Jupiter: Jurnal Pendidikan Terapan
Homepage : https://journal.diginus.id/jupiter/index
Email: jurnalpendidikanterapan@gmail.com

p-ISSN : 2985-3214; e-ISSN : 2964-3171
Jupiter, Vol. 4, No. 5, November 2026
Page 639-648

A Single-Cohort Evaluation of Online Mandarin Tutoring Using the Kirkpatrick Framework: Evidence and Measurement Limitations

Rika Fuziyah[1*], Henny Sanulita[2], Dedi Triyanto[3], Urai Salam[4], Eva Faja Ripanti[5]
[1,2,3,4,5]Universitas Tanjungpura, Pontianak, Indonesia

ARTICLE INFO

Article history:
Submitted: October 19, 2026
Final Revised: November 09, 2026
Accepted: November 11, 2026
Published: November 22, 2026

Keywords:
Program Evaluation; Kirkpatrick Model; Online Mandarin; Learning Behavior

ABSTRACT

Purpose: This study examined what the available evidence can establish about an intensive online Mandarin tutoring program when participant reactions, archived assessment scores, classroom-period behavior ratings, and internal examination outcomes are organized through the Kirkpatrick framework. **Methods:** A quantitative-dominant, single-cohort program evaluation analyzed linked records from 20 participants. Evidence comprised three non-equivalent assessment occasions, a 50-item end-of-course questionnaire with ten open questions, and observation ratings at three course stages. Descriptive statistics, exploratory normalized gain, repeated-measures ANOVA with Greenhouse–Geisser correction, pairwise comparisons, and a Friedman test were used. Open responses and observation notes were descriptively coded and compared with quantitative patterns. **Findings:** Mean archived scores were 29.15, 88.68, and 94.35 across the three occasions; the mean individual normalized gain was 0.9164. The repeated-measures tests identified differences in recorded scores, not calibrated proficiency change. Behavioral ratings increased overall, most notably for independent practice, while speaking confidence changed less consistently. All participants exceeded the provider's internal simulation threshold; this result and the post-test score came from the same final assessment. Instructor ratings were favorable, but active participation and workload were weaker areas. Fifteen participants indicated an intention to continue. **Conclusion and contribution**: The Kirkpatrick framework helped organize multiple indicators while exposing important dependencies and gaps in the evidence. The study contributes a measurement-focused account of how single-cohort online language-program evaluations should distinguish favorable reactions and internal performance from independently verified transfer, standardized proficiency, and longer-term impact. The small cohort, lack of a comparison group, non-equivalent tests, and incomplete ethics documentation limit causal and generalizable claims.

Doi: https://doi.org/10.61255/jupiter.v4i5.2347

INTRODUCTION

Online language courses are often judged through learner satisfaction and end-of-course scores. Neither indicator alone establishes whether learners can use a language beyond the tasks practiced during instruction. For providers, evaluation must therefore answer two related questions: what participants experienced and learned during the course, and whether the evidence is sufficiently valid and independent to support claims about transfer or wider program results. These questions are particularly important in intensive Mandarin tutoring, where synchronous lessons, independent practice, and examination preparation may contribute different kinds of evidence.

Research syntheses suggest that online and videoconference-based language learning can support second-language development, but they also caution against treating delivery mode as a sufficient explanation of outcomes. Yu (2022) reported positive effects of videoconferencing on second-language learning while noting limitations in the underlying experimental evidence. Zhou and Zhang (2022) identified heterogeneous outcome measures and a concentration of research on English in online and blended language learning. Indonesian studies have also examined learners' perceptions of online language learning, including before and during the COVID-19 pandemic (Rahmayanti

et al., 2022). Together, this literature supports evaluating both learning-related indicators and the conditions under which online instruction is experienced.

The Kirkpatrick framework distinguishes reaction, learning, behavior, and results. Its practical value is that it prompts evaluators to ask different questions rather than treating satisfaction or test scores as a complete account of program quality. Applications in education and training have adapted the framework to context-specific indicators and instruments (Alsalamah & Callinan, 2021; Dalimunthe, 2022; Purba & Maulana, 2023). However, the framework does not itself guarantee that indicators are valid, independent, or causally linked. Reio et al. (2017) caution against assuming that success at one level automatically establishes success at subsequent levels.

This issue is especially relevant when a program uses one final examination to represent both learning and achievement of a provider-defined target. A post-test average and a pass rate may appear to be separate indicators, but if both are calculated from the same examination they are summaries of a shared data source. They cannot independently corroborate one another. Similarly, a stated intention to continue is not equivalent to actual reenrollment, and a classroom-period behavior rating is not direct evidence of sustained language use after instruction.

