RESEARCH ARTICLE

BON VIEW PUBLISHING

Portfolio Assessment as Reflective Practice: Documenting Metacognitive Growth in Language Teacher Education

Patrícia Ferreira[1],*

[1] Foreign Languages Department, Higher School of Education of the Polytechnic Institute of Porto, Portugal and Center for Research and Innovation in Education, Portugal

Abstract: Traditional assessment models in higher education often prioritize content recall, failing to measure the complex, adaptive competencies, such as metacognition and cultural competence, essential for future educators and global citizens. This study documents the implementation of process-oriented portfolio assessment within a Language and Cultures teacher education program in Portugal. A two-year sequential explanatory mixed-methods study (2023/2024 to 2024/2025) involved one undergraduate cohort ($N = 81$) in English Culture and two graduate cohorts ($N = 40$) in Citizenship and Intercultural Education, totaling approximately 121 students. Descriptive quantitative analysis documented consistent performance patterns across cohorts when using structured, process-oriented rubrics. Qualitative analysis, focusing on reflective writing and critical thinking criteria, demonstrated significant metacognitive development and self-regulated learning, substantiated by student reflections and instructor observations. The findings affirm that portfolios can serve as an effective mechanism for fostering and documenting the process of learning rather than merely measuring its outcomes. The paper argues that integrating reflective portfolio assessment into teacher training curricula offers a pedagogically valuable alternative to traditional grading paradigms in higher education.

Keywords: portfolio assessment, reflective practice, metacognition, self-regulated learning, teacher education

1. Introduction

The modern higher education context is defined by a fundamental challenge: assessing not only what students know but what they can do with what they know. Traditional assessment models [1–3], particularly timed, high-stakes summative examinations, have been consistently criticized for rewarding superficial knowledge and information recall while failing to capture complex, adaptive skills such as critical reflection, metacognition, and intercultural competence. These latter skills are unequivocally central to 21st-century academic rigor and, critically, to the professional preparation of future teachers. This study focuses on validating the portfolio as the mechanism for fostering and measuring these 21st-century competencies, often encapsulated by the "4 C's": Critical Thinking, Communication, Collaboration, and Creativity [4].

In response, authentic assessment practices [5, 6] have gained prominence, emphasizing real-world application, sustained engagement, and documented development. Portfolios, specifically, have emerged as a powerful tool grounded in constructivist and experiential learning theories. A developmental portfolio, defined as a systematic, purposeful collection of student work that documents effort, progress, and achievement over time, offers a holistic and integrative view of the learning journey [7, 8].

1.1. Clarification of key constructs

While contemporary educational discourse often invokes the broad framework of "21st-century competencies" or the "4 C's" (Critical Thinking, Communication, Collaboration, and Creativity) as aspirational goals for higher education [4], this study focuses specifically on two interrelated competencies that are both essential for language teacher education and operationalizable through portfolio assessment: metacognition and cultural competence.

Metacognition, defined as awareness and regulation of one's own cognitive processes [9], encompasses three core components measured in this study: (1) planning (setting learning goals and strategizing approaches), (2) monitoring (tracking one's progress and identifying gaps), and (3) evaluating (reflecting critically on outcomes and determining needed adjustments). These components directly align with self-regulated learning (SRL) theory and are operationalized in the portfolio through mandatory reflective writing criteria that require students to document and analyze their learning trajectories rather than simply presenting final products.

Cultural competence, in the context of language and intercultural education, refers to the capacity to critically engage with

---
*Corresponding author: Foreign Languages Department, Higher School of Education of the Polytechnic Institute of Porto, Portugal and Center for Research and Innovation in Education, Portugal. Email: patriciaferreira@ese.ipp.pt

cultural contexts, recognize and interrogate one's own cultural assumptions, and apply intercultural frameworks to pedagogical practice [5]. For the graduate cohorts preparing to become primary school English teachers, cultural competence extends beyond theoretical knowledge to include the demonstrated ability to design culturally responsive pedagogical interventions operationalized through the portfolio's "Creativity and Pedagogical Application" criterion, which required students to create lesson plans and teaching materials that explicitly addressed intercultural learning objectives.

It is crucial to clarify what this study measures and what it does not. The portfolio rubrics employed in this research explicitly assess and weight metacognitive development (through reflective writing criteria) and cultural competence (through pedagogical application criteria). While the broader "4 C's" framework provides useful contextual grounding for understanding 21st-century educational goals, this study does not attempt to measure all four dimensions equally or independently. Communication and collaboration, while inevitably present in the learning process, were not isolated as distinct assessment criteria. Instead, this research maintains analytical focus on the two constructs most directly fostered by reflective portfolio practice and most critical for teacher preparation: the capacity for metacognitive self-regulation and the ability to apply cultural knowledge pedagogically.

