Curriculum Research
Volume 6, Issue 4
Dec. 15, 2025

Desirable Difficulties and long-term Learning Outcomes in German as a Foreign Language: Evidence from Iranian Learners

Article Type:

Original Research

Authors:

Shima Shahbazfar[1]
ORCiD: 0009-0001-9189-7946

Armin Fazelzad[2]
ORCiD: 0009-0005-7631-8466

Parastoo Panjehshahi[3]
ORCiD: 0000-0002-6794-0246

Article History:

Received: 2025.09.03

Accepted: 2025.12.07

Published: 2025.12.15

Abstract

This paper explores how the application of desirable difficulties—spacing, interleaving, retrieval, and generative learning—affects long-term retention and productive performance among Iranian learners of German as a Foreign Language (GFL). Building on a cognitive framework that distinguishes the durability of stored knowledge from its momentary accessibility, the study adopted a mixed-methods design with A1-level learners from Tehran. A total of twenty participants (N = 20) were randomly assigned to an experimental group (n = 10) receiving instruction based on desirable difficulties and a control group (n = 10) following massed-practice routines. Quantitative analysis showed that the experimental group had significantly higher retention and production skills scores in the delayed post-tests, and the results were statistically significant ($p < 0.01$). Qualitative data derived from semi-structured interviews with six volunteer learners from the experimental group indicated that participants associated increased cognitive challenge with deeper understanding and enhanced motivation. The results show that including challenging but beneficial activities helps make language knowledge more lasting and easier to use in different situations. This offers support for the idea of using spaced retrieval and interleaved practice in teaching foreign languages in a structured way to help learners develop strong and long-term language skills.

Key Words: Desirable Difficulties, Educational Design, German as a Foreign Language, Iranian Learners, Long-Term Retention

1. Department of German Language, CT.C., Islamic Azad University, Tehran, Iran. Email: shima.shahbazfar@iau.ir
2. Department of German Language, CT.C., Islamic Azad University, Tehran, Iran (Corresponding Author). Email: arm.fazelzad@iauctb.ac.ir
3. Department of German Language, CT.C., Islamic Azad University, Tehran, Iran. Email: par.panjehshahi1968@iau.ac.ir

1. Introduction

Foreign language acquisition, as extensively discussed in applied linguistics, requires sustained cognitive effort and effective memory consolidation (Ellis, 2015). In instructional contexts such as Iran, where pedagogical goals often emphasize immediate communicative performance, classroom practices frequently rely on extensive repetition and **massed (blocked) practice**, which are known to support short-term fluency gains (Bjork & Bjork, 1992; Dunlosky et al., 2013). Although these techniques can lead to rapid initial improvement and perform well on immediate assessments, **research** has repeatedly shown that knowledge acquired through **massed practice is more fragile and susceptible to accelerated forgetting when recall is delayed** (Cepeda et al., 2006; Roediger & Karpicke, 2006).

Desirable difficulties refer to learning conditions that deliberately introduce manageable challenges during the learning process in order to enhance long-term retention and transfer, even if they temporarily reduce short-term performance (Bjork & Bjork, 1992). The central premise of this framework is that learning activities requiring greater cognitive effort—such as retrieving information after a delay, discriminating between similar linguistic forms, or generating responses rather than recognizing them—strengthen underlying memory representations and improve future accessibility. From this perspective, momentary struggle or reduced fluency during practice is not a sign of ineffective instruction, but rather an indicator of deeper processing that supports durable and transferable learning outcomes (Bjork, 1994; Bjork & Bjork, 2011).

This research directly applies these concepts to the specific context of German as a Foreign Language (GFL) instruction for Persian-speaking beginners at the A1 level. A growing body of empirical research has demonstrated the effectiveness of desirable difficulties in second and foreign language learning, particularly in English-language contexts. For example, retrieval practice and spaced repetition have been shown to significantly enhance long-term vocabulary retention among EFL learners (Pan & Schmitt, 2023; Roediger & Karpicke, 2006), while interleaved practice has been found to improve grammatical discrimination and transfer across linguistic structures (Vaughn & Rawson, 2023). More recent studies have extended these findings to non-English contexts, indicating that effortful learning strategies can also support speaking accuracy and retention among Iranian EFL learners (Zhao & Li, 2025).

