ReCALL (2026), 1–19
doi:10.1017/S0958344026100561

RESEARCH ARTICLE

The effectiveness of computerized dynamic assessment in improving L2 performance: A three-level meta-analysis

Qi Lu
Ocean University of China, China (luqi@stu.ouc.edu.cn)

Mengqi Chen
Ocean University of China, China (chenmengqi2025@stu.ouc.edu.cn)

Lianrui Yang
Ocean University of China, China (lryang@ouc.edu.cn)

Shaofeng Li
The Hong Kong Polytechnic University, Hong Kong (shaofeng.li@polyu.edu.hk)

Ying Chen
Ocean University of China, China (cy@ouc.edu.cn)

Abstract

The growing body of research on the effects of computerized dynamic assessment (C-DA) on second language (L2) learning underscores the need for a comprehensive research synthesis to identify future research directions and inform the application of C-DA in L2 educational contexts. This meta-analysis employed a three-level modeling approach to examine the effectiveness of C-DA in improving L2 learners' performance. It synthesized 27 effect sizes from cake format designs, in which mediation is embedded within the test sequence, and 24 effect sizes from sandwich format designs, where mediation is delivered between a pretest and a posttest, across 35 studies published between 2000 and May 27, 2025. This study also investigated the key variables that moderate C-DA effectiveness. Findings reveal large, significant positive effects of both the cake and sandwich formats on L2 performance improvement (cake format: $g = 2.120$, $p < .001$; sandwich format: $g = 1.676$, $p < .001$), with the cake format tending to yield larger effect sizes. This may be because the cake format captures gains during mediation, whereas the sandwich format reflects post-mediation outcomes. Moderator analyses show that the number of items, test content, and learners' first language affect C-DA effectiveness in promoting L2 performance. Drawing on the synthesized findings, this study contributes to theoretical, methodological, and technological understandings of C-DA and offers suggestions for future research in this domain.

Keywords: computerized dynamic assessment; learning-promoting assessment; meta-analysis; sociocultural theory

Corresponding authors: Lianrui Yang; Email: lryang@ouc.edu.cn; Ying Chen; Email: cy@ouc.edu.cn

Cite this article: Lu, Q., Chen, M., Yang, L., Li, S. & Chen, Y. (2026). The effectiveness of computerized dynamic assessment in improving L2 performance: A three-level meta-analysis. *ReCALL* FirstView, 1–19. https://doi.org/10.1017/S0958344026100561

© The Author(s), 2026. Published by Cambridge University Press on behalf of EUROCALL, the European Association for Computer-Assisted Language Learning. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

1. Introduction

In second language (L2) learning, dynamic assessment (DA) is a pedagogical approach that dialectically integrates assessment and instruction to both evaluate and foster learner development (Poehner, 2007). In this interactive process, a mediator (e.g., an instructor or assessor) offers needs-contingent mediation – generally referred to as "intervention" and operationalized as prompts or hints by DA researchers (Sternberg & Grigorenko, 2002) – when the learner struggles. In doing so, the mediator assesses learners' current and emerging abilities through their responses while simultaneously promoting learning. Computerized dynamic assessment (C-DA) is a digital adaptation of DA, in which computers deliver standardized mediation whenever learners answer incorrectly (Poehner & Lantolf, 2013). Through quantified intervention, C-DA can probe L2 learner's potential to surpass existing abilities, with a primary focus on fostering development. Despite variation in implementation across contexts, C-DA is fundamentally characterized by two core features: the integration of assessment and mediation, and the goal of diagnosing and promoting learners' developmental potential through needs-contingent mediation.

With technological advances, research on C-DA in L2 learning has expanded rapidly (e.g., Poehner et al., 2015; Teo, 2012; Vakili & Ebadi, 2022), generally confirming its role in promoting learners' performance. However, these studies varied in terms of methodological operations such as test content, test format, and number of items, to name a few, which may influence C-DA effectiveness. Hence, to identify research orientations for L2 researchers and promote the application of C-DA in the educational context, the present study employs a meta-analysis to investigate the effect of C-DA on promoting L2 learners' performance and examine mediation-related, assessment-related, learner-related, and methodological moderators shaping its effectiveness.

2. Background

2.1. Theoretical background of DA

DA is grounded in Vygotsky's sociocultural theory (SCT), which dates back to the 1930s and was later made widely available through the English translation (Vygotsky, 1978). It was originally developed in the context of child development and educational psychology. Central to SCT is the notion of the zone of proximal development (ZPD), in which learning occurs through collaborative dialogue, mediation, and expert scaffolding. According to SCT, learners' independent performance reveals their zones of actual development (ZADs), whereas their responsiveness to assistance or mediation reveals their ZPDs. Together, the ZAD and ZPD constitute a dialectical unity of being and becoming – that is, what currently is and "what is not yet but is about to become" (Valsiner, 2012, p. 11). While traditional static assessments focus on learners' current abilities, DA emphasizes their developmental potential by examining how performance changes under mediation, thereby aligning assessment with the fundamental educational goal of fostering development.

The effectiveness of DA lies in the mediation it offers. Vygotsky conceptualized mediation as a driving force behind intellectual development, which, if appropriately sensitive to a learner's abilities, opens the ZPD, thereby triggering the internal developmental processes. Building on this premise, Campione et al. (1984) proposed the graduated prompts approach (GPA), which systematically organizes mediational prompts from implicit to explicit. Their work illustrates how DA was initially concerned with diagnosing learning potential and individual differences in children's cognitive development.

The application of DA to L2 learning emerged later, with Lantolf first introducing SCT as a theoretical framework in the 1990s to reframe L2 learning as a socially mediated, culturally embedded process. Aljaafreh and Lantolf (1994) provided an early empirical application of DA to adult L2 learning. They demonstrated that development is an uneven and mediation-dependent

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

process. Importantly, they showed that learners' progress is reflected not only in improved performance but also in the transition from explicit to implicit mediation. They further proposed a 13-level regulatory scale, which has since served as a foundational framework for subsequent DA research in L2 learning.

Within SCT, microgenesis concerns the short-term formation or unfolding of psychological processes, allowing researchers to observe development over relatively short periods of time (Wertsch, 1985). This general developmental concept was later taken up in L2 DA research to provide key evidence of how mediation is appropriately aligned with learners' emergent abilities (Poehner, 2011). Belz and Kinginger (2003) characterize the microgenetic method as "the observation of skill acquisition during a learning event" (p. 594), allowing researchers to closely examine specific instances of development.

C-DA was subsequently developed as a technology-mediated extension of DA. While sharing the same sociocultural foundation, C-DA delivers mediation through computerized procedures, making it possible to provide standardized prompts to larger groups of learners (Poehner & Lantolf, 2013). Thus, C-DA can be understood as a later development within the DA tradition, extending earlier ZPD-based and L2 DA work into technology-supported assessment contexts.

2.2. Application of C-DA in L2 learning

Mediation is central to C-DA, facilitating learning and predicting learners' ZPD, leading researchers to explore various features of mediation – including its timing, approach, mode of design, and quantity – to determine their effectiveness.

Based on when mediation is offered, Sternberg and Grigorenko (2002) proposed two general DA formats: the "cake" format and the "sandwich" format. In the cake format, mediation is embedded directly into the assessment as interventional prompts. This design yields an actual score (reflecting independent performance), a mediated score (reflecting supported performance), and a learning potential score (capturing the capacity to internalize mediation). It provides evidence of learners' performance improvement during mediation. Following Poehner and Lantolf's (2013) framework for C-DA, cake format studies have typically employed single-group designs (e.g., Bakhoda & Shabani, 2019), with statistical analyses commonly involving *t* tests comparing actual and mediated scores (e.g., Kamrood et al., 2021) and correlational analyses examining the relationship between these scores (e.g., Poehner et al., 2015). Moreover, transfer tasks are frequently incorporated to assess learners' ability to apply newly mediated knowledge in more complex or novel contexts (e.g., Izadi et al., 2023).

