VOLUME 16, NO. 2, DECEMBER 2026

ALLEMANIA

Journal homepage:
https://ejournal.upi.edu/index.php/allemania/index

Effectiveness of FelloFisch as a Domain-Specific CALL Platform in Preparing Indonesian Learners for the Goethe-Zertifikat B1 Writing Examination

Dani Hendra[1] Putrasulung Baginda[1] Ending Khoerudin[1]
[1] German Language Education Study Program, Universitas Pendidikan Indonesia, Indonesia

Article Info

Paper Type:
Research Paper

Received: August 13, 2026
Revised 1: October 17, 2026
Revised 2: November 7, 2026
Accepted: November 23, 2026
First Available online: December 01, 2026
Publication Date: December 01, 2026

Abstract

This study examined the effectiveness of FelloFisch, a domain-specific computer-assisted language learning (CALL) platform aligned with the Goethe-Institut assessment criteria, in preparing Indonesian learners for the Goethe-Zertifikat B1 writing examination. A quasi-experimental pretest–posttest control group design involved 50 German language education students at Universitas Pendidikan Indonesia in two intact classes: an experimental group using FelloFisch (n = 25) and a control group receiving conventional instruction (n = 25). Writing was assessed with a test modelled on the B1 Schreiben module and analysed using parametric and non-parametric tests, N-Gain, and Cohen's d.

Both groups were equivalent at pre-test ($p = 0.529$). The experimental group improved significantly (72.24 to 83.62; $p < 0.001$), whereas the control group improved only marginally (69.84 to 75.06; $p = 0.049$). The experimental group outperformed the control group on the post-test ($p = 0.002$; $d = 0.93$) and achieved a higher mean N-Gain (0.403 vs. −0.110; $p = 0.015$). No participant in the experimental group regressed, whereas 24% of the control group did. The findings indicate that immediate, criterion-aligned feedback makes domain-specific CALL effective for B1 writing preparation, although they are limited to a single institution.

Keywords: CALL; FelloFisch; Goethe-Zertifikat B1; German as a Foreign Language; L2 Writing; Corrective Feedback

1. INTRODUCTION

Internationally certified German proficiency has become a strategic prerequisite for academic and professional mobility, and the Goethe-Zertifikat (GZ) B1 is one of the most widely recognised certification routes (Goethe-Institut, 2023). Within the Common European Framework of Reference for Languages, the B1 level marks the threshold at which learners are expected to

[1]*Corresponding author: dani.hendra@upi.edu

produce simple connected texts on familiar topics and to describe experiences, events, and opinions (Council of Europe, 2020). Among the four components of the GZ B1 examination, however, the writing component (*Schreiben*) consistently yields the lowest average scores, particularly among learners whose first language is not European.

For Indonesian learners, the difficulty of B1 writing is compounded by sharp typological differences between Indonesian and German. Indonesian has no case system and no grammatical gender, and its word order is considerably more flexible than that of German. These differences give rise to systematic negative transfer, especially in the use of the accusative and dative cases, verb-second word order, and subordinate clause structure. At the same time, conventional writing instruction that relies on manual correction is constrained by delayed feedback, subjectivity between raters, and limited opportunities to individualise intervention (Ferris, 2011; Hyland, 2019).

Computer-assisted language learning (CALL) offers a means of addressing these constraints by providing immediate and individualised feedback (Warschauer & Healey, 1998; Golonka et al., 2014; Hubbard, 2022). FelloFisch is a CALL platform that implements the four official Goethe-Institut assessment criteria for writing, namely *Aufgabenerfüllung* (task fulfilment), *Kohärenz* (coherence), *Wortschatz* (vocabulary), and *Strukturen* (structures). Despite its explicit alignment with the examination standard, no empirical study has yet evaluated the effectiveness of FelloFisch among Indonesian learners.

This study therefore aims (1) to examine differences in B1-level German writing ability between students who used FelloFisch and students who received conventional instruction; (2) to measure the efficiency of improvement in writing ability in both groups through normalised gain (N-Gain) analysis; and (3) to determine the effect size of the FelloFisch intervention on writing examination scores assessed with the Goethe-Institut rubric.