Mandarin learning also requires attention to skill-specific assessment. Research on computer-assisted feedback has examined how visual and audio feedback can support perception and production of Mandarin lexical tones (Chen, 2024). More recent online classroom research has compared corrective-feedback approaches for Mandarin tone learning and highlighted the importance of task design and learner differences (Dong et al., 2025). These studies illustrate why broad total scores may conceal differences among listening, reading, writing, pronunciation, and spontaneous speaking. They do not, however, provide evidence about the particular cohort evaluated in the present study.

Independent practice can be interpreted in relation to self-regulated learning, which involves planning, monitoring, and evaluating learning rather than simply spending time outside class. Panadero (2017) reviews major models of self-regulated learning, while Edisherashvili et al. (2022) identify support needs across preparation, performance, and appraisal in distance-learning contexts. Chen (2022) synthesized evidence on self-regulated-learning interventions in second-language learning. These studies provide a conceptual basis for interpreting practice-related observations, but the present observation scale did not directly measure all components of self-regulation.

Affective factors may also shape participation in language learning. Teimouri et al. (2019) reported a negative association between second-language anxiety and achievement, while Botes et al. (2022) examined the relationship between foreign-language enjoyment and learning-related outcomes. The present study did not administer direct measures of anxiety or enjoyment; therefore, speaking-confidence ratings should not be treated as proxies for either construct. Instead, they are interpreted as course-period observations that can inform practical decisions about opportunities to speak and receive feedback.

Against this background, the present study evaluates one online Mandarin tutoring cohort at Cetta Course Institution through a bounded application of the Kirkpatrick framework. It asks: (1) how participants rated course features and their intention to continue; (2) how archived scores differed across three assessment occasions; (3) how three observed behaviors varied during the course; and (4) what immediate internal simulation outcomes and qualitative records indicate about program improvement. The study does not seek to validate a new Kirkpatrick model or estimate a causal treatment effect. Its contribution is methodological and practical: it demonstrates how an explicit audit of indicator quality, overlap, and evidential independence changes the interpretation of a four-level evaluation in an online Mandarin program.

METHOD

This study used a quantitative-dominant, single-cohort program-evaluation design. Archived administrative scores and structured ratings provided the primary descriptive evidence; end-of-course open responses and observation notes were used to contextualize patterns and identify improvement priorities. The qualitative material was not treated as a separate phenomenological inquiry or as evidence of thematic saturation. Integration was undertaken by comparing descriptive themes with the corresponding Kirkpatrick-level indicators and noting where the sources converged or differed.

The analysis included all 20 participants for whom linked assessment, questionnaire, and observation records were available. The program was described as a three-month course with 60 scheduled sessions of 180 minutes each (180 scheduled hours). Recorded attendance ranged from 30 to 60 sessions ($M = 54.90$, $SD = 8.03$), indicating variation in instructional exposure. The available participant records included four learners aged 15–20 years; the remaining participants were adults. Because the archive did not document a comparison group or a probability-based recruitment procedure, the cohort is treated as a complete set of available linked records rather than a representative sample of online Mandarin learners.

Assessment records were matched by participant name and then coded P1–P20 for analysis. The pre-test archive contained 30 language-response items and a recorded total. The mid-test score was calculated as the arithmetic mean of speaking and written scores. The archived post-test score was the whole-number average recorded in the final HSK 3 simulation workbook; it was not a separate assessment instrument. The simulation total was recalculated from listening, reading, and writing components, excluding the separately reported supplementary speaking score. One stored addition error was corrected during the audit. The three assessment occasions used different formats and no common-item, equating, or external validation procedure was documented; therefore, their scores were analyzed as recorded performance on successive course assessments, not as equivalent measures of Mandarin proficiency.

The end-of-program questionnaire contained 50 closed items scored from 1 to 5 and ten open-ended questions. Items were grouped into six descriptive domains: materials, methods, instructors, facilities, perceived outcomes, and online learning. Item Q48 asked about preference for combining online and face-to-face learning and was reported separately rather than interpreted as a positive or negative rating of online-course quality. Internal-consistency coefficients were recalculated for the descriptive groups. No factor analysis, confirmatory construct-validity test, or documented expert-validation record was available in the supplied archive. Accordingly, alpha coefficients are reported as exploratory item-consistency estimates, not as evidence that the domains are validated scales.

The observation archive rated speaking confidence, independent practice, and application beyond class on a four-point scale at sessions 1–4, 29–32, and 57–60. The general anchors ranged from very passive (1) to very active or consistently initiating (4); separate behavior-specific anchors were not supplied. The available records did not identify a second independent rater or provide inter-rater reliability statistics. Ratings are therefore treated as descriptive course-period observations rather than validated measures of transfer. Because the session labels refer to course sequence and are not fully consistent with the stated calendar duration, comparisons are described by observation stage rather than by exact elapsed weeks.