The sandwich format, by contrast, introduces mediation between a pretest and a posttest. Both tests follow static assessment principles, with the pretest establishing learners' baseline level and the parallel posttest capturing changes resulting from mediation. Accordingly, the sandwich format provides evidence of post-mediation improvement. While early sandwich format studies often relied on single-group designs (e.g., Teo, 2012), subsequent research has improved methodological rigor by including control groups and delayed posttests (e.g., Behbahani & Karimpour, 2025). Beigi et al. (2020) compared the effectiveness of two C-DA formats and found that learners who received the cake format C-DA outperformed their counterparts in the sandwich format C-DA group. More recently, Jin and Liu (2024) proposed a hybrid design integrating pre- and posttests into a cake format C-DA, illustrating how the two formats can be combined to provide a more comprehensive characterization of learner development.

Regarding the type of mediation provided, Lantolf and Poehner (2004) distinguished between interventionist DA and interactionist DA. Interventionist DA provides learners with a set of predesigned, standardized prompts when difficulties arise. Although standardized prompts reduce interactional flexibility, they enable quantitative descriptions of performance and facilitate cross-learner comparisons (Budoff, 1987). As a result, it is well suited to large-scale, computerized

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

implementations (Poehner & Lantolf, 2013) and has become the dominant approach in C-DA research (e.g., Meng et al., 2024; Zhang & Lu, 2019).

By contrast, interactionist DA involves open-ended, dialogic mediation tailored to individual responses, closely aligning with the learner's ZPD (Poehner & Lantolf, 2013). Advocates argue that it captures cognitive processes through real-time interaction and remains "more faithful to Vygotsky's understanding of the ZPD" (Poehner, 2005, p. 83). However, it requires skilled mediators and intensive qualitative analysis, posing challenges for current computerized systems. To address this limitation, Barabadi et al. (2018) designed a hybrid DA for English listening comprehension, combining predesigned prompts with a final teacher-delivered, interactive prompt. Their findings indicated that this hybrid design maintained assessment validity while promoting learners' listening development.

Moving beyond distinctions in mediation types, C-DA researchers have increasingly focused on optimizing mediation through prompt design. Some studies have drawn on interactionist principles to develop standardized cues (Izadi et al., 2023; Poehner et al., 2015). Although this approach is theoretically promising for improving prompt validity, empirical support remains scarce. More commonly, C-DA studies employ either strategy-based or researcher-designed prompts. Strategy-based designs organize mediation around theoretically grounded strategies relevant to the target construct (Alavi et al., 2020). In contrast, researcher-designed prompts rely on researchers' domain expertise to generate item-specific mediation. Due to its relatively low time and labor demands, this approach remains the most prevalent in C-DA research (e.g., Qin & van Compernolle, 2021).

The quantity of mediation represents another important dimension potentially influencing C-DA effectiveness and is typically operationalized through the number of prompts per item and the total number of test items. In many C-DA studies, the number of prompts (typically three to five) is constrained by the widespread use of multiple-choice questions (MCQs), which remain the dominant item format in the field. Since MCQs are often adapted from prior studies or standardized tests, opportunities for mediation are inherently restricted. To address this limitation, some researchers increase the number of response options (often to five, e.g., Poehner et al., 2015) in an attempt to reduce learner guessing and broaden mediation possibilities, whereas others retain conventional formats (typically four options, e.g., Pileh Roud & Hidri, 2021). Substantial variation also exists in the number of test items employed across studies, ranging from approximately 10 to 40. Closely related to these design features, time constraints may further shape how learners engage with and internalize mediation. However, this factor remains largely underexplored in the literature, with few notable exceptions (e.g., Teo, 2014).

Test content may also shape the effectiveness of C-DA. A clear imbalance exists in the language skills assessed, with receptive skills such as listening and reading (e.g., Teo, 2012; Zhang & Lu, 2019) far outnumbering productive skills such as speaking and writing (e.g., Ebadi & Bashir, 2020; Vakili & Ebadi, 2022). This discrepancy largely stems from the dominance of MCQs in C-DA, which are better suited to receptive skills. Moreover, the context-rich and interactive nature of pragmatics poses additional methodological challenges (Youn & Bogorevich, 2020), resulting in a scarcity of studies on C-DA for pragmatic competence (e.g., Lu et al., in press).

Turning to learner-related moderators, two dimensions warrant particular consideration: learners' first language (L1) background and academic status. Both variables have long been recognized as influential in L2 research, and prior scholarship has repeatedly called for greater participant diversity to enhance the generalizability and ecological validity of empirical findings (Ortega, 2005). Nevertheless, existing C-DA research has focused predominantly on Persian-speaking learners, followed by Chinese-speaking learners, resulting in a relatively narrow sociolinguistic sampling frame. A similar imbalance is evident in terms of academic status, with the majority of studies targeting undergraduate and graduate university students, whereas secondary-school students remain comparatively underrepresented. As a result, empirical

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

evidence on younger learners is limited, underscoring a persistent gap between long-standing calls for participant diversity and prevailing empirical practices in C-DA research.

This study also examines study quality as a potential moderator, for two reasons. First, methodological features, such as random assignment, may systematically influence reported outcomes. Second, assessing study quality can inform improvements in primary research by identifying recurring methodological weaknesses. Accordingly, this meta-analysis evaluates study quality across four dimensions: methodological rigor, transparency, ethics, and societal contribution (Plonsky, 2024).

Although C-DA research has expanded rapidly, findings remain fragmented due to heterogeneity in mediation design, assessment-related factors, learner populations, and study quality. Moreover, no meta-analytic synthesis has systematically quantified its effectiveness or examined moderators of its effects. In addition, C-DA studies often yield multiple effect sizes from the same study, for example, across different test contents or learner groups. Treating these effect sizes as statistically independent would violate a key assumption of standard two-level meta-analysis, whereas aggregating effect sizes or selecting a single effect size from each study would discard meaningful within-study information. Therefore, a three-level meta-analytic model is particularly appropriate, as it allows for the inclusion of multiple effect sizes from the same study while appropriately modeling their statistical dependence (Hox et al., 2018). To address these gaps, this study conducts a three-level meta-analysis to explore the overall effectiveness of C-DA in promoting L2 learners' performance and to identify moderating variables that influence its effectiveness. The following two research questions guided this study:

RQ1. What is the overall effect of C-DA on L2 learners' performance?
RQ2. What are the moderator variables for the effectiveness of C-DA?

3. Methods

3.1. Identifying primary studies

The literature search was conducted on May 28, 2025. Studies published between 2000 and May 27, 2025, were considered, as the first article on C-DA was published in 2000 (Guthke & Beckmann, 2000). To identify potentially relevant studies, various methods were employed, including systematic searches of electronic databases, targeted manual searches of core journals identified through expert consultation and their recurrent citation in C-DA studies (see supplementary material S1), as well as backward and forward citation tracking. The authors searched in relevant journals and electronic databases, including Education Resources Information Center (ERIC), Google Scholar, Linguistics and Language Behavior Abstracts (LLBA), PsycINFO, and ProQuest Dissertations and Theses. The following search string was used in titles, abstracts, and keywords: ("computerized" OR "computer" OR "web" OR "computer based" OR "web based" OR "computer mediated" OR "web mediated" OR "computer assisted" OR "web assisted") AND (dynamic) AND ("assessment" OR "test"). Reference lists of included studies were screened (backward searching), and citation records were examined using Google Scholar to identify newer studies citing the included articles (forward searching).

