Structure-based justifications	20% (1/5)	80% (4/5)	Preliminary observation: structural features dominated justifications
Comparative selection task (preferred the more structurally elaborate version)	0% (0/5)	80% (4/5)	Preliminary observation: same directional shift in an independent task
Mean quality rating (1–10)	M = 6.4, SD = 1.1	M = 8.6, SD = 0.9	Scale produced meaningful variation across conditions

The pilot confirmed that the instrument itself was sound. Instructions were clearly understood in both conditions, and the rating scale produced meaningful variation. The shift from content-based to structure-based justification, and the corresponding shift in

the comparative selection task, are consistent with the direction CMEJ predicts, though a sample of five cannot establish either pattern statistically or neither tests the graded interaction the model specifies. That test, together with the full inferential analysis, is reserved for the companion paper referenced above and detailed further in Section 9.

5. Opportunities and Challenges of the Compensatory Model of Evaluative Judgment

5.1 Opportunities

Positioning CMEJ against three established frameworks, as the previous section has done, addresses one kind of question: how the model relates to what already exists. A separate question remains largely unaddressed so far: what the model itself makes newly possible, and where its own explanatory reach runs out. Both questions matter for a model still at the conceptual stage, and both are taken up in turn below. CMEJ's principal opportunity is that it converts an observation the rater cognition literature has documented empirically but not yet explained mechanistically (Vaughan, 1991; Wolfe et al., 1998; Eckes, 2012), that raters weight textual features inconsistently across individuals and conditions, into a mechanism with a specified direction. Prior treatments of this inconsistency have tended to attribute it to unexplained variation between studies, rater idiosyncrasy, or genre effects considered individually. CMEJ instead proposes a single governing variable, topic familiarity, and predicts the direction of the relationship as a function of that variable, converting what had looked like an unresolved inconsistency into a testable interaction. This is the model's clearest opportunity: it does not merely add a new explanation alongside existing ones; it offers a way of reconciling findings that otherwise appear contradictory.

A second opportunity lies in the model's applied reach. Because CMEJ specifies a mechanism rather than only a correlation, it generates concrete, actionable predictions for rubric design and rater training, of the kind outlined in the implications section below, rather than remaining a purely descriptive account of what raters tend to do. A third, more modest, opportunity is disciplinary: CMEJ demonstrates one way a cognitively grounded, integrative model can be built from established second language acquisition theory without new primary data collection, offering a template other researchers might extend to related evaluative judgments, such as judgments of lexical sophistication or discourse cohesion, that raise structurally similar questions about the relationship between textual features and reader-side cognitive constraints.

5.2 Challenges and Boundary Conditions

These opportunities are real, but they should not obscure the model's present limits, several of which are conceptual rather than simply a matter of pending data collection. The most immediate limit is that CMEJ, as developed in this paper, remains untested. Its central claim, that the relationship between syntactic complexity and evaluative judgment strengthens as topic familiarity weakens, is stated with enough precision to be falsified, but falsification requires an empirical study this paper does not undertake. Until that test is run, CMEJ should be read as a theoretically motivated hypothesis rather than an established finding, a distinction the paper has tried to maintain throughout but that is worth restating explicitly here.

A second challenge concerns the model's treatment of topic familiarity itself. CMEJ treats familiarity as a single continuous variable running from knowledge-limited to knowledge-informed, which is analytically convenient but almost certainly a simplification. Topic familiarity in practice is not one thing. A reader may have strong general knowledge of a domain but limited familiarity with the specific claims a given essay makes, or the reverse. CMEJ, in its present form, does not distinguish between these kinds of familiarity, and it is an open question whether they compensate for structural complexity in the same way or to the same degree.
A third challenge concerns the model's implicit assumption that syntactic complexity is the feature most likely to be recruited under Pathway Two. This assumption was defended earlier on the grounds that complexity is perceptually salient independent of content knowledge, and that defense still holds, but it does not rule out competition from other salient features, lexical sophistication, cohesive markers, or surface-level correctness, that might be recruited instead of, or alongside, complexity, depending on the rater or the text. CMEJ, as currently specified, treats complexity as the default heuristic rather than one candidate among several, and a more complete account would need to specify why complexity should be favoured over these alternatives rather than simply asserting that it is.

A final, related challenge is cultural and disciplinary variation in what counts as a marker of quality in the first place. CMEJ was developed with reference to EFL academic writing evaluated against expectations shaped substantially by Anglophone academic norms. Whether the same mechanism holds in writing cultures with different stylistic conventions remains unaddressed, a boundary condition, not an incidental omission.