Means, sample standard deviations, ranges, and participant counts were recalculated from the supplied spreadsheets. Individual normalized gain was computed as $g = (post - pre)/(100 - pre)$; its mean is presented only as an exploratory descriptive index using the conventional high-gain label ($g \geq .70$), not as a validated Mandarin-proficiency threshold. Repeated-measures ANOVA was used to test whether recorded scores differed across the three occasions. Mauchly's test indicated a sphericity violation, so Greenhouse–Geisser-corrected degrees of freedom are reported. Bonferroni-adjusted paired comparisons and a tie-corrected Friedman test were used as complementary analyses. These tests assess within-cohort score differences; they do not address assessment equivalence, confounding, or causal program effects. Missingness was handled by complete linked records (N = 20); no imputation was performed. Open responses and observation notes were read, assigned descriptive content codes, grouped into recurring instructional topics, and compared with quantitative patterns. Coding was conducted for descriptive contextualization by the manuscript author; no independent coder agreement or saturation claim is made.

RESULTS

Participant reactions and continuation intentions

Across the 50 closed questionnaire items, the raw response average was 4.323 (SD of participant averages = 0.312). This pooled figure is reported only as a descriptive profile because the questionnaire combines different constructs and includes a format-preference item; it is not interpreted as a validated global satisfaction scale. Table 1 therefore reports domain-level summaries and exploratory alpha coefficients.

Table 1. Participant ratings by questionnaire domain

Domain	Items	M	SD	Alpha
Materials	8	4.33	0.35	0.689
Methods	8	4.24	0.50	0.827
Instructors	8	4.61	0.27	0.648
Facilities	8	4.38	0.28	0.486
Perceived outcomes	8	4.41	0.35	0.631
Online learning excluding Q48	9	4.03	0.49	0.789

Note. N = 20. Responses range from 1 to 5. Online learning excludes Q48. Alpha estimates are exploratory and based on a small cohort.

Item-level patterns qualify the favorable domain averages. Active participation (Q10) averaged 3.50, and perceived proportionality of workload within a 180-minute session (Q7) averaged 3.70. Instructor use of understandable language (Q24) averaged 4.85, and command of the subject (Q20) averaged 4.75. The contrast suggests that participants evaluated instructor competence more favorably than some aspects of participation and course pacing. The facilities

domain had the lowest alpha (0.486), so its mean should be interpreted descriptively and not as a reliable latent construct.

For Q37, six participants selected 5, nine selected 4, three selected 3, and two selected 2; none selected 1. Thus, 15 of 20 participants (75%) expressed a positive intention to continue (responses 4–5), while three were neutral and two selected a lower response. These responses record stated intentions at the end of the course; they do not establish subsequent enrollment or retention.

Recorded assessment performance

Recorded mean scores were 29.15 at baseline, 88.68 at midpoint, and 94.35 at the final occasion (Table 2). The largest numerical difference occurred before the midpoint. Because the assessment formats differed, these values describe the archived scores on three successive assessment arrangements and should not be read as a calibrated proficiency scale. The narrowing dispersion and high final scores also indicate a restricted upper range at the final occasion.

Table 2. Archived assessment scores across three occasions

Occasion	M	SD	Minimum	Maximum
Pre-test	29.15	8.39	17.00	51.00
Mid-test	88.68	6.19	78.50	99.00
Post-test	94.35	5.21	83.00	100.00

Note. N = 20. Mid-test averages speaking and written scores. Post-test is the archived rounded simulation average and is not an independent assessment.

Mean individual normalized gain was 0.9164 (SD = 0.0817; range = 0.7347–1.0000). All 20 values met the conventional descriptive high-gain category; 14 were at least .90 and seven were at least .96. Sensitivity summaries excluding maximum or very high values remain descriptive. Because the pre-test and post-test formats were not established as equivalent, normalized gain is not interpreted as a valid estimate of proficiency growth or program effectiveness.

The archived SPSS output indicated a sphericity violation, $W = .666$, $\chi^2(2) = 7.311$, $p = .026$. Reanalysis reproduced the Greenhouse–Geisser-corrected result, $F(1.499, 28.490) = 645.262$, $p < .001$, partial $\eta^2 = .971$ ($\varepsilon = .750$). All three Bonferroni-adjusted paired comparisons were statistically significant (Table 3). The tie-corrected Friedman statistic was $\chi^2(2) = 35.795$, $p < .001$. These analyses show that the recorded scores differed across occasions; they do not establish equal test difficulty, construct equivalence, or a causal effect of the program.