2. LITERATURE REVIEW

Research on CALL has long shown that digital technologies can support foreign language learning by increasing learners' exposure to target-language input, opportunities for practice, and access to feedback (Warschauer & Healey, 1998; Hubbard, 2022). In their review of technology types and their effectiveness, Golonka et al. (2014) reported that technology-based feedback generally produces moderate effects, with effect sizes of approximately $d = 0.50–0.65$. Most of the studies reviewed, however, used generic platforms rather than tools designed for a specific assessment standard. This raises the question of whether domain-specificity, that is, the alignment of a platform with the criteria of a particular examination, moderates the effectiveness of CALL.

Corrective feedback is a central mechanism through which CALL may improve second language writing. Ferris (2011) argues that the treatment of error in L2 writing is most effective when feedback is timely, consistent, and focused on learners' recurring error patterns. A meta-analysis by Kang and Han (2015) showed that written corrective feedback can improve the grammatical accuracy of L2 writing, and Rankin et al. (2022) reported efficiency and accuracy gains from digital feedback in German L2 writing. In contrast, Schmidt (2024) observed that generic artificial intelligence tools are less effective in sustaining the motivation of advanced German learners because their feedback is not sufficiently specific or challenging.

From a sociocultural perspective, learning occurs within the zone of proximal development, the distance between what learners can achieve independently and what they can achieve with assistance (Vygotsky, 1978). Scaffolding refers to the support that enables learners to complete tasks beyond their current independent level (Wood et al., 1976). A CALL platform that provides structured, criterion-based feedback can function as such a mediator, guiding learners from their actual level towards their potential developmental level.

The GZ B1 writing module assesses learners' ability to produce communicative texts such as personal e-mails, forum contributions, and semi-formal messages, in line with the communicative competences described for the B1 level in *Profile Deutsch* (Glaboniat et al., 2015) and the CEFR (Council of Europe, 2020). Performance is rated against the four Goethe-Institut criteria mentioned above (Goethe-Institut, 2023). Because FelloFisch applies the same criteria, practice on the platform is expected to transfer directly to examination performance.

Drawing on these perspectives, the present study postulates three causal pathways through which FelloFisch may improve writing performance: (1) a cognitive pathway, in which instant feedback supports the internalisation of grammatical rules; (2) a scaffolding pathway, in which the platform mediates learning within the zone of proximal development; and (3) an ecological validity pathway, in which criterion-aligned practice minimises the transfer gap between practice and assessment. To date, these pathways have not been tested empirically for a domain-specific CALL platform among Indonesian learners of German preparing for GZ B1, which constitutes the research gap addressed by this study.

3. METHODOLOGY

3.1. Research Design

This study employed a quantitative approach with a quasi-experimental pretest–posttest control group design. Because the participants were organised in existing parallel classes, intact groups were used instead of random assignment. One class served as the experimental group, which practised writing with FelloFisch, and the other served as the control group, which received conventional instruction with manual correction by the lecturer.

3.2. Participants

The participants were 50 students of the German Language Education Study Program at Universitas Pendidikan Indonesia, distributed equally across two parallel classes: Class A as the experimental group (n = 25) and Class B as the control group (n = 25). No attrition occurred during the study; all participants completed the pre-test, attended all intervention sessions, and completed the post-test, resulting in a complete dataset without missing values. The gender composition of the two classes was comparable and was not identified as a significant confounding variable.

3.3. Instruments

The main instrument was a writing test modelled on the official format of the GZ B1 *Schreiben* module and scored with the Goethe-Institut rubric, with a maximum score of 100 points. The rubric covers the four official criteria: *Aufgabenerfüllung*, *Kohärenz*, *Wortschatz*, and

Strukturen. The instrument was reviewed by an external expert from Johannes Gutenberg-Universität Mainz, Germany, with particular attention to the alignment of the writing tasks and scoring rubric with the Goethe-Institut criteria at B1 level. Scoring was calibrated through an inter-rater procedure with a target reliability of Cohen's $\kappa \geq 0.70$. In addition, a User Experience questionnaire (20 items on a five-point Likert scale) was administered to the experimental group as secondary data; its results are not reported in this article.

3.4. Procedure

Both groups first completed the pre-test. During the treatment period, the experimental group practised B1 writing tasks on FelloFisch, which provides immediate feedback annotated at specific locations in the text according to the four Goethe-Institut criteria. The control group received conventional instruction in which their writing was corrected manually by the lecturer. Because access to FelloFisch depends on a stable internet connection, scheduled practice sessions were provided in the campus language laboratory to ensure that all students in the experimental group could practise. Both groups then completed the post-test.