6. Discussion: Addressing the Research Questions

The argument developed in Sections 1 through 5 can now be related directly back to the three questions that opened this paper. The first question asked why existing second language acquisition frameworks fail to account for evaluative judgment specifically, despite each explaining a related stage of language processing. Sections 1.1 through 1.4 answered this by showing that Input Processing Theory, Cognitive Load Theory, and the Noticing Hypothesis were each developed to explain a precondition for

evaluation, extraction, capacity, and attention, without addressing the evaluative act itself, and that the literature on syntactic complexity as a construct has been developed almost entirely to describe production rather than reception. None of the four literatures reviewed, individually, was positioned to answer the question this paper asks.

The second question, together with its sub-question (2.1) concerning how that mechanism varies with the reader's prior topic knowledge, asked what cognitive mechanism governs the relationship between syntactic complexity and perceived writing quality. Sections 2 and 3 answered both directly: CMEJ proposes that topic knowledge governs a compensatory shift between two evaluative pathways, formalised as both a statistical moderation and a cognitive mediation (Section 3.1), and that this shift is graded and directional rather than threshold-based.

The third question asked what theoretical and pedagogical consequences follow from modelling this relationship as compensatory rather than fixed. Sections 4 and 5 answered this by positioning CMEJ against both the acquisition-oriented frameworks it builds on and the empirical rater cognition literature that has documented related, but conceptually distinct, patterns, before turning to what the model makes newly possible and where its own reach runs out. The specific consequences for researchers, instructors, and assessment designers are developed in the section that follows. Read together, these answers support the paper's central claim: evaluative judgment is not a residual byproduct of acquisition-related processing, but a distinct cognitive act that draws on the same limited-capacity system, governed by a mechanism this paper has now specified precisely enough to test.

7. Implications

7.1 Implications for SLA and Writing Assessment Researchers

CMEJ's contribution to the broader literature is not a new empirical finding but a reframing of what evaluative judgment is taken to be a function of. Existing accounts of syntactic complexity's relationship to perceived writing quality have largely treated that relationship as a property of the text, complexity as inherently more or less valued, independent of who is reading. CMEJ relocates the relationship into the reader, arguing that the same textual feature carries different evidential weight depending on the cognitive resources the reader brings to the encounter. This shifts the relevant research question. Rather than asking whether syntactic complexity predicts perceived quality, the CMEJ framework asks under what conditions it does, and predicts that topic familiarity is the variable governing that condition. This is a testable, falsifiable claim, and one that existing frameworks, built for a different explanatory purpose, were not positioned to generate.

This reframing also carries a direct methodological implication for how future studies of syntactic complexity are designed. Studies that treat topic familiarity as a nuisance variable, holding it constant, matching participants, or ignoring it as irrelevant to a text-level property, are, on CMEJ's account, controlling away the very condition that determines the sign of the relationship they are trying to measure. A study conducted entirely with topic-familiar raters and one conducted entirely with topic-unfamiliar raters would, under CMEJ, be expected to report different, and potentially opposite, relationships between complexity and perceived quality, not because either study erred, but because each sampled only one side of a moderated relationship. This offers one candidate explanation for why the existing literature on complexity and perceived quality has proven so difficult to reconcile across studies: inconsistent findings may reflect unmeasured variation in participants' topic familiarity relative to the texts they rated, rather than inconsistency in the phenomenon itself. The same logic extends beyond syntactic complexity specifically; wherever a textual feature's relationship to perceived quality has proven similarly inconsistent, lexical sophistication and cohesion markers among the more obvious candidates, CMEJ's reframing suggests the inconsistency itself may be diagnostic of an unmeasured reader-side moderator, rather than evidence that the feature bears no reliable relationship to quality at all.

7.2 Implications for EFL Instructors, Raters, and Assessment Designers

Evaluative judgment, as CMEJ frames it, is never a judgment of a text in the abstract. It is a judgment of specific textual features, weighted against one another under real cognitive constraints. Syntactic complexity is the feature this paper has centred throughout, not because it is the only such feature, but because it is the one most likely to be both perceptually salient and structurally visible independent of subject-matter expertise (Lu, 2010). A rater can register sentence length and clausal density without understanding the essay's content at all, in a way that is far harder to do for, say, argument coherence or evidential accuracy. This is precisely what makes complexity the feature evaluative judgment defaults to once content-based appraisal becomes too costly to sustain. It is also what gives the model direct consequences for how academic writing is evaluated in EFL contexts, both by teachers and by the rating instruments teachers rely on. If evaluative judgment shifts towards form-based heuristics precisely when a rater's topic knowledge is limited, then rater training and rubric design face a specific, previously unnamed risk. Raters are least equipped to resist superficial cues exactly when they are being asked to evaluate writing on unfamiliar subject matter, which is a routine condition in EFL classrooms where a single teacher rates essays across a wide range of student-chosen topics. Empirical work on rater behaviour lends independent support to this concern. Kim (2020) documents a halo effect in L2 writing assessment, in which raters' judgments on one criterion bleed into their judgments on ostensibly separate criteria.

