PPLING
International Applied Linguistics Research Conference
https://doi.org/10.29140/9781763711686-7

What the Survey Could Not See: Feature Coverage in an AI-Assisted EFL Reading Instrument

Le Thi Thach-Thao[1], **Thanh-Nhan Le**[2]
[1]Victoria University of Wellington, New Zealand
[2]Graduate Institute of Cross-Cultural Studies, Fu Jen Catholic University, Taiwan

Abstract

Computer-assisted language learning (CALL) acceptance surveys operationalise a platform as a fixed inventory of features; that inventory is a design decision reporting rarely tests. This pilot reports one consequence. Two intact English as a foreign language (EFL) classes (Vietnam, n = 15; Taiwan, n = 30) completed teacher-assigned reading on a shared Gemini-built platform and a 44-item six-point survey measuring three platform features: timer, explanations, and vocabulary list. No subscale measured a text-highlighting function, and no item or interview question named it. Four of five interviewees raised it unprompted, two calling it the feature they valued most; 6 of 15 Vietnamese open responses volunteered it against 1 of 30 Taiwanese (Fisher's exact, p = .003). Because reliability and fit statistics evaluate only the items present, coverage failures of this kind are invisible from inside an instrument, and only the qualitative strand detected this one. Enumerating available affordances before administration is a low-cost safeguard.

Keywords: content validity; CALL instrument design; text highlighting; L2 reading; mixed methods

E-mail: jessie.thachthao@gmail.com, 413038052@cloud.fju.edu.tw

In Y. Wang, J. Matthews, & B. Zou (Eds.), *Proceedings of the 2nd International Applied Linguistics Research Conference, 2026* (pp. 48–54). Castledown.

Background

Technology acceptance instruments in computer-assisted language learning (CALL) translate a platform into a set of rateable features. Researchers decide which features enter that set before data collection and seldom defend the choice in print, yet it fixes the limit of what any subsequent model can represent. The dominant frameworks are silent here. Acceptance and fit models specify relations among usefulness, effort, enjoyment, and task-technology correspondence (Davis, 1989; Goodhue & Thompson, 1995; Venkatesh et al., 2003; Venkatesh et al., 2012), and cognitive fit theory specifies how a representation should match a task (Vessey, 1991); none prescribes how researchers should enumerate that inventory. Its adequacy is a content-validity question: whether the item set represents the domain it claims to cover (Haynes et al., 1995; Messick, 1995).

Text highlighting makes an instructive test case, because it is neither novel nor marginal. Dunlosky et al. (2013) reviewed it because students report relying on it heavily, and rated its utility low. Meta-analysis since finds that learner-generated highlighting improves memory for text but not reliably comprehension (Ponce et al., 2022), while Yue et al. (2015) trace its persistence to learners' metacognitive beliefs. In CALL, De Ridder (2000) examined its effect on attention allocation during reading; Simard (2009) links it to textual enhancement in second language (L2) acquisition; and digital annotation systems built around it support digital and L2 reading performance (Chen & Chen, 2014; Chen et al., 2014). A heavily used strategy of contested benefit is exactly what an acceptance instrument should capture, not omit.

This paper reports a single finding from a pilot deployment: learners attributed value to a platform affordance that the instrument did not measure, and only the qualitative strand detected the omission.

Method

Design. An explanatory sequential pilot. Two intact classes completed teacher-assigned reading tasks on a shared Gemini-built platform, each across one term: a college-preparatory IELTS course in Vietnam (n = 15) and an advanced freshman English course in Taiwan (n = 30). We chose the two sites for access rather than representativeness: the Taiwanese class was our own, and we had a collaborator in Vietnam. We compare them because both answered the same instrument about the same platform; since they also differed in country, teacher, course type, and the timing of feature access, the comparison shows that salience diverged, not why. Of 49 responses returned,

45 were complete; we excluded incomplete cases listwise. We grouped by administration wave rather than first-language (L1) self-report, because free-text L1 responses were inconsistent.

Platform. The platform was not a commercial product: we built it with Google Gemini as browser-based reading tasks shared as links, each scoring answers on submission. Gemini generated each task's interface when it was built; the platform itself contains no large language model (LLM), so any AI that students consulted during tasks (e.g., ChatGPT, Gemini, Claude, or Grok) was a separate tool of their choosing. The timer, explanations, and vocabulary list were built into every task. We could revise the platform at any point: text highlighting, absent at launch, was added at a learner's suggestion, and students used it at their discretion, without teacher control.