Table 3. Pairwise comparisons of archived assessment scores

Comparison	Mean increase	t(19)	Adjusted p	dz
Pre to mid	59.53	28.94	< .001	6.47
Mid to post	5.68	4.08	.002	0.91
Pre to post	65.20	26.69	< .001	5.97

Note. N = 20. Bonferroni adjustment covers three comparisons. Effect sizes quantify within-person recorded-score differences and are not treatment effects.

Behavior during the course

The composite observation mean increased from 1.62 at the beginning to 2.53 at the middle and 3.12 at the end. Independent-practice ratings showed the largest raw increase (+2.00), followed by application beyond class (+1.55) and speaking confidence (+0.95). Independent-practice ratings increased for all 20 participants. Speaking-confidence ratings increased for 13, remained unchanged for five, and declined for two (Table 4). These patterns are descriptive because the observation rubric used general rather than construct-specific anchors and had no documented independent-rater reliability.

Table 4. Behavioral indicators across three occasions

Indicator	Beginning M (SD)	Middle M (SD)	End M (SD)	Up / same / down
Speaking confidence	2.20 (0.62)	2.65 (0.59)	3.15 (0.81)	13 / 5 / 2
Independent practice	1.35 (0.49)	2.50 (0.51)	3.35 (0.75)	20 / 0 / 0
Application beyond class	1.30 (0.47)	2.45 (0.51)	2.85 (0.88)	17 / 3 / 0

Note. N = 20. Scale = 1–4. Counts compare beginning with end. Outside-class application is an observer-rated indicator without independent verification.

Open responses and observation notes provided context for the quantitative patterns. P1 requested more conversation practice and examples of everyday use. P5 described pronunciation and tones as difficult and requested additional

examples and practice. P11 requested more Hanzi exercises and one-to-one speaking. P13 identified long explanations without intervening practice as a source of reduced focus. P4 and P16 described lesson recordings as useful for review. These illustrative comments point to practical design priorities; they are not treated as exhaustive themes or evidence that a particular feature caused an outcome.

Immediate examination outcomes and evidence integration

All 20 participants exceeded the provider's internal simulation threshold of 180/300. After recalculating totals from listening, reading, and writing, the mean was 282.45 (SD = 15.22; range = 250–300). Speaking was assessed separately as a supplementary component. Writing and speaking had lower means and wider dispersion than listening and reading (Table 5). Because component difficulty and scoring were not equated, these differences are descriptive rather than proof of corresponding differences in underlying skill.

Table 5. Audited final simulation results

Component	M	SD	Range
Listening	98.35	3.51	90–100
Reading	97.60	5.31	80–100
Writing	86.50	12.68	50–100
Speaking reported separately	85.75	9.88	60–98
Corrected total	282.45	15.22	250–300

Note. N = 20. Each component is scored on a 100-point scale. The 300-point total excludes speaking and corrects one stored addition error. This is an internal simulation, not official certification.

Using the unrounded corrected total divided by three as the final average yielded M = 94.15 and a mean individual normalized gain of 0.9136 (SD = 0.0799). This sensitivity calculation is close to the archived rounded-score analysis, indicating that rounding and the single addition error do not account for the descriptive score pattern. However, the post-test average and the internal simulation pass rate derive from the same final assessment source. They must not be presented as independent corroboration.

Table 6. Interpretation of evidence across Kirkpatrick levels

Level	Evidence in this cohort	Boundary of inference
Reaction	Favorable instructor and satisfaction ratings; weaker participation and workload ratings	Self-reports; heterogeneous questionnaire groups
Learning	Higher recorded scores across three occasions	Different formats; final score near upper limit
Behavior	More independent practice; uneven confidence changes	Ratings during the course; no independent transfer verification
Results	20/20 internal passes; 15/20 positive continuation responses	Pass indicator shares final-test data; intention is not reenrollment

DISCUSSION

Interpreting high scores within the limits of the assessments

The principal quantitative finding is that the archived scores differed substantially across the three assessment occasions. This pattern was visible in descriptive means and in both parametric and rank-based repeated-measures analyses. For the provider, the records indicate strong performance on the course's later assessment tasks. The result is meaningful as a description of this cohort's recorded performance, but its interpretation depends on the properties of the assessments.

The three score occasions were not equivalent forms. The baseline archive contained 30 language-response items, the midpoint combined speaking and written scores, and the final score was derived from the internal HSK 3 simulation. A common 0–100 scale does not establish common content coverage, difficulty, or scoring standards. The statistically significant differences therefore concern scores recorded under different assessment arrangements, not a calibrated trajectory of Mandarin proficiency. Future evaluations should use parallel forms, common anchor tasks, or a documented linking procedure before interpreting score change as comparable growth.