3.2. Inclusion/exclusion criteria

To be eligible, studies had to meet the following criteria:

1. Studies focused on L2 learning.
2. Studies were written in English, as most C-DA research is published in English, and this restriction ensured reliable and consistent coding.

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

3. Studies were published in peer-reviewed journals or as PhD and master's theses. The latter were included to mitigate potential publication bias (Hunter & Schmidt, 2004), as they often contain rigorous research conducted under the supervision of field experts.
4. Studies were original empirical research; replication studies, reviews, and book reviews were excluded.
5. Studies examined the effects of C-DA on improving L2 performance.
6. For studies using the cake format, prompt quality and consistency were supported by adhering to the widely recognized GPA. These studies also reported both actual and mediated scores, employing computational methods aligned with established research paradigms in the field. In contrast, studies employing the sandwich format were required to adopt an experimental or quasi-experimental design, operationalizing C-DA effectiveness through either pretest–posttest comparisons within the same group or posttest comparisons between experimental and control groups.
7. Studies provided sufficient data to allow for the extraction of effect sizes.
8. When a single study reported more than one experiment, each experiment contributed an effect size to the meta-analysis.
9. Mediation was delivered by a computer system.

Initially, 124 potentially relevant studies were identified. After further screening and applying the inclusion criteria outlined above, 89 studies were excluded for one or more of the following reasons: 10 studies (e.g., McNeil, 2018) were not original empirical research; 42 studies (e.g., Wang, 2010) targeted domains other than L2 learning; two studies (e.g., Nicholas et al., 2024) employed non-standard scoring procedures; eight studies (e.g., Ghahderijani et al., 2021) provided mediation delivered by teachers rather than by computers; 11 studies (e.g., Ebadi & Goodarzi, 2023) did not examine the effectiveness of computerized mediation in improving learner performance; and 16 studies (e.g., Vakili & Ebadi, 2022) failed to report sufficient data for calculating Hedges's *g*. The procedure for study selection is illustrated in Figure 1.

3.3. Study coding

The development of the coding scheme underwent several rounds of revisions before the final version was established. Initially, in line with standard meta-analysis procedures, we coded bibliographic information (e.g., author, publication date, journal) and sample characteristics (e.g., learners' academic status, native language) for the selected studies. Next, we reviewed both existing research syntheses (Geng et al., 2022) and 30% of the studies included in the final analysis to identify key independent variables, dependent variables, and methodological characteristics. Based on these insights, a preliminary coding scheme was developed. As we proceeded with coding the remaining studies, the scheme was further refined. The finalized coding protocol is presented in Table 1.

To ensure coding reliability, two researchers familiar with C-DA research conducted the coding. The two researchers first studied and discussed the coding scheme. One researcher then independently coded the studies, while the second researcher cross-checked the coding. Each independent sample was assigned one effect size, and when a study reported multiple effect sizes, each set was coded independently. Intercoder reliability was calculated for categorical variables using Cohen's kappa and for continuous variables using the intraclass correlation coefficient. Across the 17 coding categories, kappa coefficients ranged from .69 to 1.00. Detailed reliability coefficients for each coding category are reported in supplementary material S2. Discrepancies were resolved through discussion between the two coders until full agreement was reached.

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

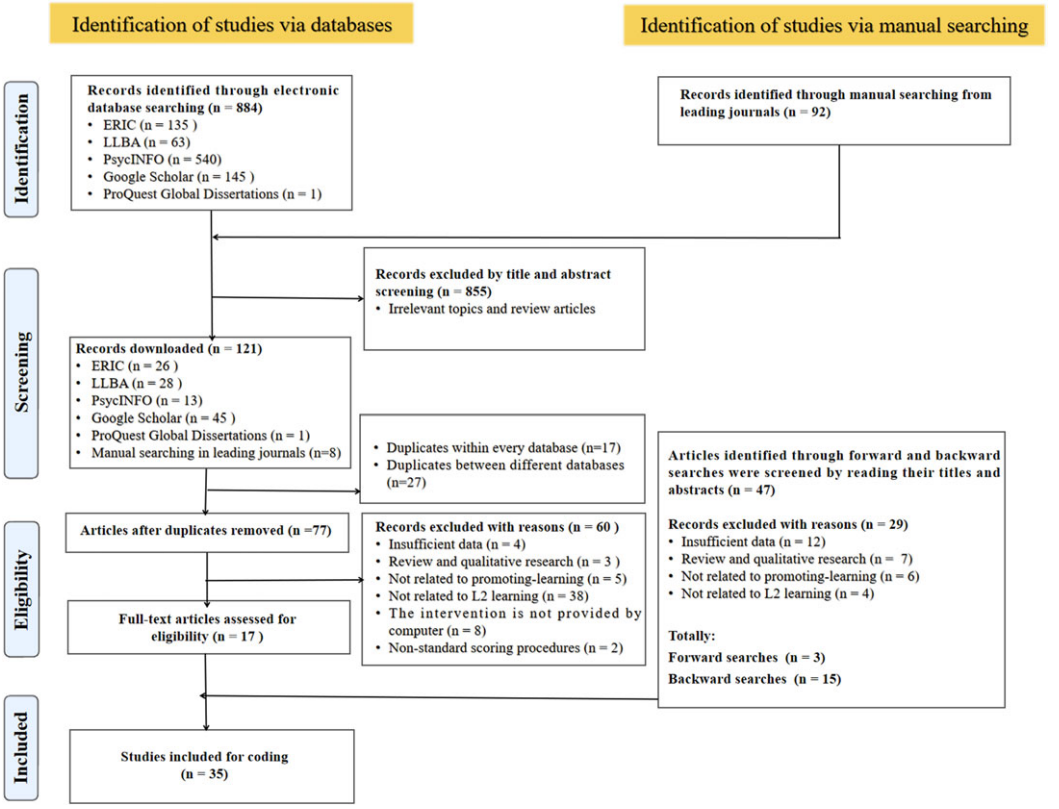


Figure 1. The procedure for study selection.

3.4. Data analysis

Hedges's *g* was chosen as the effect size measure due to the small sample sizes present in this analysis. According to Plonsky and Oswald (2014), values below 0.4 are considered small, those between 0.4 and 0.7 are moderate, and values above 0.7 indicate large effects.

Given the theoretical distinction between during-mediation and post-mediation performance improvement, cake format and sandwich format studies were treated as separate subgroups. In the cake format, effect sizes were derived from comparisons between mediated and actual scores (27 effect sizes calculated from paired-samples *t* values and sample sizes), reflecting learners' responsiveness to mediation. These outcomes were interpreted as indicators of during-mediation performance improvement. In the sandwich format, effect sizes were computed from either pretest–posttest comparisons (4 effect sizes based on sample sizes, means, and standard deviations) or experimental–control group comparisons at posttest (20 effect sizes based on posttest sample sizes, means, and standard deviations). To combine effect sizes across study designs within a sandwich format, Morris and DeShon's (2002) formulas were used to transform effect sizes into a comparable metric (see supplementary material S3 for details). These outcomes captured learners' performance after mediation and were interpreted as post-mediation performance improvement. Because the two formats operationalize different dimensions of performance improvement, effect sizes were synthesized separately.