3.5. Data Analysis

Normality was tested with the Shapiro–Wilk test and homogeneity of variance with Levene's test. Depending on the results, within-group differences between pre-test and post-test were examined with the paired t-test or the Wilcoxon signed-rank test, and between-group differences with the independent t-test or the Mann–Whitney U test. The efficiency of improvement was calculated as normalised gain, g = (post-test − pre-test) / (maximum score − pre-test), and categorised as high ($g \geq 0.70$), medium ($0.30 \leq g < 0.70$), or low ($g < 0.30$) (Hake, 1998); negative values indicate score regression. The magnitude of the treatment effect was calculated with Cohen's d and interpreted as small (0.20), medium (0.50), or large (0.80) (Cohen, 1988). All tests used a significance level of $\alpha = 0.05$.

4. RESULTS AND DISCUSSION

The data comprised the pre-test and post-test scores of both groups and the individual N-Gain values. The results are presented in six parts: descriptive statistics, assumption tests, within-group differences, between-group differences and effect size, the distribution of N-Gain and cases of score regression, and a discussion of the findings.

4.1. Descriptive Statistics

Table 1 summarises the descriptive statistics for both groups. The absolute improvement in Class A (11.38 points; 15.75%) was more than twice that in Class B (5.22 points; 7.47%), indicating that FelloFisch had a substantial differential effect on writing performance.

Table 1. Descriptive Statistics by Group

Group / Variable	n	Mean	SD	Min	Max	Gain	% Gain
Class A – Pre-test	25	72.24	12.664	45.0	92.5	11.38	15.75%

Group / Variable	n	Mean	SD	Min	Max	Gain	% Gain
Class A – Post-test	25	83.62	7.776	72.0	98.5	—	—
N-Gain Class A	25	0.403	0.165	0.000	0.880	Medium	—
Class B – Pre-test	25	69.84	14.041	52.0	92.0	5.22	7.47%
Class B – Post-test	25	75.06	10.379	52.0	93.5	—	—
N-Gain Class B	25	−0.110	1.049	−4.000	0.640	Highly variable	—

The post-test standard deviation of Class A (7.776) was considerably smaller than that of Class B (10.379). This indicates that FelloFisch not only raised the mean score but also reduced the variability of achievement, so that improvement occurred consistently across the experimental group. The mean N-Gain of Class A (0.403, medium category) contrasts sharply with that of Class B (−0.110), which was negative because several members of the control group regressed from pre-test to post-test.

4.2. Assumption Tests

Before the statistical tests were selected, the normality of each variable was examined with the Shapiro–Wilk test (Table 2) and homogeneity of variance with Levene's test (Table 3).

Table 2. Results of the Shapiro–Wilk Normality Test

Variable / Group	n	W	p-value	Decision	Distribution	Implication
Class A – Pre-test	25	0.9649	0.5206	H_0 accepted	Normal	Parametric
Class A – Post-test	25	0.9279	0.0777	H_0 accepted	Normal	Parametric
Class B – Pre-test	25	0.8869	0.0096*	H_0 rejected	Not normal	Non-parametric
Class B – Post-test	25	0.9620	0.4555	H_0 accepted	Normal	Parametric
N-Gain Class A	25	0.9339	0.1067	H_0 accepted	Normal	Parametric
N-Gain Class B	25	0.6343	< 0.001*	H_0 rejected	Not normal	Non-parametric

Note: * $p < 0.05$.

Two of the six variables were not normally distributed: the Class B pre-test ($W = 0.887$, $p = 0.010$) and the Class B N-Gain ($W = 0.634$, $p < 0.001$). The non-normality of the Class B pre-test suggests a skewed or bimodal distribution of initial scores, which is typical of heterogeneous classes that include both highly proficient and weak students. The extreme non-normality of the Class B N-Gain reflects a highly asymmetrical distribution caused by strongly negative N-Gain values, as low as −4.0, for several individuals.

Table 3. Results of Levene's Test of Homogeneity of Variance

Variable	Groups Compared	Levene's F	df	p-value	Decision
Pre-test	Class A vs. Class B	0.6802	1, 48	0.4136	Homogeneous
Post-test	Class A vs. Class B	0.6509	1, 48	0.4238	Homogeneous
N-Gain	Class A vs. Class B	5.0196	1, 48	0.0297*	Not homogeneous

Note: * $p < 0.05$.