Instrument. The survey comprised 44 six-point forced-choice items with no neutral midpoint: 35 across 11 construct subscales, and 9 indexing motivational orientation. Three subscales of three items each measured platform features: timer, explanations, and vocabulary list. Branching on use of such external AI meant no respondent saw all 44, and we analyse no Likert item here. Two open items followed: the first named the three features explicitly and asked which helped most and least; the second was unrestricted, asking what most supported and most limited the respondent's reading improvement.

Interviews. We conducted five semi-structured interviews in Vietnamese with a purposive subsample of the Vietnamese class; transcribed segments ran from 31 to 46 minutes. The opening protocol block named the same three features and no others.

Coding. We screened open responses and transcripts for references to platform affordances in Vietnamese, Chinese, or English, counting a reference as learner-initiated only where neither the eliciting item nor a prior interviewer turn had named the affordance. We tested the group contrast with Fisher's exact test, two-tailed, at .05.

Results

No item, subscale, or interview question referred to text highlighting. Learners referred to it anyway.

Four of the five interviewees raised highlighting without prompting, in response to open questions about overall experience or platform familiarity; two did so at their first substantive turn. Two named it the feature they valued most: one reported requesting it mid-term and called it the feature that impressed him most (Eric, 20:58); another named it first among her impressions and returned to it at the close as the part she liked best (Mandy, 9:25,

36:40). A third described the platform through its highlighting and note functions before the interviewer had named any feature (Samantha, 45:37); a fourth volunteered it when asked only whether the tools had become familiar (Abner, 33:29). The fifth never mentioned it.

The open responses showed the same salience in the Vietnamese class, and almost none of it in the Taiwanese class. Six of 15 Vietnamese respondents volunteered highlighting, against 1 of 30 Taiwanese (Table 1 and Figure 1), a significant difference by Fisher's exact test, $p = .003$, odds ratio = 19.33. One further Taiwanese response reported being unable to use a pen to mark the text; counted as a mention, the contrast becomes 6 of 15 versus 2 of 30, $p = .011$, and the conclusion holds.

Table 1. Unprompted Mentions of Text Highlighting in Open Survey Responses, by Class

Class	Mentioned highlighting	Did not mention	Total
Vietnam (IELTS preparation)	6	9	15
Taiwan (freshman English)	1	29	30
Total	7	38	45

Note. Fisher's exact test, two-tailed, $p = .003$, odds ratio = 19.33. Neither open item named highlighting, so learners initiated every mention.

Discussion

Highlighting was absent at launch (see Method); one interviewee recalled requesting it part-way through the Vietnamese term (Eric, 20:58, 23:25). The omission is ordinary rather than careless: instruments are fixed early, deployments evolve, and nothing in a conventional analysis pipeline reports the divergence.

The consequence is construct under-representation (Haynes et al., 1995; Messick, 1995). Any model we estimate from these subscales omits one of the four affordances learners reported using, a text-marking strategy with documented effects on attention and memory for text (De Ridder, 2000; Ponce et al., 2022). No internal statistic could have flagged this. Reliability coefficients, factor structures, and fit indices evaluate the items that are present; they cannot register an item set's silence. Coverage failures are visible only from outside the instrument.

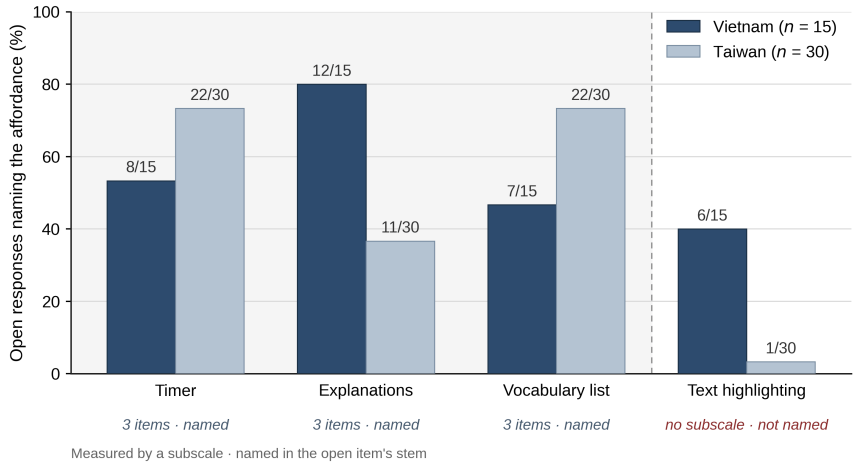

Figure 1. Open-Response Mentions of Four Platform Affordances, by Class and Measurement Status