Since several studies contributed multiple effect sizes, we implemented a three-level meta-analysis model to account for statistical dependence among estimates and to quantify within- and between-study heterogeneity (Hox et al., 2018). The model specified three variance components:

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

Table 1. Coding scheme

	Features	Descriptors
Identification	Author	Author
	Year of publication	Year of publication
	Type of publication	Article/Dissertation
Learner context	Learners' academic status	High school/Undergraduates, etc.
	Learners' L1	Chinese/English/Persian, etc.
Mediation	DA format	Cake/Sandwich/Cake+Sandwich
	DA approach	Interactionist/Interventionist
	Maximum number of prompts	3/4/5, etc.
	Prompts design method	Interactionist DA/Strategy-based design/Researcher-design, etc.
	Number of C-DA sessions	The number of C-DA sessions
Assessment context	Number of items	The number of items
	Assessment time constraints	Yes/No
	Target language	Chinese/English/French/Persian
	Test content	Grammar/Listening/Pragmatics, etc.
	Test format	MDCT/Mixed types/Open-ended questions, etc.
Study quality	Random assignment	Yes/No
	Transfer items	Yes/No
	Confidence interval	Yes/No
	Institutional approval	Yes/No
	Pedagogical implication/Theoretical contribution	Yes/No

Note. DA = dynamic assessment; C-DA = computerized dynamic assessment; MDCT = multiple-choice discourse completion test.

sampling variance at the effect size level (Level 1), clustering of effect sizes within studies (Level 2), and variability across studies (Level 3). All analyses were conducted in R using the metafor package (Viechtbauer, 2010).

To ensure the robustness of the results, leave-one-out sensitivity analyses, following Borenstein et al. (2009), were conducted. Each study was sequentially removed, and the pooled effect size was recalculated to evaluate the influence of individual studies on the overall results. These analyses identified Shabani (2022, ES2)[1] as an influential case exerting a disproportionate impact on the pooled effect size. Therefore, it was excluded from subsequent analyses. A total of 35 studies, comprising 51 effect sizes, were retained for further analysis. A three-level random-effects model was selected for the meta-analysis because it allows the true effect size to vary across studies, in contrast to fixed-effects models that assume a single common effect size (Zhou & Zhou, 2025). Given the methodological heterogeneity among the included studies, including variation in

[1]When multiple effect sizes were extracted from a single publication, they were distinguished using effect size labels (e.g., ES1, ES2). For example, "Shabani (2022, ES1)" and "Shabani (2022, ES2)" refer to different effect sizes derived from the same study. By contrast, labels such as "Ebadi et al. (2018a)" and "Ebadi et al. (2018b)" denote different publications by the same author(s) in the same year.<FN>

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

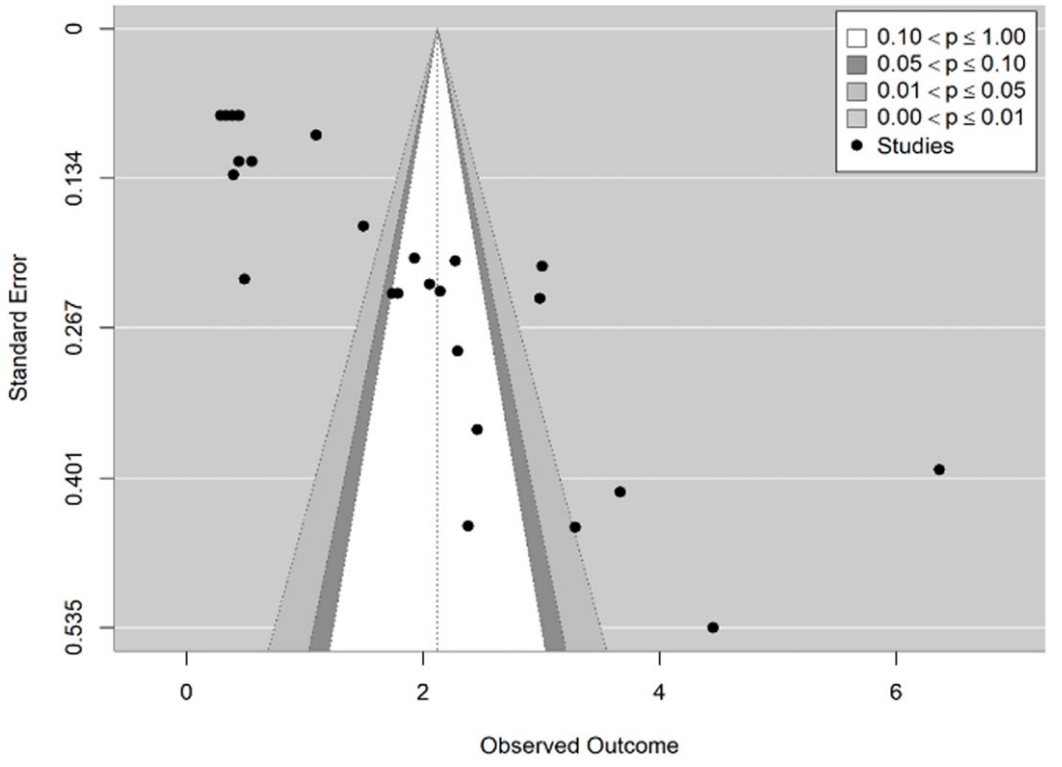


Figure 2. Funnel plot of Hedges's *g* against standard error for cake format C-DA studies.

participant characteristics and research contexts, this approach was deemed more appropriate for aggregating effect sizes.

Publication bias poses a potential threat to the validity of meta-analytic findings. To assess this risk, we examined contour-enhanced funnel plots and conducted Egger's regression tests for both cake and sandwich format analyses. The results indicated significant funnel plot asymmetry for both the cake format (intercept = 10.064, $SE = 1.188$, $p < .001$; see Figure 2) and the sandwich format (intercept = 9.781, $SE = 1.016$, $p < .001$; see Figure 3), suggesting possible publication bias. However, as Egger's test is not fully developed for multilevel meta-analytic models (Hou & Min, 2025), these results should be interpreted with caution. Robustness checks, including fail-safe *N* and trim-and-fill analyses, indicated that the overall conclusions were stable. Although effect sizes were attenuated after adjustment, all corrected estimates remained statistically significant, indicating that the positive effect of C-DA is robust, albeit with a potentially inflated magnitude (see supplementary material S4 for full details).

4. Results

4.1. Random-effect model results

The overall results of the random-effects model analysis are presented in Table 2. Large and statistically significant effect sizes were observed for both formats, with Hedges's $g = 2.120$ (95% CI [1.405, 2.836], $p < .001$) for the cake format and Hedges's $g = 1.676$ (95% CI [1.108, 2.244], $p < .001$) for the sandwich format. For cake format studies, this effect reflects the magnitude of during-mediation performance improvement, whereas for sandwich format studies, it reflects