The quantitative component of this study (Section 3.3) documents performance patterns based on rubric criteria that operationalize these two constructs, while the qualitative component (also Section 3.3) explores how the portfolio process specifically fosters metacognitive awareness and cultural competence through student reflections and documented developmental trajectories. This focused approach allows for deeper, more rigorous analysis than would be possible with a diffuse attempt to measure all possible 21st-century competencies simultaneously.

2. Literature Review

The pedagogical strength of the portfolio [10, 11] is rooted in its capacity as an instrument of self-formation and reflection, its power to foster metacognition, or "thinking about one's thinking" and drive SRL [12]: planning, monitoring, and evaluating performance [9]. This process transforms the assignment from a mere compilation of artifacts into an engine for deeper cognitive engagement, directly impacting the development of higher-order skills.

2.1. Reflective practice and professional development: theoretical foundations

The portfolio's mandatory reflective component is fundamentally aligned with theories of professional development [13–15]. Consequently, the required reflective essays transform practical experience into an "educational experience" by demanding students to analyze and rationally reconstruct their actions. Similarly, Donald Schön's central theory, presented in *The Reflective Practitioner*, emphasizes the importance of the Reflection-on-Action process, the act of reviewing and analyzing practice outside the immediate classroom context [16]. This is the crucial moment where practical experience is intentionally linked with theoretical frameworks, fundamentally grounding and transforming the professional identity of the learner.

Furthermore, authors like Idália Sá-Chaves [17] highlight that the portfolio enables the learner to speak for themselves, assign unique meaning to experiences, and interpret them personally.

The portfolio thus becomes a personal and professional instrument that provides tangible evidence of the reflective process, demonstrating the learner's capacity for self-evaluation and self-regulation.

In recent years, scholarship has expanded this understanding of portfolios by highlighting their role not only in professional development but also in fostering a sense of belonging in higher education. Because portfolio work requires students to articulate their experiences, values, and identities, it naturally creates a space where learners can situate themselves within their academic and professional communities. Boyd and Feeting [18] argue that the e-portfolio can act as a "*locus for the convergence of the foundations of belonging,*" offering students a structured yet personal environment in which reflective practice becomes intertwined with identity affirmation, relational connection, and academic integration. In this sense, the reflective portfolio does more than document learning: it supports students in recognizing themselves as legitimate participants within their disciplinary and institutional contexts, strengthening both engagement and wellbeing.

This reflective mandate also takes on a formalized, investigative character, particularly relevant in teacher training programs [19, 20]. The reflection embedded in the portfolio promotes an investigative attitude. Learners are encouraged to act as researchers of their own practice, identifying problems, testing hypotheses (interventions), and analyzing the results, which fosters an attitude of continuous professional deepening and improvement. Correspondingly, Villas Boas [21] underscores the portfolio's relevance in documenting the student's process, or the journey they undertake. As an instrument of formative assessment, it enables the continuous monitoring of the learner's progress, actively encouraging the development of autonomy and criticality throughout the course. She argued that the portfolio is a robust alternative to traditional tests, enabling a more holistic and contextualized evaluation of student skills and knowledge. This aligns with Paul Black and Dylan Wiliam's [22] seminal work on formative assessment, *Inside the Black Box*, which highlights the importance of assessment methods that promote continuous learning and provide detailed feedback.

Other examples of recent scholarship further deepen this understanding. In their extensive reviews, Yang and Wong [8] and Walland and Shaw [23] illuminate the complexities surrounding the implementation of e-portfolios in higher education. Their work not only maps the technological, pedagogical, and ethical barriers that institutions and learners encounter but also offers thoughtful strategies for navigating these challenges. Their analyses invite us to consider e-portfolio adoption as an evolving, context-sensitive process, one that requires continual reflection, adaptation, and institutional commitment.

While the theoretical case for portfolio assessment is well-established through foundational works on reflective practice [16, 24, 25] and professional development [17], recent empirical research provides compelling quantitative and qualitative evidence of portfolio effectiveness in fostering the competencies essential for teacher preparation.

2.2. Recent empirical evidence (2020–2025)

This section synthesizes recent empirical studies (2020–2025) examining portfolio assessment's impact on metacognitive development, documents implementation realities across diverse educational contexts, and identifies critical research gaps that the current study addresses.