Despite these advances, empirical evidence examining the combined application of spacing, interleaving, and active retrieval within German as a Foreign Language (GFL) instruction, particularly among Persian-speaking beginners, remains scarce. The present study addressed

this gap by investigating whether structured desirable-difficulty-based instruction led to more durable retention of grammatical structures and vocabulary, and whether these gains could be transferred to complex productive skills such as speaking and writing within the Iranian educational context.

2. Review of the Related Literature

2.1. The New Theory of Disuse (NTOD)

The theoretical underpinning of this study rests primarily on Bjork and Bjork's (1992) New Theory of Disuse (NTOD). This model revolutionized the understanding of memory consolidation by proposing two distinct parameters governing memory traces:

Storage Strength (SS): This represents the enduring strength of the memory trace encoded in long-term memory. A high SS indicates that the information is deeply rooted and resistant to forgetting.

Retrieval Strength (RS): This denotes the immediate accessibility of the information at a given moment. High RS allows for quick, almost automatic recall.

Massed practice (cramming) is highly effective at rapidly boosting retrieval strength (RS), often creating a subjective sense of proficiency immediately after study. However, because it contributes little to the gradual accumulation of storage strength (SS), such gains tend to be short-lived and result in poor long-term retention (Bjork & Bjork, 1992; Dunlosky et al., 2013). Conversely, desirable difficulties are intentionally designed to slow the initial growth of RS; by requiring effortful searching and reconstruction of memory traces, these conditions promote stronger and more durable storage strength, thereby supporting long-term learning (Bjork, 1994; Bjork & Bjork, 2011).

2.2. Operationalizing Desirable Difficulties

The principle of desirable difficulties is operationalized through several specific laboratory and classroom techniques, four of which are central to this investigation:

Spacing (Distributed Practice): Instead of reviewing material in one long session, learning sessions are spread out over increasing intervals (Cepeda et al., 2006). This forces the memory system to work harder to retrieve moderately faded information.

Interleaving (Mixing Practice): Rather than practicing one skill or grammatical structure exhaustively before moving to the next (blocked practice), interleaving involves mixing tasks from

different, related concepts within a single study session. This forces the learner to continuously discriminate between concepts and select the appropriate retrieval strategy, enhancing pattern recognition (Borromeo-Ferri et al., 2021).

Retrieval Practice (Testing Effect): The act of successfully recalling information without cues (e.g., self-testing, flashcards, short quizzes) is significantly more beneficial for long-term retention than restudying the material (Roediger & Karpicke, 2006). This directly strengthens SS. Recent studies have expanded the framework of Desirable Difficulties to foreign language learning, showing that making learning effortful enhances long-term retention and transfer across linguistic domains (Bjork & deWinstanley, 2022; Pan & Schmitt, 2023; Vaughn & Rawson, 2023).

Generative Learning: This involves applying learned forms to create novel sentences, paragraphs, or solutions that were not explicitly taught, requiring the learner to synthesize existing knowledge flexibly (DeKeyser, 2007).

Studies in cognitive and educational fields have repeatedly shown that introducing challenging but beneficial obstacles can improve how well information is remembered and applied over time. For instance, Dunlosky et al. (2013) synthesized extensive experimental evidence showing that spacing and retrieval practice reliably outperform massed rehearsal across learning tasks. In the context of second and foreign language learning, empirical studies have reported similar benefits. For example, retrieval-based practice and spaced repetition have been shown to significantly improve long-term vocabulary retention (Pan & Schmitt, 2023; Roediger & Karpicke, 2006), while interleaved practice has been found to enhance grammatical discrimination and flexible application of linguistic forms (Vaughn & Rawson, 2023). Recent studies conducted in classrooms show that carefully planned teaching methods help students engage more deeply and better develop useful skills like speaking and writing (Tullis & Finley, 2024).

Despite this growing body of evidence, most empirical investigations have been conducted in laboratory settings or within English as a Foreign Language (EFL) contexts, often focusing on isolated techniques rather than their combined, systematic application. Consequently, there remains a shortage of localized, classroom-based studies examining how multiple desirable difficulty strategies operate together within German as a Foreign Language (GFL) instruction, particularly for Persian-speaking beginner learners in Iranian educational contexts.