This is an effect the study links to raters' inability to fully disentangle overlapping textual signals under standard rating conditions. CMEJ offers a cognitive account of why such bleed-through would be more likely under exactly the conditions rating scales rarely control for: rater unfamiliarity with the essay's topic. Two practical adjustments follow from this. First, rating rubrics could be designed to require raters to justify a syntactic-complexity-based judgment in writing, converting an automatic heuristic into a deliberate decision the rater is accountable for, a step that CMEJ predicts would be disproportionately useful precisely when rater topic knowledge is low. Second, rater training programs could incorporate topic-familiarity checks prior to rating, flagging essays on unfamiliar subject matter for double-rating or for assignment to a rater with relevant background knowledge, rather than treating rater assignment as topic-agnostic.

7.3 Implications for Assessment Validity

The pattern CMEJ predicts is not only a matter of cognitive interest; it is a validity concern in Messick's (1995) sense. Messick's account of construct-irrelevant variance describes exactly the risk CMEJ specifies a mechanism for: score variance attributable not to the construct an assessment is meant to measure, here, writing quality, but to something systematically unrelated to it, here, a rater's own topic familiarity with a given essay. If two raters of comparable training and severity score the same essay differently because one happens to know the topic and the other does not, and if that difference operates through a shift toward structural heuristics rather than through a difference in genuine severity, the resulting score variance is construct-irrelevant by definition, however reliable it might appear across repeated ratings by the same rater.

Kane's (2013) argument-based approach to validation sharpens the practical stakes of this concern. Under Kane's framework, a scoring inference, the claim that an assigned score reflects the writer's underlying ability rather than incidental features of the rating occasion, requires evidence that raters apply criteria consistently across conditions that should not matter to the construct. Topic familiarity is not part of what an academic-writing rubric claims to measure; if CMEJ's prediction holds, it is nonetheless doing measurable work in determining the score. This would constitute a specific, previously unarticulated threat to the scoring inference in Kane's chain, one that recent empirical work has begun to substantiate directly. Trüb, Lohmann, Möller, and Keller (2026) found that ratings of syntactic complexity itself were significantly predicted by unrelated criteria, such as vocabulary and orthography, rather than by complexity as independently judged, and Lohmann et al. (2025) similarly documented halo, frame-of-reference, and strategy effects operating jointly across multiple rated criteria in essay assessment. Neither study manipulated topic familiarity directly, but both confirm, in the same assessment contexts CMEJ addresses, that raters' judgments of structural and linguistic criteria are demonstrably contaminated by factors the rubric does not intend to measure. CMEJ's contribution to this validity literature is to specify one candidate source of that contamination, and to predict when it should be strongest: precisely when a rater's topic knowledge is lowest.

8. Limitations of the Current Paper

Beyond the boundary conditions identified for the model itself (Section 5.2), certain limitations attach to this paper as a piece of theoretical work. The synthesis offered here draws on three frameworks selected for their prominence and established standing within second language acquisition research. Other cognitively oriented accounts of language processing were not incorporated, and the argument does not claim that these three are the only frameworks that bear on evaluative judgment, only that they are individually silent on it in the way documented in Sections 1 and 2. This paper also does not operationalise its central constructs: topic familiarity and syntactic complexity are treated throughout as theoretical variables rather than measured ones, and translating them into instruments a rating study could use is identified here as necessary but is not undertaken. Finally, the account is developed with specific reference to EFL academic writing evaluated by human raters. Its extension to other genres, to automated scoring systems, or to first-language evaluative judgment, is neither claimed nor tested.