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

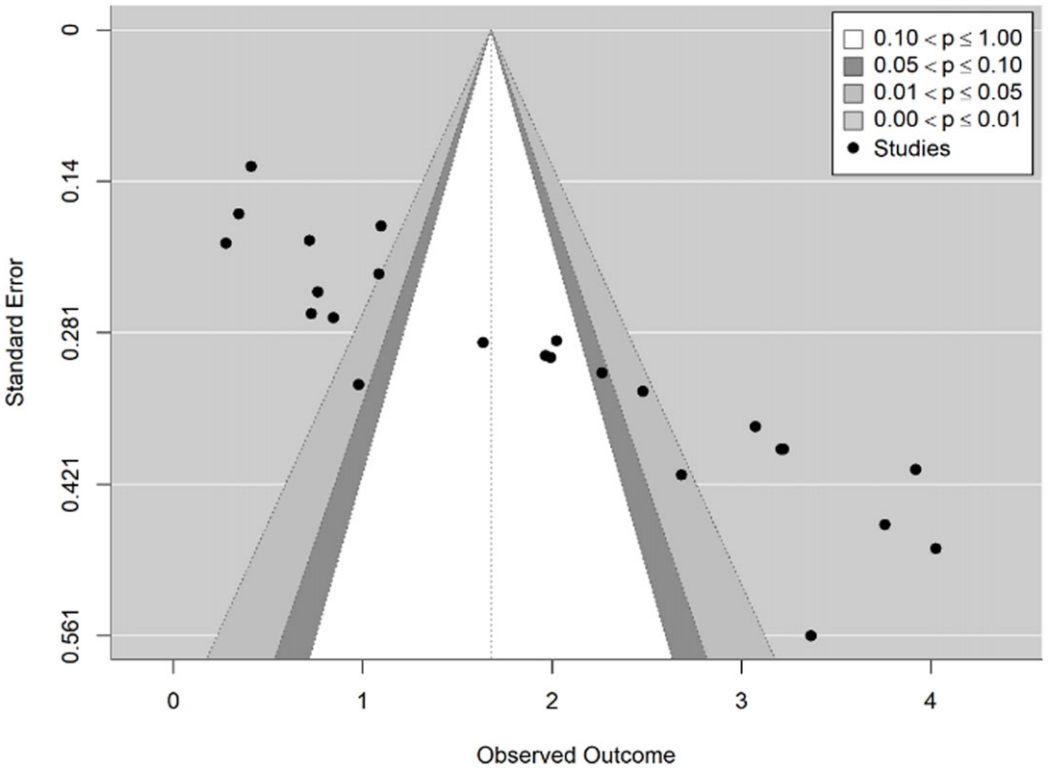


Figure 3. Funnel plot of Hedges's *g* against standard error for sandwich format C-DA studies.

changes in post-mediation performance. The cake format showed numerically larger effect sizes. The analysis also indicated substantial heterogeneity among studies in the cake group ($Q = 963.469$, $p < .001$) and the sandwich group ($Q = 368.629$, $p < .001$), suggesting the need for moderator analysis in both groups. In the cake group, the between-study variance was substantial, accounting for approximately 99.13% of the total variance in observed effect sizes, whereas sampling variance accounted for 0.87%, and no systematic within-study variance was detected. The likelihood ratio test indicated that the between-study systematic variance was statistically significant ($\chi^2 = 40.048$, $df = 1$, $p < .001$), suggesting that moderator variables may explain variability in effect sizes. In the sandwich group, the between-study variance was also substantial, accounting for approximately 89.66% of the total variance, with 5.83% attributed to systematic within-study variance and 4.51% to sampling variance. The likelihood ratio test showed that the between-study systematic variance was statistically significant ($\chi^2 = 8.052$, $df = 1$, $p = .005$), whereas the within-study variance was not ($\chi^2 = 1.219$, $df = 1$, $p = .270$), indicating that potential moderator effects are more likely to operate at the between-study level.

4.2. Moderating variables

Table 3 summarizes the results of categorical moderator analyses for both the cake and sandwich C-DA meta-analyses, whereas Tables 4 and 5 present the results of continuous moderator analyses for cake and sandwich C-DA formats, respectively.

In the cake group, meta-regression analyses indicated that the number of test items significantly predicted effect sizes ($\beta = 0.066$, $SE = 0.026$, $p = .010$), suggesting that studies with

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

Table 2. Results of C-DA's overall effectiveness under a random-effects model

C-DA format	Effect size	*SE*	95% CI	*p*	σ^2_v	σ^2_u	σ^2_e
Cake	2.120	0.365	[1.405, 2.836]	.000	99.138	0.000	0.862
Sandwich	1.676	0.290	[1.108, 2.244]	.000	89.659	5.830	4.511

Note. C-DA = computerized dynamic assessment; *SE* = standard error; CI = confidence interval; σ^2_v = between-study variance; σ^2_u = within-study variance; σ^2_e = typical sampling variance (calculated using Higgins and Thompson's, 2002, formula).

more test items tended to report larger effects. Specifically, each additional test item was associated with an average increase in effect size of 0.066, with the 95% CI [0.015, 0.116] excluding zero, further supporting the robustness of this relationship.

In contrast, in the sandwich group, learners' L1 and test content significantly moderated effect sizes. Learners' L1 exerted a significant moderating effect ($Q = 15.005$, $p < .001$), with studies involving Persian-speaking learners showing larger effects ($g = 2.470$, 95% CI [1.888, 3.053]) than those involving Chinese-speaking learners ($g = 0.674$, 95% CI [−0.024, 1.371]).

Test content was also a significant moderator in the sandwich group ($Q = 8.703$, $p = .034$). Effect sizes were largest for vocabulary ($g = 2.855$, 95% CI [1.604, 4.106]) and grammar outcomes ($g = 2.190$, 95% CI [1.319, 3.062]), followed by writing outcomes ($g = 1.466$, 95% CI [−0.010, 2.942]), whereas reading outcomes showed smaller effects ($g = 0.736$, 95% CI [−0.192, 1.664]).

Following the significant moderator effect of test content, post hoc pairwise contrasts were conducted using Tamhane's T2 procedure to correct for heterogeneity of variances and multiple testing. The results showed that grammar outcomes yielded significantly larger effect sizes than reading outcomes (mean difference = 1.78, $p = .023$, 95% CI [0.22, 3.34]). No other pairwise contrasts were statistically significant (all $p > .05$). Despite the large mean effect size for vocabulary outcomes, contrasts involving vocabulary were not significant, likely due to the small number of studies and wide confidence intervals in this subgroup.

Other moderators, including learners' academic status, DA approach, maximum number of prompts, prompt design method, number of C-DA sessions, assessment time constraints, target language, test format, random assignment, transfer items, transparency, institutional approval, and pedagogical or theoretical contribution, were not statistically significant (all $p > .05$). Given the limited number of studies in several subgroups, these findings should be interpreted with caution. Overall, these results suggest that substantive factors (learner characteristics and outcome domains) appear to play a more prominent role than methodological features in shaping the effectiveness of C-DA. Detailed characteristics and effect sizes for each included study are reported in supplementary material S5.

5. Discussion and implications

In this meta-analysis, we synthesized 27 effect sizes from cake format C-DA studies and 24 effect sizes from sandwich format C-DA studies to investigate the effectiveness of C-DA in enhancing L2 learners' performance. Beyond estimating overall effects, the meta-analysis aimed to identify potential sources of variability and to inform future research on the conceptualization and operationalization of C-DA. The findings yield several theoretical, methodological, and technological implications.

5.1. Theoretical implications

The cake format showed numerically larger effect sizes than the sandwich format, suggesting that during-mediation performance gains may be more readily captured than post-mediation performance outcomes. This finding is consistent with primary studies (e.g., Beigi et al., 2020).

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

Table 3. Moderator analysis for categorical variables

	Cake						Sandwich					
Moderator	*df*	*Q*	*p*	*k*	*n*	*g* [95%CI]	*df*	*Q*	*p*	*k*	*n*	*g* [95%CI]
Target language	1	0.618	.421									
English				13	20	2.282 [1.453, 3.112]						
Chinese				4	6	1.597 [0.102, 3.092]						
Learners' L1							1	15.005	.000			
Persian										9	15	2.470 [1.888, 3.053]
Chinese										6	6	0.674 [−0.024, 1.371]
Maximum number of prompts	2	1.887	.389				1	2.480	.115			
3				4	9	1.262 [−0.258, 2.782]				3	3	0.608 [−0.501, 1.717]
4				7	10	2.6 [1.436, 3.764]				9	12	1.637 [0.996, 2.279]
5				5	6	2.194 [0.819, 3.57]						
Prompts design method	3	2.947	.400									
Researcher design				6	10	2.721 [1.433, 4.009]						
Interactionist DA				4	8	1.403 [−0.161, 2.968]						
Strategy-based design				3	3	1.344 [−0.482, 3.17]						
Feedback-based and strategy-based design				2	4	3.012 [0.779, 5.245]						
DA format							1	0.042	.837			
Sandwich										13	19	1.715 [1.027, 2.404]
Cake and sandwich										5	5	1.578 [0.464, 2.692]
Assessment time constraints	1	1.127	.288				1	2.890	.089			
Yes				9	14	1.758 [0.78, 2.736]				7	8	1.088 [0.222, 1.954]
No				8	13	2.532 [1.489, 3.575]				11	16	2.051 [1.356, 2.747]
Test content	3	6.974	.073				3	8.703	.034			