To address this gap, the present study empirically examined the integrated implementation of spacing, interleaving, retrieval practice, and generative learning within a six-week GFL program at the A1 level. By combining quantitative measures of retention and

productive performance with qualitative learner perspectives, this study contributed context-specific evidence to the desirable difficulties literature and extended its applicability beyond dominant EFL settings. Accordingly, the study pursued the following research objectives: (a), whether instruction grounded in desirable difficulties led to greater long-term retention of grammatical structures and vocabulary compared to massed practice. (b), whether gains in retention could be transferred to improved productive skills, specifically speaking and writing. (c) learners' perceptions of effortful learning strategies in GFL instruction.

Based on these objectives, the study attempted to address the following research questions:

RQ1. Does the application of desirable difficulties result in significantly higher delayed post-test performance than conventional massed practice among Iranian GFL learners?

RQ2. Do learners exposed to desirable difficulties demonstrate superior performance in productive skills (speaking and writing)?

RQ3. How do learners perceive the cognitive challenge associated with desirable difficulty-based instruction?

3. Method

3.1. Design

This study adopted a mixed-methods design, integrating quantitative and qualitative data to comprehensively examine the effects of desirable difficulties on learning outcomes in German as a Foreign Language (GFL). A mixed-methods approach was selected because quantitative measures alone could capture differences in retention and productive performance but could not sufficiently explain learners' experiences of cognitive effort and perceived usefulness of the instructional approach.

The quantitative phase employed a quasi-experimental pretest–posttest control group design, comparing an experimental group instructed through desirable difficulties with a control group following conventional massed practice. Immediate and delayed post-tests were used to assess short-term performance, long-term retention, and transfer to productive skills. The qualitative phase consisted of semi-structured interviews conducted with a purposively selected subset of learners from the experimental group. These interviews explored learners' perceptions of effortful learning, cognitive challenge, and the perceived impact of desirable difficulty-based

instruction on understanding and motivation. Integrating qualitative data provided deeper interpretation of the quantitative findings and increased the explanatory power of the study.

3.2. Participants

The study population comprised twenty (N = 20) native Persian speakers enrolled in an intensive, beginner-level (A1) German course at a private language center in Tehran, Iran. Participants were selected through convenience sampling from intact classes and voluntarily agreed to take part in the study. Eligibility criteria required that learners had no prior formal instruction in German, which was confirmed through a screening questionnaire administered prior to the intervention. Participants were randomly assigned to either an experimental group (n = 10), receiving instruction based on desirable difficulties, or a control group (n = 10), following conventional massed-practice instruction.

All participants were adult learners (M = 22.5 years, SD = 1.8). To ensure baseline comparability between groups, participants' initial homogeneity was assessed using scores from a preliminary Persian language aptitude test and their stated motivations for learning German (e.g., university entrance, employment). Based on these measures, no systematic differences were observed between the two groups prior to the intervention. The sample consisted of 14 female and 6 male learners, with a comparable gender distribution across the experimental and control groups. Informed consent was obtained from all participants prior to data collection.

3.3. Instruments and Materials

Three types of instruments were employed in this study to measure learners' receptive knowledge, controlled production, and complex productive performance. All instruments were aligned with A1-level objectives as specified by the Common European Framework of Reference (CEFR).

Immediate Post-Test (IPT) and Delayed Post-Test (DPT) Learners' short-term achievement and long-term retention were assessed using two parallel, researcher-developed tests: an Immediate Post-Test (IPT) and a Delayed Post-Test (DPT). Both tests consisted of 100 items equally weighted across receptive knowledge and controlled productive use of grammatical and lexical structures covered during the instructional period. The receptive component included multiple-choice items and cloze tests targeting A1-level grammar and vocabulary. The controlled production component consisted of short sentence transformation tasks requiring learners to apply learned grammatical forms accurately. Each test required approximately 45 minutes to complete. All items were scored dichotomously (1 = correct, 0 = incorrect), resulting in a maximum

possible score of 100. The IPT was administered immediately following the six-week intervention, whereas the DPT was administered four weeks later without any intervening review or practice. Parallel versions of the tests were employed to minimize test–retest effects.