The normalized-gain estimate has the same limitation. The mean individual value of 0.9164 summarizes the relationship between the archived baseline and final scores under the selected formula. Conventional gain categories are not Mandarin proficiency standards and should not be described as validated effectiveness thresholds. In addition, normalized gain is sensitive to the initial score and the scale's upper limit. It is therefore retained as a supplementary descriptive index, not as the primary evidence that the program caused learning gains.

The Greenhouse–Geisser-corrected ANOVA and Friedman test provide convergent evidence that the recorded score distributions differed across occasions. The correction addresses the sphericity assumption; it does not correct non-equivalent tests, the absence of a comparison group, or unmeasured influences such as differences in attendance and prior learning. Likewise, partial η^2 and paired-sample effect sizes quantify within-cohort differences on these recorded scores. They are not estimates of a treatment effect and should not be compared directly with controlled intervention effects.

The smaller numerical increase from midpoint to final assessment may reflect reduced room for improvement near the upper end of the scale, differences in task demands, or both. The available archive cannot separate these explanations. A more discriminating assessment design should include sufficiently challenging tasks and comparable measures at each occasion, with skill-specific scoring criteria. Such a design would make future results more useful for identifying learners who need additional support.

Behavioral development and opportunities to use Mandarin

Independent-practice ratings increased more than the other observed behaviors, whereas speaking confidence developed less consistently. Although every participant's independent-practice rating increased, two participants' speaking-confidence ratings declined. This contrast supports reporting each indicator separately rather than relying only on a composite score. It does not establish that independent practice was effective self-regulation, because the observation form did not separately measure planning, monitoring, or reflection.

Self-regulated-learning literature offers a framework for interpreting, but not explaining causally, the observed practice pattern. Panadero (2017) describes regulation as a coordinated process, and Edisherashvili et al. (2022) identify support needs across phases of distance learning. Chen (2022) synthesized evidence on self-regulated-learning interventions in second-language learning. Participants' comments about lesson recordings suggest that review resources were valued, but the archive does not show how frequently recordings were used or whether review involved active retrieval and reflection. Providers could connect recordings with short practice goals and simple learning logs in future cohorts.

Research on Mandarin pronunciation provides a more directly relevant basis for skill-specific instructional recommendations. Chen (2024) examined computer-aided visual and audio feedback for Mandarin tone perception and production, while Dong et al. (2025) investigated corrective-feedback approaches in online classroom Mandarin tone learning. These studies support the relevance of targeted pronunciation practice and feedback as design options. They do not establish that such interventions were present in, or caused outcomes for, the cohort evaluated here.

Speaking confidence should also be interpreted narrowly. Meta-analytic evidence links second-language anxiety with achievement (Teimouri et al., 2019), and foreign-language enjoyment is associated with communication and learning outcomes (Botes et al., 2022). However, the present observation ratings did not measure anxiety or enjoyment. It would therefore be speculative to attribute lower confidence to either emotional construct. Participants' requests for conversation, individual speaking opportunities, and shorter explanations instead provide direct, actionable evidence for piloting more structured speaking tasks.

The application-beyond-class indicator is weaker evidence of transfer. An online classroom observer has limited access to participants' everyday language use, and the available notes do not independently verify all reported activity outside class. Kirkpatrick-based evaluations in other educational settings illustrate the importance of specifying the source and operational definition of each level's evidence (Dalimunthe, 2022; Purba & Maulana, 2023). In this study, the indicator should be described as a course-period rating relevant to possible transfer, not as demonstrated sustained transfer. Future work could collect consented practice logs or task products and follow participants after the course.

What reaction and simulation outcomes reveal about program quality?

Favorable instructor ratings coexisted with weaker ratings for active participation and workload. This distinction suggests that instructor competence and course design should not be collapsed into a single judgment. The facilities domain also had the lowest internal-consistency coefficient (α = .486), while several other domains showed only modest consistency. Because the questionnaire's construct validity was not established, domain means and alpha values are exploratory descriptions of item responses, not validated latent measures. The item-level results and open comments are consequently more useful for immediate program improvement than a single global satisfaction score.

The response to Q48 also requires a separate interpretation. Preference for a combined online and face-to-face format does not necessarily indicate dissatisfaction with every aspect of online learning. The item was therefore excluded from the online-learning domain rather than reverse-scored. Comments about focus, connectivity, and long explanations suggest testing shorter instructional segments and more frequent interactive tasks. A future evaluation could compare these design changes using consistent measures.