(*Continued*)

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

Table 3. (*Continued*)

	Cake						Sandwich					
Moderator	*df*	*Q*	*p*	*k*	*n*	*g* [95%CI]	*df*	*Q*	*p*	*k*	*n*	*g* [95%CI]
Listening				7	13	1.528 [0.742, 2.313]						
Reading				7	7	1.689 [0.903, 2.475]				5	5	0.736 [−0.192, 1.664]
Pragmatics				3	4	3.251 [1.699, 4.804]						
Grammar				2	3	3.473 [1.569, 5.378]				6	9	2.190 [1.319, 3.062]
Vocabulary										3	3	2.855 [1.604, 4.106]
Writing										2	4	1.466 [−0.010, 2.942]
Test format	2	1.256	.534									
MDCT				11	19	2.105 [1.176, 3.035]						
Open-ended questions				3	3	1.342 [−0.444, 3.128]						
Mixed types				2	4	2.943 [0.756, 5.13]						
Transfer items	1	0.406	.524				1	2.432	.119			
Yes				7	11	2.406 [1.266, 3.547]				5	5	0.979 [−0.052, 2.010]
No				10	16	1.924 [0.974, 2.874]				13	19	1.947 [1.302, 2.591]
Confidence interval	1	1.436	.231				1	0.604	.437			
Yes				4	9	2.899 [1.441, 4.358]				8	11	1.933 [1.068, 2.797]
No				13	18	1.88 [1.07, 2.689]				10	13	1.473 [0.702, 2.245]
Random assignment							1	0.041	.839			
Yes										12	17	1.905 [1.160, 2.650]
No										3	4	1.733 [0.247, 3.219]
Institutional approval	1	0.201	.654				1	0.034	.855			
Yes				2	6	1.665 [−0.464, 3.794]				4	6	1.778 [0.549, 3.008]
No				15	21	2.184 [1.4, 2.967]				14	18	1.648 [0.982, 2.314]

Note. k = number of studies; n = number of effect sizes; g = Hedges's g; DA = dynamic assessment; MDCT = multiple-choice discourse completion test.

Table 4. Moderator analysis for continuous variables in the cake group

				95% CI		
	k	Coefficient	*SE*	Lower	Upper	*p*
Number of items	27	0.066	0.026	0.015	0.116	.010

Table 5. Moderator analysis for continuous variables in the sandwich group

				95% CI		
	k	Coefficient	*SE*	Lower	Upper	*p*
Number of items	24	0.026	0.014	0.002	0.054	.073
Number of C-DA sessions	24	−0.020	0.064	0.146	0.106	.760

Note. C-DA = computerized dynamic assessment.

One possible explanation is that cake format designs provide mediation immediately after learner errors, resulting in stronger performance improvement, whereas sandwich format designs typically provide mediation before the posttest, with no hints delivered during testing. The large positive effects of C-DA in improving L2 learners' performance is consistent with SCT, which conceptualizes development as mediated through interaction with more knowledgeable others (Aljaafreh & Lantolf, 1994). C-DA operationalizes this mediation by providing learners with graduated prompts as they work through tasks. C-DA therefore enhances learners' task performance through strategic support when learners encounter difficulties, enabling them to apply strategies or knowledge that they could not deploy independently.

Additionally, although C-DA effectiveness was generally robust across learner, mediation, and assessment contexts, several moderator effects were identified. Specifically, the number of items, test content, and learners' L1 emerged as significant moderators.

The number of items in the cake format C-DA significantly influenced its effectiveness in improving learners' performance. This is closely related to Vygotsky's (1978) concept of microgenesis, which focuses on the moment-to-moment changes in learners' performance over a very short period of time (Poehner, 2011). As development is an ongoing process (Poehner, 2005), an optimal number of items allows learners to progress through their ZPD with repeated opportunities for mediation. Too few items may limit opportunities for mediation, constraining observable performance improvement. However, given the limited sample size examined in this meta-analysis, these findings should be interpreted with caution. In particular, indefinitely increasing items is not recommended, as it may raise cognitive load and reduce mediation effectiveness. Balancing item quantity with learner proficiency is therefore crucial for maintaining learners within their ZPD and supporting performance improvement during C-DA interactions.

In terms of test content, the effectiveness of the sandwich format C-DA varied across domains. This variation might arise from the knowledge involved in different domains. Individuals may not have a single ZPD for general cognitive development but rather ZPDs specific to various content domains (Guthke, 1993). Thus, it is important to design mediational prompts oriented toward specific content to enhance the quality of the intervention. Additionally, C-DA research has focused disproportionately on receptive skills, such as listening and reading, while other dimensions of language ability, particularly pragmatic competence, remain underexplored (Lu et al., in press). To align with learners' developmental needs, future research should examine a broader range of targeted domains to enhance the diversity and representativeness of assessment content.

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

One notable finding of this study is the significant moderating effect of learners' L1 on the effectiveness of sandwich format C-DA. This may be partly attributable to the linguistic distance between the L1 and the target language, which can affect how mediation is internalized. Variations in linguistic structure, cultural norms, and communicative conventions may influence how salient and accessible mediation appears to learners from different language backgrounds (Lantolf & Poehner, 2004), thereby shaping their engagement with C-DA. Moreover, the effectiveness of C-DA may depend on learners' autonomy and willingness to participate in exploratory learning. From a sociocultural viewpoint, development is mediated not only through tools and signs but also through broader social and institutional contexts (Vygotsky, 1978). Educational cultures that prioritize teacher-centered instruction may limit learners' readiness to engage with the scaffolded mediations essential to C-DA. Thus, the internalization of mediation is not merely cognitive but also culturally embedded, highlighting the need to account for sociocultural and pedagogical backgrounds in C-DA implementation.

5.2. Methodological implications

First, this study underscores the methodological importance of distinguishing between cake- and sandwich-format C-DA in meta-analysis. Although both operationalizations address the broader question of how mediation facilitates L2 learners' performance, they differ in the timing at which development is captured. Future studies may integrate both formats by embedding the cake format mediation between pre- and posttests, thereby providing a more comprehensive depiction of learners' performance trajectories over time.

Second, 65.7% of the included studies did not incorporate transfer items. Transfer items constitute an important component of C-DA, as they allow researchers to examine whether learners can apply previously mediated principles to novel or more complex contexts (Qin & van Compernolle, 2021). However, the absence of transfer items does not necessarily invalidate a study, as C-DA can still provide valuable insights into learners' mediated performance within a given context. Rather, the inclusion of transfer items can enhance the explanatory power of C-DA outcomes. We therefore recommend that future C-DA studies systematically incorporate transfer items where appropriate.

Third, although many studies reported random assignment, internal validity remains a concern. In classroom-based studies, individual-level randomization is often impractical, leading researchers to rely on intact classes. While this practice helps preserve ecological validity, it may weaken causal inference because preexisting class differences cannot be fully controlled. Future research should therefore prioritize group-level randomization in classroom-based settings where individual-level randomization is not feasible, balancing methodological rigor with classroom authenticity.

This study also identified notable weaknesses in reporting practices. Only 34.3% of the included studies reported confidence intervals for inferential statistics, limiting the interpretability of the findings. Transparent reporting of uncertainty estimates is essential for evaluating the robustness of primary study results and for enabling accurate synthesis in future meta-analyses (Plonsky, 2024). We therefore encourage researchers to adhere more closely to established reporting guidelines to enhance the interpretability, reproducibility, and cumulative value of C-DA research.