Productive Skills Assessment (PSA) Learners' complex productive performance was assessed through a Productive Skills Assessment (PSA), administered exclusively during the delayed post-test phase to measure the transfer of learning to spontaneous language use. The PSA consisted of two independent tasks: a speaking task and a writing task. The speaking task took the form of a semi-structured interview conducted individually with each participant.

The semi-structured interviews were also used as the third tool for research, giving qualitative information about how learners felt about the level of mental effort, the difficulty, and the teaching method. The interview was guided by a fixed set of prompts focusing on familiar A1-level topics (e.g., self-introduction, daily routines, personal preferences). While the prompts were predetermined, follow-up questions were used to elicit spontaneous responses and encourage extended production. Each interview lasted approximately 8–10 minutes and was audio-recorded for scoring purposes. The writing task required participants to produce a short descriptive essay of approximately 150 words on a familiar topic aligned with the instructional content (e.g., describing their daily routine or living environment). Participants were given 25 minutes to complete the task without access to reference materials. Before the interviews started, the participants were told why the interviews were happening, promised that their information would stay private and not be shared, and were made clear that they could choose not to take part. They signed a document agreeing to take part, and all the audio recordings were only used for research.

Scoring procedures and inter-rater reliability performance in both the speaking and writing tasks was evaluated using a holistic analytic rubric focusing on two dimensions: accuracy (grammatical correctness and appropriate use of lexical items) and effectiveness (clarity of expression and successful communication of meaning). Each dimension was scored on a 10-point scale, yielding a maximum combined score of 20 for each task. All PSA performances were independently rated by two trained native German speakers who were blinded to participants' group assignment. Prior to scoring, the raters were familiarized with the rubric and jointly rated a subset of samples to establish scoring consistency. Inter-rater reliability was calculated using Pearson correlation coefficients, yielding satisfactory agreement for both speaking ($r = 0.87$) and writing ($r = 0.84$). Discrepancies in scoring were resolved through discussion.

3.4. Procedure

A six-week intervention was implemented. Instruction for both groups was delivered by the same instructor to control for potential teacher effects. The instructor followed identical curricular content and instructional objectives for both groups, with differences limited to the practice conditions specified by the experimental design. Instruction took place over 18 instructional sessions, with three sessions per week. Each session lasted approximately 90 minutes, resulting in a total of 27 hours of classroom instruction for each group. The instructional syllabus for both groups was identical in terms of content coverage, instructional objectives, and total exposure time.

The curriculum targeted core A1-level competencies as specified by the Common European Framework of Reference (CEFR), with a particular focus on foundational grammar (e.g., basic sentence structure, modal verbs, case marking), essential vocabulary sets, and controlled communicative use of these forms. Listening and reading activities were integrated as supportive input skills, while speaking and writing were emphasized primarily as outcome measures. Given the study's primary focus on examining the effects of desirable difficulties on long-term retention and transfer to productive performance, the six-week instructional period was deemed sufficient for introducing and practicing targeted A1 structures while maintaining experimental control. The study did not aim to provide full balanced development of all language skills, but rather to investigate how differing practice schedules influenced the durability and functional use of newly learned grammatical and lexical knowledge.

All participants completed a German A1 pre-test aligned with CEFR descriptors, consisting of 60 multiple-choice and cloze items assessing basic grammatical structures and essential vocabulary. The test was administered solely to ensure baseline homogeneity across groups and was not included in the main statistical analyses. Following the pre-test, the 20 selected participants were randomly assigned to experimental and control groups, with ten students in each group.

Experimental Group (DD; n=10): This group engaged in using desirable difficulties as detailed below:

- **Spacing:** Previously taught grammatical structures were systematically reviewed after delayed intervals rather than immediately following initial instruction. For example, grammatical structures introduced in Week 1 (e.g., nominative and accusative case marking) were revisited only in Weeks 3 and 5 through brief review tasks and practice exercises. The

teacher did not provide advance reminders or summaries before these reviews; instead, learners were required to retrieve prior knowledge independently. The teacher's role was limited to providing feedback after task completion.