Research syntheses by Yu (2022) and Zhou and Zhang (2022) support attention to interaction and assessment design in online and blended language learning. They do not justify attributing this cohort's recorded performance to videoconferencing itself or assuming that blended delivery would necessarily improve outcomes. The attendance range of 30–60 sessions further indicates that participants did not receive identical instructional exposure. Because the available analysis did not estimate attendance-adjusted effects, attendance should be treated as a contextual factor rather than an explanation of individual outcomes.

The internal simulation provides evidence that all participants exceeded the provider-defined threshold. The component profile also shows that writing and supplementary speaking scores were more variable than listening and reading. Because the component tasks were not equated, the differences should guide closer inspection of task demands and practice allocation, not be interpreted as calibrated differences in underlying competence. The simulation was internal and should not be described as an official HSK result or proof of readiness for external certification.

Fifteen participants expressed an intention to continue, but intention is not actual reenrollment. This indicator can inform follow-up communication and program planning, but it cannot be reported as a retention rate. Likewise, the internal simulation pass rate and the post-test mean are not independent results because they share the same final assessment source. These distinctions are central to interpreting the Kirkpatrick levels without overstating the strength of the available evidence.

Contribution to evaluation and implications for future evidence

The study's principal contribution is not a new or validated version of the Kirkpatrick model. Rather, it provides a measurement-focused demonstration of how the framework can organize a small online Mandarin program evaluation while making indicator overlap and evidential limits explicit. Previous educational applications show that the framework can be adapted to different training contexts (Alsalamah & Callinan, 2021; Dalimunthe, 2022; Purba & Maulana, 2023). The present case adds a language-program example in which the same final assessment contributes to both the post-test and internal pass indicator, while behavior ratings and continuation intentions do not independently establish post-course transfer or retention.

A key implication is that having several indicators is not the same as having independent evidence sources. A post-test mean and a pass percentage may answer different descriptive questions, but both were generated from the same final examination. Similarly, behavior ratings collected during instruction do not establish durable transfer, and favorable reactions do not guarantee learning outcomes. This interpretation is consistent with Reio et al. (2017), who caution against treating the four levels as an automatic causal sequence. The framework is most useful here as an organizing structure for evidence and improvement decisions.

The qualitative records added context where numerical summaries alone were insufficient. Requests for more conversation practice, pronunciation examples, Hanzi exercises, and shorter explanations identify specific areas for course redesign; positive comments about recordings identify a resource learners valued. These are illustrative descriptive categories derived from the available open responses and observation notes. The small dataset does not support claims of thematic saturation, population prevalence, or causal relationships between reactions, behavior, and assessment scores.

Several limitations constrain interpretation. The cohort comprised 20 participants, had no comparison group, and was not shown to be representative of a wider learner population. Assessment formats differed across occasions, the final scores were concentrated near the upper end, and the questionnaire had no documented construct-validation procedure. Some domain alpha coefficients were modest or low. Behavioral ratings used general anchors and had no documented second-rater reliability; observation labels were inconsistent with calendar-based timing. Attendance varied, outside-class language use was not independently verified, and the final simulation was not an official examination. The supplied archive also did not contain verifiable ethics-approval or participant-consent documentation. These limitations are reported rather than reconstructed through assumptions.

Future evaluation should prioritize four improvements: (1) use parallel assessment forms or common anchor tasks and report skill-specific scoring criteria; (2) establish questionnaire content and construct validity before interpreting domain scores, and shorten or refine items with weak consistency; (3) operationalize behavioral indicators with specific anchors and obtain independent ratings for a subset; and (4) follow participants after the course to document actual language use, reenrollment, and external examination outcomes. Attendance and exposure should also be recorded consistently. These steps would strengthen the independence, comparability, and practical value of evidence at each Kirkpatrick level.

CONCLUSION

The evaluation found favorable instructor ratings, higher recorded scores across three assessment occasions, and increased course-period ratings for independent practice and other behaviors. All 20 participants exceeded the provider's internal simulation threshold, while 15 expressed an intention to continue. The findings also identify improvement needs in active participation, workload, speaking opportunities, and skill-specific practice. The study's contribution is to show how the Kirkpatrick framework can organize evidence while exposing limits in its independence and measurement quality. Because assessment formats differed, the final post-test and simulation shared one source, and no comparison group was available, the score differences cannot establish causal program effectiveness, calibrated proficiency growth, official HSK attainment, or sustained transfer. Future evaluations should prioritize comparable assessments, validated and shorter questionnaires, behaviorally anchored observation with independent ratings, and post-course follow-up. The conclusions are specific to this 20-participant cohort and should not be generalized beyond the available evidence.