Recent meta-analytical research also provides robust empirical support for the portfolio's capacity to foster SRL and metacognitive awareness [26]. Guntur and Purnomo's [27] comprehensive meta-analysis examined SRL interventions across 15 empirical studies encompassing 5519 participants in online and blended learning environments. Their findings revealed that interventions incorporating structured reflection, the cornerstone of portfolio assessment, yielded moderate-to-strong effect sizes (d = 0.66) on learning outcomes. Critically, 53.3% of the most successful interventions employed mixed strategies combining cognitive task engagement with metacognitive reflection, precisely the dual structure inherent in portfolio work, where students both create artifacts (cognitive) and reflect upon their creation process (metacognitive). This meta-analytical evidence substantiates the theoretical claim that portfolios function as engines for deeper cognitive engagement by requiring students to document and analyze their own learning trajectories.

Quasi-experimental studies further corroborate these meta-analytical findings. Xu et al. [28], in a controlled study with 68 college students, demonstrated that explicit metacognitive support significantly enhanced students' SRL capacities and overall learning experiences. Students receiving structured metacognitive prompts, analogous to portfolio reflection requirements, showed marked improvements in monitoring and controlling their learning processes compared to control groups. Similarly, Arianto and Hanif's [29] factorial design study with 100 primary school students confirmed that both metacognitive strategies and SRL approaches positively influence problem-solving skills and self-efficacy, with synergistic effects when implemented in combination. These findings are particularly relevant for teacher education, as future educators must develop both metacognitive awareness and self-efficacy to model these competencies for their own students.

The Education Endowment Foundation's comprehensive evidence review [30] synthesizes decades of research on metacognitive and self-regulatory instruction, providing authoritative support for explicit teaching of these skills. Their analysis across multiple age groups and subject areas confirmed that metacognitive skills do not develop spontaneously but require intentional pedagogical structures, precisely what portfolio assessment's mandatory reflective components provide. The review emphasizes that effective metacognitive interventions share key characteristics: they make thinking processes explicit, provide structured opportunities for reflection, and offer timely feedback on both task performance and the learning process itself. This evidence base directly validates the portfolio's three-stage scaffolding protocol (initial submission, mid-term reflection, final synthesis) as aligned with empirically supported best practices in metacognitive instruction.

Empirical research documenting portfolio implementation in teacher education contexts reveals both significant developmental benefits and persistent practical challenges, providing a nuanced picture essential for realistic adoption planning.

Multiple recent studies confirm the portfolio assessment's positive impact on dimensions critical for teacher preparation. A large-scale study in Ghana involving 300 pre-service teachers demonstrated that the majority perceive e-portfolios as beneficial for their professional training, particularly in developing reflective capacity and creating evidence of teaching competence [31]. Notably, this study revealed discipline-specific variations, with pre-service teachers showing more positive perceptions than nursing students, suggesting that portfolio assessment may be particularly well-suited to professions requiring ongoing reflective practice.

Research on affective dimensions further illuminates portfolio benefits. A 2024 study published in *Language Testing in Asia* examined portfolio assessment in technology-enhanced environments, revealing significant positive impacts on academic emotion regulation, mindfulness, and attitudes toward learning. The study demonstrated that portfolio work, particularly when accompanied by constructive formative feedback, enhances learners' confidence while reducing anxiety, critical outcomes for teacher candidates who must navigate the emotional demands of classroom practice. Similarly, a quasi-experimental study with 58 Saudi Arabian EFL learners (2024) documented that portfolio-based assessment significantly enhanced grit tendencies, intrinsic motivation, and willingness to communicate compared to traditional testing, with medium-to-large effect sizes across all measures. These affective and motivational outcomes are particularly valuable in teacher education, where persistence, self-discipline, and communication skills directly predict professional effectiveness.

Comparative research provides empirical evidence of portfolio advantages over traditional assessment. Zaabalawi and Zaabalawi's [32] study with 60 university engineering students examined student perceptions of portfolio assessment versus exams across academic, mental health, and professional dimensions. Students overwhelmingly preferred portfolios, citing reduced test anxiety, opportunities for authentic skill demonstration, and better alignment with real-world professional demands. While this study focused on engineering education, its findings resonate with teacher education contexts where authentic demonstration of pedagogical skills holds greater validity than decontextualized examination performance.

Despite documented benefits, empirical research consistently identifies barriers that must be addressed for successful implementation. Yıldırım and Batdi's [33] multi-method study, while confirming portfolio effectiveness in enhancing academic achievement and 21st-century skills in primary education, documented critical obstacles including lesson duration constraints, technology access inequities, and the need for extensive advance planning and resource allocation. Their research emphasized that portfolio implementation cannot be grafted onto existing curricula as a simple substitution; it requires fundamental redesign of instructional time and support structures.