- **Interleaving:** Practice activities were designed to mix multiple grammatical structures within a single session. For instance, learners completed worksheets in which sentences requiring Präteritum verb forms, separable verbs, and adjective declensions appeared in random order rather than in separate blocks. Students first identified which grammatical rule applied before producing the correct form. The teacher monitored performance and provided delayed corrective feedback, encouraging learners to explain their rule selection when errors occurred.
- **Retrieval Practice:** Each instructional week included mandatory retrieval-based activities at the beginning of selected sessions. Learners were asked to perform short "brain dump" tasks, in which they wrote everything they could recall about previously studied grammar rules and vocabulary without access to notes or textbooks. These were followed by brief, ungraded quizzes consisting of short-answer or sentence-completion items. The teacher did not correct errors immediately, using the tasks solely to prompt effortful recall rather than formal evaluation.
- **Generative Learning:** Learners regularly engaged in productive tasks that required generating novel language beyond rote repetition. For example, students wrote short dialogues or descriptive paragraphs (e.g., describing their apartment or daily routine) that explicitly required the use of grammatical structures taught in different weeks, such as locative prepositions, adjective endings, and separable verbs. Students worked individually or in pairs, while the teacher acted as a facilitator, providing prompts and post-task feedback without modeling complete responses in advance.

Control Group (CP; n = 10): The control group followed a conventional blocked-practice instructional approach. Each session began with the explicit presentation of a new grammatical structure by the teacher, including rule explanation and model sentences written on the board. This was followed by extensive guided practice focusing on one grammatical feature at a time. Practice activities included repetitive sentence construction, fill-in-the-blank exercises, verb conjugation tables, and pattern-completion worksheets targeting a single structure (e.g., only Präteritum forms or only separable verbs) within each session. Students practiced the same form repeatedly until a high level of immediate accuracy was achieved. The teacher played an active, directive role by providing frequent explanations, modeling correct responses, and supplying immediate corrective feedback after each student response. Errors were corrected instantly to

prevent persistence of incorrect forms, and correct answers were often provided when students hesitated. Review sessions were typically massed at the end of each instructional unit. During these sessions, previously taught material was revisited through summary explanations, repetition drills, and short practice exercises designed to reinforce recently learned forms. Students' primary role was to apply explicitly taught rules accurately, with an emphasis on fluency and correctness during practice, rather than on effortful retrieval or rule selection.

Following the intervention, the following post-tests were administered:

1. Immediate Post-Test (IPT) which was administered immediately following the six-week intervention to measure short-term proficiency achieved by both groups.

2. Delayed Post-Test (DPT): which was administered four weeks after the instruction concluded (i.e., without any specific study or review time allocated during this four-week gap) to measure long-term retention. Both post-tests employed parallel versions of the same German A1 proficiency test. The two tests were matched in format, content coverage, and level of difficulty, and included identical task types as described in Section 3.3 (i.e., multiple-choice items, cloze tests, and controlled sentence transformation tasks). The same scoring procedures were applied to both tests, and all post-tests were administered under standardized classroom conditions without access to instructional materials. Testing conditions were identical for both groups. The four-week delay was selected to reduce retrieval strength while allowing storage strength to be meaningfully assessed, in line with the predictions of the New Theory of Disuse. Finally, six representative participants from the experimental group were invited to voluntary participate in the semi-structured interviews to elicit their reflections on the activities they were engaged in during the intervention.

3.5. Data Analysis

The pre-test scores were used solely to establish baseline equivalence between the experimental and control groups and were not included in the main statistical analyses. Quantitative data from the Immediate Post-Test (IPT) and Delayed Post-Test (DPT) were analyzed using independent-samples t-tests to compare group differences in short-term learning outcomes and long-term retention. Retention loss was calculated by comparing IPT and DPT scores within each group.

For the Productive Skills Assessment (PSA), total scores for speaking and writing were computed by summing the accuracy and effectiveness ratings assigned by two independent raters. Independent-samples t-tests were conducted to examine group differences in productive

performance. Inter-rater reliability coefficients were calculated prior to analysis and indicated acceptable agreement levels. Significance was set at $p < .05$ for all statistical tests.