REFERENCE

Alsalamah, A., & Callinan, C. (2021). Adaptation of Kirkpatrick's four-level model of training criteria to evaluate training programmes for head teachers. *Education Sciences*, *11*(3), 116. https://doi.org/10.3390/educsci11030116

Alsalamah, A., & Callinan, C. (2022). The Kirkpatrick model for training evaluation: Bibliometric analysis after 60 years (1959–2020). *Industrial and Commercial Training*, *54*(1), 36–63. https://doi.org/10.1108/ict-12-2020-0115

Amin, M., Sibuea, A. M., & Mustaqim, B. (2022). The effectiveness of online learning using e-learning during pandemic Covid-19. *Journal of Education Technology*, *6*(2). https://doi.org/10.23887/jet.v6i2.44125

Arthur, R. (2018). Evaluasi program diklat karya tulis ilmiah untuk widyaiswara Pusbangtendik Kemdikbud. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 35–48. https://doi.org/10.21831/pep.v22i1.16749

Botes, E., Dewaele, J.-M., & Greiff, S. (2022). Taking stock: A meta-analysis of the effects of foreign language enjoyment. *Studies in Second Language Learning and Teaching*, *12*(2), 205–232. https://doi.org/10.14746/ssllt.2022.12.2.3

Chen, J. (2022). The effectiveness of self-regulated learning (SRL) interventions on L2 learning achievement, strategy employment and self-efficacy: A meta-analytic study. *Frontiers in Psychology*, *13*, 1021101. https://doi.org/10.3389/fpsyg.2022.1021101

Chen, M. (2024). Computer-aided feedback on the pronunciation of Mandarin Chinese tones: Using Praat to promote multimedia foreign language learning. Computer Assisted Language Learning, 37(3), 363–388. https://doi.org/10.1080/09588221.2022.2037652

Dalimunthe, M. B. (2022). Kirkpatrick four-level model evaluation: An evaluation scale on the preservice teacher's internship program. Journal of Education Research and Evaluation, 6(2), 367–376. https://doi.org/10.23887/jere.v6i2.43535

Dewi, L. R., & Kartowagiran, B. (2018). An evaluation of internship program by using Kirkpatrick evaluation model. *REID (Research and Evaluation in Education)*, *4*(2), 155–163. https://doi.org/10.21831/reid.v4i2.22495

Dong, Z. R., Han, C., & Li, S. (2025). Feedback explicitness, working memory, and explicit knowledge in online classroom-based second language Mandarin tone learning. Language Teaching Research. https://doi.org/10.1177/13621688251378559

Edisherashvili, N., Saks, K., Pedaste, M., & Leijen, Ä. (2022). Supporting self-regulated learning in distance learning contexts at higher education level: Systematic literature review. *Frontiers in Psychology*, *12*, 792422. https://doi.org/10.3389/fpsyg.2021.792422

Ernawati, E., Dewi, M., Rosalina, L., Ferdian, F., & Andres, F. S. (2022). An evaluation of e-learning of entrepreneurship course: Learning alternative during Covid-19 pandemic for university students. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *26*(1). https://doi.org/10.21831/pep.v26i1.44073

Fahruddin, F., Jana, P., Setiawan, J., Rochmat, S., Aman, A., & Yuliantri, R. D. A. (2022). Student perception of online learning media platform during the Covid-19 pandemic. *Journal of Education Technology*, *6*(1), 126–132. https://doi.org/10.23887/jet.v6i1.42738

Firmansyah, F., Rahayu, W., & Nurjannah, N. (2020). Evaluation of the entrepreneurship education program through extracurricular activities of Student Company. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(1). https://doi.org/10.21831/pep.v24i1.19783

McCrudden, M. T., Marchand, G., & Schutz, P. (2019). Mixed methods in educational psychology inquiry. *Contemporary Educational Psychology*, *57*, 1–8. https://doi.org/10.1016/j.cedpsych.2019.01.008

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Mukhlis, R., & Hiryanto. (2025). Evaluasi belajar pelatihan pemandu wisata gunung di Asosiasi Pemandu Gunung Indonesia. *Diklus: Jurnal Pendidikan Luar Sekolah*, *9*(1), 94–104. https://doi.org/10.21831/diklus.v9i1.86162

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Purba, A., & Maulana, A. D. (2023). The evaluation of field teaching practice using Kirkpatrick's model. FOUNDASIA, 14(1), 27–41. https://doi.org/10.21831/foundasia.v14i1.57790

Rahmayanti, P., Suwastini, N. K. A., Dantes, G. R., & Kultsum, U. (2022). Indonesian college students' perception toward language online learning before and during COVID-19 pandemic. Journal of Education Technology, 6(1), 56–66. https://doi.org/10.23887/jet.v6i1.41561

Ramadhan, M. A., Sumarsono, R. A., Achmad, A. A., & Cisse, A. (2022). Implementation of Kirkpatrick evaluation model in Building Information Modeling (BIM) training program. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *28*(2). https://doi.org/10.21831/jptk.v28i2.43555