The variances of the pre-test and post-test scores were homogeneous ($F = 0.680$, $p = 0.414$; $F = 0.651$, $p = 0.424$), which justifies the use of the independent t-test assuming equal variances. In contrast, the variances of N-Gain were not homogeneous ($F = 5.020$, $p = 0.030$), which confirms the choice of the Mann–Whitney U test for the between-group comparison of N-Gain. This

heterogeneity reflects fundamentally different patterns of improvement: a narrow, consistent distribution in Class A and a wide, extreme distribution in Class B.

4.3. Within-Group Differences (Pre-test vs. Post-test)

The first stage of hypothesis testing examined whether each group improved significantly from pre-test to post-test (Table 4).

Table 4. Within-Group Differences between Pre-test and Post-test

Group	Test	M Pre	M Post	Statistic	df	p-value	Decision
Class A (Experimental)	Paired t-test	72.24	83.62	t = −8.812	24	< 0.001	Significant
Class B (Control)	Wilcoxon signed-rank	69.84	75.06	W = 81.0	—	0.049	Significant

In Class A, the paired t-test showed a highly significant improvement ($t(24) = -8.812$, $p < 0.001$). A t-value of this magnitude indicates that the improvement of 11.38 points is not an artefact of sampling variability but reflects a strong treatment effect. The 95% confidence interval of the mean difference ranged from 8.68 to 14.08 and lay entirely above zero, confirming the direction of improvement.

In Class B, the Wilcoxon signed-rank test detected a statistically significant improvement ($W = 81$, $p = 0.049$), but only by a very narrow margin below $\alpha = 0.05$. A mean improvement of 5.22 points with a p-value close to the threshold indicates that conventional instruction did produce some improvement, but with a much weaker and less consistent effect than in Class A.

4.4. Between-Group Differences and Effect Size

The second stage examined whether the differences observed between Class A and Class B in the post-test and N-Gain could be attributed to the treatment rather than to differences in initial conditions (Table 5).

Table 5. Between-Group Differences

Variable	Test	M Class A	M Class B	Statistic	df	p-value	Decision
Pre-test	Independent t-test	72.24	69.84	t = 0.635	48	0.529	Not significant
Post-test	Independent t-test	83.62	75.06	t = 3.300	48	0.002	Significant
N-Gain	Mann–Whitney U	0.403	−0.110	U = 438.5	—	0.015	Significant

There was no significant difference between the pre-test scores of Class A ($M = 72.24$) and Class B ($M = 69.84$) ($t(48) = 0.635$, $p = 0.529$). This confirms that both groups started from equivalent initial conditions, a crucial prerequisite for inferring that post-test differences result from the treatment rather than from differences in initial ability.

In contrast, there was a highly significant difference between the post-test scores of Class A ($M = 83.62$) and Class B ($M = 75.06$) ($t(48) = 3.300$, $p = 0.002$). The mean difference of 8.56 points strongly supports the hypothesis that the use of FelloFisch contributes to greater

improvement in examination scores than conventional instruction. With regard to the efficiency of improvement, the Mann–Whitney U test detected a significant difference in the distribution of N-Gain between the groups ($U = 438.5$, $p = 0.015$). A U-value this high within the possible range of 0–625 for $n_1 = n_2 = 25$ indicates the dominance of Class A's N-Gain over that of Class B. FelloFisch thus produced not only larger absolute score changes but also more efficient improvement relative to each learner's maximum potential.

Table 6. Effect Size (Cohen's d)

Variable	Cohen's d	Category	Interpretation
Post-test	0.9335	Large	A difference of 0.93 SD between the experimental and control groups
N-Gain	0.6840	Medium	A difference of 0.68 SD in the efficiency of improvement

The Cohen's d of 0.93 for the post-test is a notable finding because the effect is not only statistically significant but also of considerable practical relevance. For comparison, Hattie's (2009) synthesis of more than 800 meta-analyses of educational interventions reported an average effect size of $d = 0.40$. The effect found in this study is more than twice that average, so the FelloFisch intervention can be regarded as a highly effective educational intervention.