5.3. Technological implications

Beyond theoretical and methodological insights, the findings also have important technological implications for C-DA and computer-assisted language learning. Technology enables standardized delivery of graduated mediation, thereby operationalizing needs-contingent support to a large number of learners in ways that are difficult to achieve in human-mediated DA and enabling quantifiable assessment outcomes. Besides, computerized platforms can log interaction

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

traces and offer fine-grained evidence of developmental processes. These affordances position C-DA as more than a digital version of DA; rather, it constitutes a distinct assessment ecology in which technological mediation reshapes how learner performance and learning potential are elicited and interpreted. From a pedagogical perspective, the large effects observed in this meta-analysis suggest that technology-enhanced mediation holds considerable promise for supporting learner performance.

C-DA typically positions the computer as the mediator, enabling standardized and scalable interventions for large cohorts of learners. However, such mediation is often insufficiently tailored to individual learners, which constitutes a key limitation of current C-DA implementations. Looking forward, recent advances in large language models (LLMs) offer promising opportunities for the next generation of C-DA systems. Leveraging LLMs' contextual understanding and generative abilities, future C-DA platforms could integrate interventionist and interactionist approaches by providing individualized, dialogic mediation while maintaining scalability. Fine-tuned LLM-based systems may deliver context-sensitive prompts and dynamically adapt mediation to learners' needs, thereby enhancing C-DA efficiency.

6. Limitations and future research

Although the present meta-analysis provides important insights into the effectiveness of C-DA, several limitations should be acknowledged, each of which points to directions for future research.

First, Egger's regression test indicated potential publication bias, suggesting that the overall effect size estimates may be inflated due to the underrepresentation of null or nonsignificant findings. It is important to note, however, that formal statistical procedures for assessing publication bias in multilevel meta-analytic models are currently limited (Hou & Min, 2025). Accordingly, publication bias was evaluated using conventional two-level approaches, and the findings should be interpreted with caution. Nevertheless, robustness analyses using fail-safe *N* and trim-and-fill procedures indicated that the overall conclusions remained stable. Future methodological work is needed to develop bias-detection techniques tailored to multilevel models.

Second, the scope of the included literature may limit the generalizability of the findings. The focus on English-language publications, peer-reviewed journal articles, and dissertations may have excluded relevant studies reported in other languages or formats, such as conference proceedings and book chapters. Future meta-analyses should consider broader sources to provide a more comprehensive picture of the field.

Finally, the relatively small number of eligible studies reflects the emerging nature of C-DA research and constrains the precision of moderator analyses. This highlights the need for a larger body of methodologically rigorous primary studies with more transparent reporting of design features and mediation procedures. Such efforts would facilitate more fine-grained meta-analytic investigations in this field.

7. Conclusion

This meta-analysis investigated 51 effect sizes from 35 studies published between 2000 and May 27, 2025, to examine the effectiveness of C-DA in enhancing L2 learners' performance and to identify key moderators of these effects. Overall, the findings provide robust quantitative evidence that C-DA substantially enhances learners' performance, while also revealing systematic variation associated with mediation, assessment context, and learner background. Notably, both cake and sandwich formats showed strong positive effects, although the cake format tended to yield a larger effect size, likely reflecting differences in mediation timing. In addition, factors such as item number, test content, and learners' L1 were found to significantly shape C-DA outcomes, highlighting the importance of theoretically informed design in optimizing mediated learning.

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

Taken together, these findings highlight the potential of C-DA as a powerful technology-mediated assessment framework for promoting learner development. At the same time, they point to the need for continued methodological refinement to better understand how mediation operates across different contexts and learner populations.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/S0958344026100561

Data availability statement. All data and supplementary materials are publicly available via the Open Science Framework (OSF): https://osf.io/62b8s/.

Authorship contribution statement. Qi Lu: Conceptualization-Equal, Formal analysis-Equal, Investigation-Equal, Methodology-Equal, Validation-Equal, Writing - original draft-Equal, Writing - review & editing-Equal; Mengqi Chen: Data curation-Equal, Formal analysis-Equal, Investigation-Equal, Methodology-Equal, Validation-Equal, Visualization-Equal; Lianrui Yang: Funding acquisition-Equal, Project administration-Equal, Resources-Equal, Supervision-Equal; Shaofeng Li: Conceptualization-Supporting, Writing - review & editing-Equal; Ying Chen: Conceptualization-Equal, Funding acquisition-Equal, Project administration-Equal, Resources-Equal, Supervision-Equal, Writing - review & editing-Equal.

Funding disclosure statement. This research was supported by the National Planning Office for Philosophy and Social Sciences of the People's Republic of China (17AYY023 and 20BYY108).

Competing interests statement. The authors declare no competing interests.

Ethical statement. Ethical approval was not required.

GenAI use disclosure statement. ChatGPT (version 4, https://chat.openai.com) was used to revise some sentences for grammatical accuracy. The authors reviewed and approved all suggestions and take full responsibility for the final text.

References

Alavi, S. M., Shahsavar, M., & Norouzi, M. H. (2020). Diagnosing EFL learners' development of pragmatic competence implementing computerized dynamic assessment. *Issues in Language Teaching*, 9(1), 117–149. https://doi.org/10.22054/ilt.2020.42653.400

Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal*, 78(4), 465–483. https://doi.org/10.1111/j.1540-4781.1994.tb02064.x

Bakhoda, I., & Shabani, K. (2019). Bringing L2 learners' learning preferences in the mediating process through computerized dynamic assessment. *Computer Assisted Language Learning*, 32(3), 210–236. https://doi.org/10.1080/09588221.2018.1483950

Barabadi, E., Khajavy, G. H., & Kamrood, A. M. (2018). Applying interventionist and interactionist approaches to dynamic assessment for L2 listening comprehension. *International Journal of Instruction*, 11(3), 681–700. https://doi.org/10.12973/iji.2018.11346a

Behbahani, H. K., & Karimpour, S. (2025). Exploring the impact of computerized dynamic assessment on the explicit and implicit knowledge of reflexive pronouns: The mediating role of brain dominance. *System*, 128, Article 103537. https://doi.org/10.1016/j.system.2024.103537

Beigi, A. M., Basirloo, S. R., Molaei, A., & Yazdani, S. (2020). Effect of interventionist computerized dynamic assessment on learners' grammatical English tenses: Analysis of cake and sandwich formats. *International Journal of Foreign Language Teaching and Research*, 8(32), 79–92.

Belz, J. A., & Kinginger, C. (2003). Discourse options and the development of pragmatic competence by classroom learners of German: The case of address forms. *Language Learning*, 53(4), 591–647. https://doi.org/10.1046/j.1467-9922.2003.00238.x

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley-Blackwell. https://doi.org/10.1002/9780470743386

Budoff, M. (1987). The validity of learning potential assessment. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 53–81). Guilford Press.

Campione, J. C., Brown, A. L., Ferrara, R. A., & Bryant, N. R. (1984). The zone of proximal development: Implications for individual differences and learning. *New Directions for Child Development*, 23, 77–91. https://doi.org/10.1002/cd.23219842308

Ebadi, S., & Bashir, S. (2020). An exploration into EFL learners' writing skills via mobile-based dynamic assessment. *Education and Information Technologies*, 26(2), 1995–2016. https://doi.org/10.1007/s10639-020-10348-4

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

Ebadi, S., & Goodarzi, A. (2023). Exploring perceptual learning-style preferences of English language non-gainers in computerized dynamic reading assessment. *Reading Psychology*, 44(1), 95–115. https://doi.org/10.1080/02702711.2022.2141393

Geng, X., Meng, Y., & Du, W. (2022). Research on computerized L2 dynamic assessment: Development and framework. *Modern Foreign Languages*, 45(1), 126–136.