Note. Bars show the percentage of each class's open responses naming the affordance; counts appear above. Three-item subscales measured the first three, and the open item's stem named them; nothing in the instrument named text highlighting.

That is what the qualitative strand supplied. The closed items could not have surfaced highlighting. The open items recovered it partially, in 7 of 45 cases, two of them inside the item that named only the other three; the interviews made it unavoidable, at 4 of 5, and offered an account of its absence. The qualitative strand did not elaborate a quantitative result but contradicted a premise of the quantitative design, which is integration in the initiation rather than the complementarity sense (Greene et al., 1989).

The between-class asymmetry warrants restraint. Both cohorts had the highlighting function available, so access alone does not explain the gap. The design cannot separate the remaining explanations: novelty, since the Vietnamese class, which used the platform first, gained the feature mid-term and so met it as a change, whereas the Taiwanese class had it from the start; task type, since keyword scanning is central to IELTS reading and less so to freshman English; and instructional framing by different teachers. No cultural interpretation is warranted, and we offer none. The asymmetry matters because feature salience is not a stable property of a platform, which is why a shared instrument cannot assume a shared feature set.

The practical implication is narrow. Before administering a platform acceptance survey, researchers should enumerate and log the affordances available to each cohort, and include one open item that invites feature-level com-

ment without naming any feature. The revised instrument adds a highlighting subscale and a feature-exposure field.

Limitations. This is a pilot with a small qualitative sample from one site, and the interviewer differed for one participant. Mention counts index salience, not use: learners who highlighted silently are invisible. Highlighting is a representation-manipulation affordance, squarely within what cognitive fit theory addresses (Vessey, 1991), yet outside what the instrument asked about.

Conclusion

A survey can be internally sound and still look in the wrong place. In this pilot the affordance learners raised first and valued most was the one no subscale measured, and we detected the gap only because learners had room to speak outside the item set. Enumerating available affordances, and leaving one question unframed by them, is a low-cost safeguard against measuring a platform learners did not use.

References

Chen, C.-M., & Chen, F.-Y. (2014). Enhancing digital reading performance with a collaborative reading annotation system. *Computers & Education, 77*, 67–81. https://doi.org/10.1016/j.compedu.2014.04.010

Chen, C.-M., Wang, J.-Y., & Chen, Y.-C. (2014). Facilitating English-language reading performance by a digital reading annotation system with self-regulated learning mechanisms. *Educational Technology & Society, 17*(1), 102–114. https://www.jstor.org/stable/jeductechsoci.17.1.102

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

De Ridder, I. (2000). Are we conditioned to follow links? Highlights in CALL materials and their impact on the reading process. *Computer Assisted Language Learning, 13*(2), 183–195. https://doi.org/10.1076/0958-8221(200004)13:2;1-D;FT183

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213–236. https://doi.org/

10.2307/249689

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255–274. https://doi.org/10.3102/01623737011003255

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

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

Ponce, H. R., Mayer, R. E., & Méndez, E. E. (2022). Effects of learner-generated highlighting and instructor-provided highlighting on learning from text: A meta-analysis. *Educational Psychology Review, 34*(2), 989–1024. https://doi.org/10.1007/s10648-021-09654-1

Simard, D. (2009). Differential effects of textual enhancement formats on intake. *System, 37*(1), 124–135. https://doi.org/10.1016/j.system.2008.06.005

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly, 36*(1), 157–178. https://doi.org/10.2307/41410412

Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences, 22*(2), 219–240. https://doi.org/10.1111/j.1540-5915.1991.tb00344.x

Yue, C. L., Storm, B. C., Kornell, N., & Bjork, E. L. (2015). Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review, 27*(1), 69–78. https://doi.org/10.1007/s10648-014-9277-z