4.5. Distribution of N-Gain and Cases of Score Regression

Table 7. Distribution of N-Gain by Category

N-Gain Category	Class A (n)	Class A (%)	Class B (n)	Class B (%)	Difference (A − B)
High ($g \geq 0.70$)	1	4.0%	0	0.0%	+4.0%
Medium ($0.30 \leq g < 0.70$)	22	88.0%	8	32.0%	+56.0%
Low ($0 \leq g < 0.30$)	2	8.0%	11	44.0%	−36.0%
Negative ($g < 0$)	0	0.0%	6	24.0%	−24.0%
Total	25	100%	25	100%	—

The distribution of N-Gain reveals a strongly contrasting pattern between the groups (Table 7). In Class A, 92% of the students (23 of 25) were in the medium or high category, and none regressed (0% negative N-Gain). This consistency strongly indicates that the effectiveness of FelloFisch applies across the group and does not benefit only learners with particular characteristics. In Class B, 24% of the students (6 of 25) regressed with a negative N-Gain and 44% were in the low category. Only 32% reached the medium category, and none reached the high category. This distribution reflects a fundamental inconsistency in the effectiveness of conventional instruction.

The six students in Class B whose scores regressed (N-Gain < 0) were analysed individually to understand why conventional instruction failed in their case (Table 8). Four of the six had pre-test scores of 80 or higher, suggesting that conventional instruction has a systematic weakness in maintaining and developing the ability of high-scoring learners. Without feedback that identifies and addresses specific weaknesses at higher levels of competence, already proficient learners tend to stagnate or even regress.

Table 8. Cases of Score Regression in Class B

No.	Student	Pre	Post	N-Gain	Possible Cause
1	S1	91.0	77.5	−1.50	Very high pre-test score; ceiling effect without specific feedback
2	S2	55.5	54.5	−0.02	Score almost stagnant; low engagement in practice
3	S3	90.0	67.0	−2.30	Dramatic regression (90 → 67); decline in structural accuracy
4	S4	85.0	82.0	−0.20	Moderate decline from a high initial score
5	S5	84.5	74.5	−0.65	Considerable regression (−10 points) without targeted intervention
6	S6	92.0	60.0	−4.00	Deepest regression (−32 points); stagnation of a proficient learner

Note: Student names have been anonymised.

4.6. Discussion

All operational hypotheses derived from the theoretical framework were confirmed by the empirical data through the three postulated causal pathways. The cognitive pathway, which links instant feedback to the internalisation of grammatical rules, is supported by the significant improvement in Class A ($t(24) = -8.812$, $p < 0.001$) together with the narrowing of the standard deviation from 12.664 at pre-test to 7.776 at post-test. This pattern suggests that the instant feedback provided by FelloFisch effectively corrected recurring error patterns (Ferris, 2011; Kang & Han, 2015).

The scaffolding pathway, which positions FelloFisch as a mediator within the zone of proximal development (Vygotsky, 1978; Wood et al., 1976), is supported by the mean N-Gain of 0.403 in Class A compared with −0.110 in Class B (Mann–Whitney U, $p = 0.015$). FelloFisch therefore appears to move learners beyond their actual level towards their potential developmental level more efficiently than conventional instruction. The ecological validity pathway, which links criterion-aligned design to a minimal transfer gap, is supported by the large effect size ($d = 0.93$): practice aligned with the Goethe-Institut rubric transferred directly to higher scores on an instrument that uses the same rubric.

These findings are consistent with, and in some respects exceed, the results of previous CALL research. The effect size of $d = 0.93$ is higher than the average effect of technology-based feedback of approximately $d = 0.50$–0.65 reported by Golonka et al. (2014). This difference may be explained by domain-specificity: most studies in that review used generic platforms, whereas FelloFisch was designed specifically for the Goethe-Zertifikat standard. The finding strengthens the argument that domain-specificity is a significant moderating variable in determining the effectiveness of CALL.

The negative N-Gain values among high-scoring learners in Class B add a perspective that has rarely been discussed in the literature. Schmidt (2024) noted that generic AI is ineffective in sustaining the motivation of advanced learners because its feedback is not sufficiently specific or challenging. FelloFisch, with its structured assessment criteria and clear targets based on the GZ B1 rubric, appears to have maintained the engagement of learners across the full range of ability, as shown by the absence of negative N-Gain in Class A.