Ghahderijani, B. H., Namaziandost, E., Tavakoli, M., Kumar, T., & Magizov, R. (2021). The comparative effect of group dynamic assessment (GDA) and computerized dynamic assessment (C-DA) on Iranian upper-intermediate EFL learners' speaking complexity, accuracy, and fluency (CAF). *Language Testing in Asia*, *11*, Article 25. https://doi.org/10.1186/s40468-021-00144-3

Guthke, J. (1993). Current trends in theories and assessment of intelligence. In J. H. M. Hamers, K. Sijtsma, & A. J. J. M. Ruijssenaars (Eds.), *Learning potential assessment: Theoretical, methodological and practical issues* (pp. 13–20). Swets & Zeitlinger.

Guthke, J., & Beckmann, J. F. (2000). The learning test concept and its applications in practice. In C. S. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 17–69). JAI.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hou, Z., & Min, S. (2025). Dialogue-based computer-assisted language learning systems for second language speaking development: A three-level meta-analysis. *ReCALL*. Advance online publication. https://doi.org/10.1017/S0958344025100268

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). SAGE Publications. https://doi.org/10.4135/9781412985031

Izadi, M., Izadi, M., & Heidari, F. (2023). The potential of an adaptive computerized dynamic assessment tutor in diagnosing and assessing learners' listening comprehension. *Education and Information Technologies*, 29(4), 3637–3661. https://doi.org/10.1007/s10639-023-11871-w

Jin, C., & Liu, Y. (2024). Diagnosing and promoting learners' L2 inferential reading development through hybrid computerised dynamic assessment in the Chinese EFL classroom. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2024.2421521

Kamrood, A. M., Davoudi, M., Ghaniabadi, S., & Amirian, S. M. R. (2021). Diagnosing L2 learners' development through online computerized dynamic assessment. *Computer Assisted Language Learning*, 34(7), 868–897. https://doi.org/10.1080/09588221.2019.1645181

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics*, 1(1), 49–72. https://doi.org/10.1558/japl.1.1.49.55872

Lu, Q., Chen, Y., & Yang, L. (in press). Developing and validating a computerized dynamic diagnostic assessment of pragmatic competence for Chinese learners of English. *Language Testing*.

McNeil, L. (2018). Understanding and addressing the challenges of learning computer-mediated dynamic assessment: A teacher education study. *Language Teaching Research*, 22(3), 289–309. https://doi.org/10.1177/1362168816668675

Meng, Y., Fu, H., & Wang, C. (2024). The effectiveness of computerized listening dynamic assessment: Attribute-based mediation model. *Language Learning & Technology*, 28(1), 1–28. https://doi.org/10.64152/10125/73580

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7(1), 105–125. https://doi.org/10.1037/1082-989X.7.1.105

Nicholas, A., Blake, J., Perkins, J., & Mozgovoy, M. (2024). Evaluating the effectiveness of a computerised dynamic assessment of L2 English email requests. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2024.2374775

Ortega, L. (2005). For what and for whom is our research? The ethical as transformative lens in instructed SLA. *The Modern Language Journal*, 89(3), 427–443. https://doi.org/10.1111/j.1540-4781.2005.00315.x

Pileh Roud, L. F., & Hidri, S. (2021). Toward a sociocultural approach to computerized dynamic assessment of the TOEFL iBT listening comprehension test. *Education and Information Technologies*, 26(4), 4943–4968. https://doi.org/10.1007/s10639-021-10498-z

Plonsky, L. (2024). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*, 44, 4–18. https://doi.org/10.1017/S0267190524000059

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. https://doi.org/10.1111/lang.12079

Poehner, M. E. (2005). *Dynamic assessment of oral proficiency among advanced L2 learners of French* [Doctoral dissertation, The Pennsylvania State University]. Electronic Theses and Dissertations for Graduate School. https://etda.libraries.psu.edu/catalog/6627

Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *The Modern Language Journal*, 91(3), 323–340. https://doi.org/10.1111/j.1540-4781.2007.00583.x

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press

Poehner, M. E. (2011). Dynamic assessment: Fairness through the prism of mediation. *Assessment in Education: Principles, Policy & Practice*, 18(2), 99–112. https://doi.org/10.1080/0969594X.2011.567090
Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). *Language Teaching Research*, 17(3), 323–342. https://doi.org/10.1177/1362168813482935
Poehner, M. E., Zhang, J., & Lu, X. (2015). Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing*, 32(3), 337–357. https://doi.org/10.1177/0265532214560390
Qin, T., & van Compernolle, R. A. (2021). Computerized dynamic assessment of implicature comprehension in L2 Chinese. *Language Learning & Technology*, 25(2), 55–74. https://doi.org/10.64152/10125/73433
Shabani, K. (2022). Diagnostic and developmental potentials of computerized dynamic assessment (C-DA) for L2 vocabulary. *Interdisciplinary Studies in English Language Teaching*, 1(2), 130–149. https://doi.org/10.22080/ISELT.2022.23570.1035
Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge University Press.
Teo, A. (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning & Technology*, 16(3), 10–20. https://doi.org/10.64152/10125/44292
Teo, A. (2014). Beyond traditional testing: Exploring the use of computerized dynamic assessment to improve EFL learners' reading. *Arab World English Journal*, 5(1), 42–58.
Vakili, S., & Ebadi, S. (2022). Exploring EFL learners' developmental errors in academic writing through face-to-face and computer-mediated dynamic assessment. *Computer Assisted Language Learning*, 35(3), 345–380. https://doi.org/10.1080/09588221.2019.1698616
Valsiner, J. (2012). *A guided science: History of psychology in the mirror of its making*. Transaction Publishers.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. https://doi.org/10.18637/jss.v036.i03
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
Wang, T.-H. (2010). Web-based dynamic assessment: Taking assessment as teaching and learning strategy for improving students' e-Learning effectiveness. *Computers & Education*, 54(4), 1157–1166. https://doi.org/10.1016/j.compedu.2009.11.001
Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Harvard University Press.
Youn, S. J., & Bogorevich, V. (2020). Assessment in L2 pragmatics. In N. Taguchi (Ed.), *The Routledge handbook of second language acquisition and pragmatics* (pp. 308–321). Routledge. https://doi.org/10.4324/9781351164085-20
Zhang, J., & Lu, X. (2019). Measuring and supporting second language development using computerized dynamic assessment. *Language and Sociocultural Theory*, 6(1), 92–115. https://doi.org/10.1558/lst.31710
Zhou, Y., & Zhou, M. (2025). A meta-analysis on mobile-assisted vocabulary learning: Do mobile applications help? *ReCALL*. Advance online publication. https://doi.org/10.1017/S0958344025100335

About the authors

Qi Lu is a postgraduate student at the College of Foreign Languages at Ocean University of China. Her research interests include L2 pragmatics, dynamic assessment, and local grammar.

Mengqi Chen is a postgraduate student at the College of Foreign Languages at Ocean University of China. Her research interests include second language acquisition, language assessment, diagnostic assessment, and computerized dynamic assessment.

Lianrui Yang is a professor at the College of Foreign Languages at Ocean University of China. His current research interests are second language acquisition, interlanguage linguistics, AI-supported language learning and teaching, and quantitative linguistics.

Shaofeng Li is a professor at the Department of English and Communication, the Hong Kong Polytechnic University, China. His research interests include language aptitude, working memory, form-focused instruction, task-based language teaching and learning, corrective feedback, and research methods (including meta-analysis).

Ying Chen is a professor of applied linguistics at Ocean University of China. Her research interests include language assessment, second language acquisition, and L2 pragmatics.

https://doi.org/10.1017/S0958344026100561 Published online by Cambridge University Press