The data collected from the interviews were audio recorded, transcribed verbatim, and then analyzed using thematic analysis. This method, based on a six-step process developed by Braun and Clarke in 2006, helped develop common themes and patterns about how learners viewed mental effort, difficulties in learning, and the benefits they felt from what they had learnt.

4. Results

This section reports the quantitative and qualitative findings of the study, comparing the Desirable Difficulties (DD) group and the Control Practice (CP) group in terms of long-term retention and productive skill development.

4.1. Quantitative Findings

Long-Term Retention Performance (IPT vs. DPT)

After confirming normality and homogeneity of variances, an independent-samples t-test was conducted. Initial Post-Test (IPT) results indicated minimal differences between the two groups, suggesting comparable immediate learning outcomes following instruction. An independent-samples *t*-test confirmed that the difference was not statistically significant ($p > .05$), supporting the assumption that desirable difficulties do not accelerate short-term acquisition.

Table 1.

Mean Scores and Retention Rates for IPT and DPT by Group

Group	IPT Mean (SD)	DPT Mean (SD)	Retention Rate (%)	Retention Drop (%)
Control Practice (CP)	88.56 (4.23)	68.23 (5.11)	68.9	31.1
Desirable Difficulties (DD)	88.17 (4.08)	82.75 (4.67)	75.9	6.7

Note. IPT = Immediate Post-Test; DPT = Delayed Post-Test. Scores are reported out of 100. Retention rate reflects the proportion of IPT performance maintained at DPT.

As shown in Table 1, learners in the DD group demonstrated significantly higher retention scores after a four-week interval compared to the CP group. While the DD group retained approximately 75.9% of their previously acquired knowledge, the CP group exhibited a markedly larger decline, retaining only about two-thirds of the learned material. These results give evidence for the New Theory of Disuse, showing that actively recalling information increases how well it's stored and helps keep learning results from fading over time.

An independent-samples *t*-test indicated that this difference was statistically significant, $t(18) = 4.15$, $p < .001$, with a large effect size (Cohen's $d = 1.89$). The descriptive and inferential statistics for the delayed post-test are shown in Table 2.

Table 2.

Independent-Samples t-Test Results for Delayed Post-Test Performance

Group	Mean	SD	*t*(18)	*p*	Cohen's *d*
Control Practice (CP)	68.23	5.11	4.15	< .001	1.89
Desirable Difficulties (DD)	82.75	4.67			

Note. DPT = Delayed Post-Test. Values represent group means and standard deviations. Effect size is reported as Cohen's d.

Productive Skills Assessment (PSA)

Productive language skills were assessed through speaking and writing tasks designed to measure both accuracy and communicative effectiveness. Descriptive statistics for both groups are presented in Table 3.

Table 3.

Productive Skills Assessment (PSA) Scores by Skill and Group

Skill	Group	Accuracy	Effectiveness	Total
Speaking	CP	5.1	6.3	11.4
Speaking	DD	6.8	7.5	14.3
Writing	CP	4.8	5.1	10.7
Writing	DD	6.8	7.1	13.9

Note. Maximum possible score per skill = 20. PSA scores represent combined rater judgments of grammatical accuracy and communicative effectiveness.

As shown in Table 3, The DD group performed better than the CP group in both speaking and writing tasks, scoring higher in total PSA across all assessment areas. In both speaking and writing, the DD group achieved higher total PSA scores compared to the CP group.

Following checks for normality and homogeneity of variances, the independent-samples t-test was run. The descriptive and inferential statistics for the Productive Skills Assessment (PSA) are presented in Table 4.

Table 4.

Independent-Samples t-Test Results for Productive Skills Assessment (PSA)

Measure	Group	Mean	SD	*t*(18)	*p*	Cohen's *d*
PSA (Total)	Control Practice (CP)	11.05	1.1			
	Desirable Difficulties (DD)	14.1	1.25	3.02	< .01	1.35

Note. PSA = Productive Skills Assessment. Scores represent aggregated speaking and writing performance (combined accuracy and communicative effectiveness). Maximum possible score = 20.