The Indonesian context adds an important dimension to the interpretation of these findings. Because Indonesian has no case system, no grammatical gender, and a far more flexible word order than German, Indonesian learners are prone to systematic negative transfer, particularly in the accusative and dative cases, verb-second word order, and subordinate clause structure. By annotating errors at specific locations in the text, FelloFisch appears to interrupt this cycle of negative transfer more effectively than general manual correction. To the best of our knowledge, this study is the first to document the effectiveness of a domain-specific CALL intervention for Indonesian-speaking learners in the context of GZ B1 certification, an empirical contribution of direct relevance to German language study programmes in Indonesia.

5. CONCLUSION

This study examined the effectiveness of FelloFisch, a domain-specific CALL platform, in preparing Indonesian learners for the Goethe-Zertifikat B1 writing examination. Starting from equivalent initial conditions, students who practised with FelloFisch improved significantly more than students who received conventional instruction with manual correction, as shown by higher post-test scores (83.62 vs. 75.06; $p = 0.002$), a higher mean N-Gain (0.403 vs. −0.110; $p = 0.015$), and a large effect size ($d = 0.93$). The improvement in the experimental group was also more consistent: 92% of the students reached the medium or high N-Gain category and none regressed, whereas 24% of the control group regressed, most of whom were high-scoring learners.

These findings support the three postulated pathways through which FelloFisch operates: instant feedback that helps learners internalise grammatical rules, scaffolding within the zone of proximal development, and criterion-aligned practice that transfers directly to examination performance. They also suggest that domain-specificity is an important moderator of CALL effectiveness and that criterion-based, location-specific feedback can help Indonesian learners overcome negative transfer in areas such as case marking, verb-second word order, and subordinate clauses. Pedagogically, the results support the integration of FelloFisch into Goethe-Zertifikat B1 preparation courses in German language study programmes.

This study has several limitations. The use of intact classes did not allow full random assignment, so selection bias cannot be ruled out entirely, although the equivalence of the groups at pre-test was verified statistically. The study was conducted in a single study programme at one institution with 50 participants, which limits the generalisability of the findings. Future research should complete the analysis of the User Experience data and its relationship with individual N-Gain, decompose scores according to the four Goethe-Institut criteria to identify the criteria most responsive to the intervention, and extend the design to multi-institutional samples and higher certification levels such as B2 within a longitudinal framework.

Author contribution statement

Author 1: Conceptualization, Methodology, Instrument Design, Formal Analysis, Writing – Original Draft, Project Administration.
Author 2: Validation of the Instrument based on the Goethe-Institut Rubric, Writing – Review & Editing.
Author 3: Investigation (Pre-test and Post-test Data Collection), Data Curation (User Experience Questionnaire).

Acknowledgements

This research was funded by the Universitas Pendidikan Indonesia research grant (Hibah Penelitian UPI) under the Research Group Collaboration Programme. The authors would like to thank Prof. Dr. Marion Grein of Johannes Gutenberg-Universität Mainz, Germany, for her expert validation of the research instrument, methodological guidance, and internal review of the manuscript. The authors also thank the students of the German Language Education Study Program at Universitas Pendidikan Indonesia who participated in this study.

REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.

Ferris, D. R. (2011). *Treatment of error in second language student writing* (2nd ed.). University of Michigan Press.

Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2015). *Profile deutsch: Lernzielbestimmungen, Kannbeschreibungen, kommunikative Mittel* (Neubearbeitung). Langenscheidt.

Goethe-Institut. (2023). *Goethe-Zertifikat B1: Prüfungsziele, Testbeschreibung*. Goethe-Institut e.V.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning, 27*(1), 70–105. https://doi.org/10.1080/09588221.2012.700315

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64–74. https://doi.org/10.1119/1.18809

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hubbard, P. (2022). An invitation to CALL: Foundations of computer-assisted language learning. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). Wiley-Blackwell.

Hyland, K. (2019). *Second language writing* (2nd ed.). Cambridge University Press.

Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal, 99*(1), 1–18. https://doi.org/10.1111/modl.12189

Rankin, Y. A., McNeil, M., & Shute, V. J. (2022). Digital feedback in German L2 writing: Efficiency and accuracy gains in a computer-assisted environment. *Language Learning & Technology, 26*(2), 45–68.

Schmidt, J. (2024). Artificial intelligence in German language education: Opportunities and limitations of generative AI for DaF learners. *International Journal of Educational Technology, 18*(3), 45–60.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching, 31*(2), 57–71. https://doi.org/10.1017/S0261444800012970

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x