Reio, T. G., Rocco, T. S., Smith, D. H., & Chang, E. (2017). A critique of Kirkpatrick's evaluation model. *New Horizons in Adult Education and Human Resource Development*, *29*(2), 35–53. https://doi.org/10.1002/nha3.20178

Simangunsong, J., & Purnomo, L. (2021). An evaluation of the Petty Officer Training program of the Republic of Indonesia police using the Kirkpatrick model. *REID (Research and Evaluation in Education)*, *7*(2), 88–105. https://doi.org/10.21831/reid.v7i2.36937

Sudiyatno, S., & Wulandari, I. (2020). Evaluation of education and training programs in Solo Technopark Central Java in Indonesia. *REID (Research and Evaluation in Education)*, *6*(2), 150–159. https://doi.org/10.21831/reid.v6i2.36794

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, *41*(2), 363–387. https://doi.org/10.1017/s0272263118000311

Waruwu, M. (2021). Evaluation of the online mode-based the candidate of headmaster training program by using the Kirkpatrick evaluation model. *Jurnal Pendidikan Dan Pengajaran*, *54*(3), 434–445. https://doi.org/10.23887/jpp.v54i3.35482

Wijaya, H., Sumule, L., Weismann, I. T. J., Supartini, T., & Tari, E. (2021). Online learning evaluation in higher education: Study survey method. *Journal of Education Technology*, *5*(3), 401–408. https://doi.org/10.23887/jet.v5i3.35466

Yu, L.-T. (2022). The effect of videoconferencing on second-language learning: A meta-analysis. *Behavioral Sciences*, *12*(6), 169. https://doi.org/10.3390/bs12060169

Zhou, T., & Zhang, W. (2022). Effectiveness study on online or blended language learning based on student achievement: A systematic review of empirical studies. *Sustainability*, *14*(12), 7303. https://doi.org/10.3390/su14127303

ACKNOWLEDGEMENT

The author thanks Universitas Tanjungpura and the Master's Program in Educational Technology for academic support. Appreciation is extended to Henny Sanulita and Dedi Triyanto for supervisory guidance, and to Cetta Course Institution and the participating learners for their contributions to the program evaluation.

AUTHOR CONTRIBUTION STATEMENT

RF is responsible for the study records, analysis and interpretation, manuscript preparation, and verification of the final article. ChatGPT (OpenAI) was used to assist with consistency checks across supplied files, statistical recalculation, bibliographic checks, and English-language revision. AI assistance did not collect participant data or replace the author's responsibility for the study. The author is responsible for checking the underlying records, analyses, interpretations, references, ethical disclosures, and final manuscript.

AI DISCLOSURE STATEMENT

ChatGPT (OpenAI) was used to assist with checking consistency across supplied files, developing code for statistical recalculation, verifying bibliographic metadata, and revising the English manuscript. AI assistance did not collect

participant data. The author is responsible for verifying the analyses, interpretations, references, ethical statements, and final content before submission.

***Rika Fauziah (Corresponding Author)**
Master's Program in Educational Technology, Faculty of Teacher Training and Education,
Universitas Tanjungpura, Pontianak
Jl. Prof. Dr. H. Hadari Nawawi, Bansir Laut, Kecamatan Pontianak Tenggara, Kota Pontianak, Kalimantan Barat 78124, Indonesia
Email: f2151251068@student.untan.ac.id

Henny Sanulita
Universitas Tanjungpura, Pontianak
Jl. Prof. Dr. H. Hadari Nawawi, Bansir Laut, Kecamatan Pontianak Tenggara, Kota Pontianak, Kalimantan Barat 78124, Indonesia
Email: henny.sanulita@fkip.untan.ac.id

Dedi Triyanto
Universitas Tanjungpura, Pontianak
Jl. Prof. Dr. H. Hadari Nawawi, Bansir Laut, Kecamatan Pontianak Tenggara, Kota Pontianak, Kalimantan Barat 78124, Indonesia
Email: dedi.triyanto@siskom.untan.ac.id

Urai Salam
Universitas Tanjungpura, Pontianak
Jl. Prof. Dr. H. Hadari Nawawi, Bansir Laut, Kecamatan Pontianak Tenggara, Kota Pontianak, Kalimantan Barat 78124, Indonesia
Email: urai@untan.ac.id

Eva Faja Ripanti
Universitas Tanjungpura, Pontianak
Jl. Prof. Dr. H. Hadari Nawawi, Bansir Laut, Kecamatan Pontianak Tenggara, Kota Pontianak, Kalimantan Barat 78124, Indonesia
Email: evaripanti@untan.ac.id