Rater comments further indicated that learners in the DD group demonstrated greater syntactic flexibility and more confident use of verb tense and case marking, particularly in tasks requiring spontaneous production. Although minor grammatical inaccuracies persisted, these did not substantially impede communicative effectiveness.

4.2. Qualitative Findings

To further explore learners' instructional experiences, semi-structured interviews were conducted with a subset of participants from the DD group (n = 6). The interviews were audio-recorded, transcribed verbatim, and analyzed using thematic analysis following Braun and Clarke's (2006) six-step framework. The analysis yielded three recurrent themes, summarized below.

Theme 1: Initial Cognitive Difficulty

Participants consistently reported that early instructional stages felt demanding and occasionally confusing, particularly during interleaved retrieval tasks. For example, participant 4 noted that; "At first, recalling the dative case verbs without looking felt stressful and confusing".

Theme 2: Delayed Perceived Benefit

Despite early difficulty, learners later recognized the long-term benefits of effortful practice. As participant 6 expressed: "Weeks later, when the test came, I remembered it naturally, without forcing myself".

Theme 3: Enhanced Metacognitive Awareness

Several learners described a shift in how they perceived learning, viewing struggle as an integral part of progress rather than a sign of failure. For instance, participant 2 expressed that: "I used to think that if learning was hard, it meant I was doing something wrong. But then I realized that facing challenges actually helped me learn more and remember things better".

These findings align with mastery-oriented learning perspectives (Dweck, 1986), in which

sustained effort and productive struggle contribute to deeper learning and motivation.

5. Discussion

The present study investigated the effects of integrating desirable difficulties into German as a Foreign Language (GFL) instruction for beginner Iranian learners. The results of both quantitative and qualitative data supported the idea that instruction methods which require active recall can be very effective, especially when it comes to remembering information over time and using it in real-life language situations.

First, the quantitative results revealed a dissociation between short-term performance and long-term retention. While immediate post-test outcomes showed no statistically significant differences between groups, delayed post-test scores favored the desirable difficulties (DD) group, indicating a substantially lower rate of knowledge decay (6.7% vs. 31.1%). This pattern aligns with core assumptions of the New Theory of Disuse (Bjork & Bjork, 1992), indicating that retrieval effort may play an important role in strengthening long-term memory representations, even when short-term retrieval strength appears comparable across instructional conditions. Within this framework, learning activities that require effortful retrieval after a delay are expected to promote more durable storage strength, resulting in greater resistance to forgetting over time.

Second, the higher performance of the DD group on the Productive Skills Assessment (PSA) suggests that desirable difficulties may contribute not only to improved retention of linguistic forms, but also to enhanced transfer to complex communicative tasks. Learners exposed to spacing, interleaving, and retrieval-based practice exhibited greater syntactic flexibility and communicative effectiveness in both speaking and writing tasks. These findings are consistent with prior research indicating that retrieval-based practice supports flexible language use rather than surface-level fluency driven by repeated exposure (Tullis & Finley, 2024; Van Merriënboer & Sweller, 2010). In contrast, the control group's reliance on blocked practice may have fostered temporary fluency driven by immediate accessibility, which deteriorated once retrieval pathways weakened.

The qualitative findings further helped illuminate the mechanisms underlying these quantitative outcomes. Participants initially perceived effort-inducing instructional activities as demanding or confusing; however, this early cognitive challenge came to be perceived as beneficial when learners observed improved recall and greater confidence in delayed assessments. Such reflections are the characteristic of mastery-oriented learning, in which

learners view struggle as an integral component of learning rather than a signal of failure (Dweck, 1986). The alignment between learners' experiences and objective performance gains contributes to the explanatory coherence of the findings.

From a pedagogical perspective, these results suggest that instructors may need to be prepared to tolerate slower initial progress when implementing desirable difficulties in GFL instruction. Rather than treating retrieval-based and interleaved activities as supplementary or optional, these strategies may be more effective when systematically embedded into instructional design to promote durable learning and transferable communicative competence.