9. Future Directions

Three directions follow directly from the limitations above. The most immediate is empirical: testing whether the interaction CMEJ predicts between topic familiarity and reliance on syntactic complexity holds under direct measurement, using a rating design that manipulates topic familiarity while holding text and rater constant. A second direction is instrumental: developing and validating the measures, a topic familiarity scale and a graded index of syntactic complexity, that such a test would require, since neither presently exists in a form suited to this specific model. Yasuda's (2026) critique of purely quantitative complexity indices is a useful caution for this work, since a graded index built for CMEJ would need to capture perceived structural salience, not only counted structures. A third direction is extension: examining whether the same compensatory mechanism generalises to evaluative judgments beyond syntactic complexity, such as lexical sophistication or cohesion, and whether it holds across the cultural and disciplinary boundary conditions identified in Section 5.2. None of these directions is pursued here, as each defines a specific, falsifiable next step rather than a general call for further research.

10. Conclusion

This paper began from a narrow observation. Three well-established frameworks in second language acquisition (Input Processing Theory, Cognitive Load Theory, and the Noticing Hypothesis) each explain a stage of how EFL learners process written input, and none explains what those learners do once a qualitative judgment about that input is required. The Compensatory Model of Evaluative Judgment addresses that gap directly, proposing that evaluative judgment is governed by a single, cognitively grounded mechanism. As topic familiarity decreases, readers compensate for diminished content-based appraisal capacity by relying increasingly on syntactic complexity as a heuristic proxy for quality. The model formalises this mechanism as both a statistical moderation and a cognitive mediation, giving it a form that is simultaneously testable and explanatory, a distinction existing accounts, built for other purposes, were not positioned to draw. Positioned within the broader literature, CMEJ does not compete with the frameworks it builds on, but it extends them past the point at which each was designed to stop. Whether the predicted interaction between topic knowledge and syntactic complexity holds under direct measurement, and whether the suggested pedagogical adjustments measurably reduce the bias CMEJ predicts, remain outside this paper's scope but follow directly from it.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381
[2] Eckes, T., & Jin, K.-Y. (2024). Examining illusory halo effects across successive writing assessments: An issue of stability and change. *Journal of Applied Measurement*, *25*(1/2), 75–95.
[3] Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, *10*(1), 18–26. https://doi.org/10.1007/s13162-020-00161-0
[4] 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
[5] Kim, H. (2020). Effects of rating criteria order on the halo effect in L2 writing assessment: A many-facet Rasch measurement analysis. *Language Testing in Asia*, *10*(16), 1–23. https://doi.org/10.1186/s40468-020-00115-0
[6] Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, *43*(4), 781–812.
[7] Lohmann, J. F., Lötscher, F., Keller, S. D., Fleckenstein, J., Jansen, T., & Möller, J. (2025). Testing teacher judgment effects comprehensively: Accuracy, halo, frame of reference, strategy, and personality effects in the assessment of student essays. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000969
[8] 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
[9] MacInnis, D. J. (2011). A framework for conceptual contributions in marketing. *Journal of Marketing*, *75*(4), 136–154.
[10] Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741
[11] Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492
[12] Schmidt, R. W. (2001). *Attention and awareness in foreign language learning*. University of Hawai'i Press.
[13] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.
[14] Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory* (Vol. 1). Springer.
[15] Trüb, R., Lohmann, J., Möller, J., & Keller, S. D. (2026). Pursuing fair writing assessment: Halo effects in primary school foreign language writing in grade six. *Assessing Writing*, *68*, Article 101036. https://doi.org/10.1016/j.asw.2026.101036
[16] Trüb, R., Möller, J., Lohmann, J., Jansen, T., & Keller, S. D. (2025). Judgment accuracy in primary school EFL writing assessment: Do text characteristics matter? *Assessing Writing*, *66*, Article 100957. https://doi.org/10.1016/j.asw.2025.100957
[17] VanPatten, B. (2015). Input processing in adult second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 113–134). Routledge.
[18] Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Ablex.
[19] Wolfe, E. W., Kao, C.-W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465–492. https://doi.org/10.1177/0741088398015004002
[20] Wu, W., & Qin, J. (2025). Creativity, topic familiarity, and L2 performance in story continuation writing tasks. *International Journal of Applied Linguistics*, *35*, 1572–1584. https://doi.org/10.1111/ijal.12737
[21] Yasuda, S. (2024). Does "more complexity" equal "better writing"? Investigating the relationship between form-based complexity and meaning-based complexity in high school EFL learners' argumentative writing. *Assessing Writing*, *61*, 1–21.
[22] Yasuda, S. (2026). Beyond quantification: Rethinking the measurement of syntactic complexity in L2 writing. *International Journal of Applied Linguistics*. Advance online publication. https://doi.org/10.1111/ijal.70175