6. Conclusion

This study investigated the effects of incorporating desirable difficulties, specifically spacing, interleaving, and retrieval practice, on long-term retention and the functional application of language knowledge among GFL learners. Quantitative findings demonstrated that while immediate post-test performance did not differ significantly between groups, learners exposed to desirable difficulties exhibited substantially higher delayed retention and significantly stronger productive skills in both speaking and writing tasks. Complementary qualitative data further indicated that learners initially perceived these instructional conditions as more demanding, yet later recognized their contribution to deeper understanding, increased self-monitoring, and more durable learning outcomes.

From a pedagogical perspective, these findings underscore the importance of prioritizing long-term learning gains over short-term performance. Instructional practices that generate immediate fluency or accuracy may create an illusion of effectiveness, whereas effort-inducing learning conditions appear to foster more robust and transferable language competence. For GFL programs, the systematic integration of spacing, interleaving, and retrieval-based activities may therefore serve as a viable means of enhancing both retention and productive language use, particularly when instructional goals emphasize sustainability and communicative functionality.

The study shows there's a big difference between what works in the short term and what leads to real, long-lasting learning. While practicing a lot quickly might seem better in the short run, it does not help build lasting language skills. For programs that aim to develop true fluency, using learning methods that require effort is not just helpful, it is essential. This idea matches recent research on flipped learning that involves working together or competing (Marashi & Mokhlesi, 2025). The study also corroborates the idea of using teaching methods that create mild

challenges, which help build strong and lasting language abilities. By treating these short-term difficulties as a normal and useful part of learning and not as a sign that something is wrong, teachers can create learning environments that lead to deeper understanding, better memory, and more reliable language skills.

Several limitations of the present study should be acknowledged. First, the sample size was relatively small and limited to A1-level learners within a specific instructional context, which may restrict the generalizability of the findings. Second, productive skills were assessed within a controlled classroom setting, potentially limiting the extent to which results reflect real-world communicative performance. Finally, learner perceptions were explored through interviews with a subset of participants, which, while informative, may not fully capture the range of experiences across the entire cohort.

References

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *Psychological Review, 99*(1), 35–45. https://doi.org/10.1037/0033-295X.99.1.35

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition, 9*(4), 475–479. https://doi.org/10.1016/j.jarmac.2020.08.005

Bjork, R. A., & deWinstanley, P. A. (2022). Why making learning effortful improves long-term retention: Advances in the desirable difficulties framework. *Journal of Cognitive Psychology, 34*(2), 145–162. https://doi.org/10.1080/20445911.2022.2087741

Borromeo Ferri, R., Pede, S., & Lipowsky, F. (2021). Nested learning in procedural and conceptual tasks. *Journal für Mathematik-Didaktik, 42*(1), 1–23. https://doi.org/10.1007/s13138-021-00201-2

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

DeKeyser, R. M. (2007). *Skill acquisition theory*. Routledge.

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

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*(10), 1040–1048.

Ellis, R. (2015). Cognitive approaches to second language acquisition. *Annual Review of Applied Linguistics, 35*, 145–165. https://doi.org/10.1017/S0267190514000245

Marashi, H., & Mokhlesi, N. (2025). Comparing the impact of cooperative and competitive flipped learning on EFL learners' speaking performance. *Curriculum Research Journal, 6*(3), 1–18. https://doi.org/10.71703/cure.2025.1211195

Pan, S. C., & Schmitt, A. (2023). Optimizing retrieval practice and spacing for foreign language vocabulary learning. *Applied Psycholinguistics, 44*(3), 601–623. https://doi.org/10.1017/S0142716423000190

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Tullis, J. G., & Finley, J. R. (2024). Cognitive effort and engagement in multimedia language learning: Testing the desirable difficulties hypothesis. *Computer Assisted Language Learning, 37*(5), 1085–1104. https://doi.org/10.1080/09588221.2024.2365498

Vaughn, K. E., & Rawson, K. A. (2023). A framework for applying desirable difficulties to second language instruction. *Language Teaching Research, 29*(4), 587–605. https://doi.org/10.1177/13621688231101597

Zhao, Q., & Li, H. (2025). Retrieval-based difficulty and speaking accuracy among EFL learners: Evidence from Iranian contexts. *Curriculum Research Journal, 6*(1), 1–15. https://doi.org/10.71703/cure.2025.1211189