The workload issue emerges as the most significant barrier across multiple studies. Teacher education studies [34] found that e-portfolio adoption remains "quite uneven" across institutions, with most at early implementation stages despite widespread acknowledgment of pedagogical benefits. Their analysis identified faculty workload concerns as the primary impediment to scaling portfolio assessment, noting that institutions have failed to provide adequate recognition or compensation for the substantially increased assessment labor required. This finding is echoed in the Ghana study, which documented that the time-consuming nature of portfolio maintenance and evaluation creates practical sustainability concerns, particularly in under-resourced contexts.

Technical and evaluative challenges compound workload concerns. Multiple studies identify the absence of standardized evaluation criteria as a significant barrier [31, 34], noting that faculty often lack training in developing and applying analytical rubrics for complex, process-oriented work. Additionally, unequal technology access creates equity concerns, particularly in contexts where students' digital literacy and resource availability vary substantially. These implementation realities underscore that portfolio assessment's pedagogical promise can only be realized when accompanied by systemic institutional support, including faculty development, technological infrastructure, and workload recognition.

Recent systematic reviews reveal critical gaps in portfolio assessment research that this study directly addresses. A 2025 systematic scoping review analyzing 97 empirical articles (2014–2024) on assessment practices in initial teacher education identified concerning patterns: considerable methodological variability, an almost complete absence of comparative studies, and limited attention to how assessment practices function as embedded elements within program design rather than as isolated tools. The review emphasized that while individual innovative assessment instruments are widely studied, research capturing the in situ processes of implementation, the messiness of authentic classroom adoption, the scaffolding adjustments required across iterations, and the interplay between formative and summative functions remains critically under-researched.

This study contributes to addressing three specific gaps identified in recent literature:

Gap 1: Comparative Implementation Data. While numerous studies document portfolio use in single contexts or cohorts, comparative research examining consistency across different academic levels and student populations remains scarce. This study provides rare comparative data documenting portfolio implementation across undergraduate and graduate levels, across two distinct courses with different learning objectives, and across two iterations of the same course with refined criteria. This comparative approach allows for examination of how portfolio assessment functions across varying contextual demands, addressing the systematic review's call for research on assessment adaptability.

Gap 2: Assessment as Embedded Process. Most portfolio research treats assessment as an isolated intervention, examining outcomes without documenting the ongoing scaffolding, student struggles, and instructor adjustments that characterize authentic implementation. This study employs a naturalistic, longitudinal design spanning two academic years, with portfolios serving as the primary course assessment rather than an experimental add-on. By documenting both quantitative performance patterns and qualitative developmental processes, including student resistance and the "feedback paradox," the study provides the kind of implementation realism that systematic reviews identify as missing from current literature.

Gap 3: Integration of Process and Outcome Data. The 2025 systematic review noted that assessment research tends to focus either on psychometric properties (reliability, validity) or on learning processes (reflection, development) but rarely integrates both perspectives. This study's mixed-methods approach intentionally bridges this divide: descriptive quantitative data documents grading consistency while acknowledging the limitations of cross-level comparisons, while rich qualitative data captures the developmental processes underlying observed performance patterns. This integrated approach offers a more complete picture of what portfolio assessment accomplishes and requires.

Boyd and Feeting's [18] recent conceptual work on e-portfolios as "loci for the convergence of the foundations of belonging" suggests an additional dimension worthy of empirical investigation: portfolio work's role in identity formation and institutional integration. While not a primary focus of this study, the qualitative data revealing students' transformation from "students of teaching" to "teachers-as-professionals" through portfolio work resonates with Boyd and Feeting's theoretical framework, suggesting promising directions for future research on portfolio assessment's role in professional identity development within teacher education.

By providing detailed documentation of portfolio implementation across multiple cohorts and academic levels, this study contributes empirical evidence to support practitioners and administrators considering portfolio adoption while acknowledging the persistent challenges that require institutional commitment rather than individual faculty heroics to overcome.

3. Research Methodology

This section outlines the methodological framework employed to investigate portfolio assessment implementation and its impacts on metacognitive development in language teacher education. The study adopts a sequential explanatory mixed-methods design, integrating quantitative performance data with qualitative evidence of developmental processes. We begin by describing the overall research design and theoretical positioning, followed by a detailed specification of participants, instruments, and analytical procedures.

3.1. Research design

This study documents the implementation and outcomes of portfolio assessment in interdisciplinary Language and Cultures programs at a Portuguese university. The investigation is grounded in the practical application of this methodology across two academic years, from 2023/2024 to 2024/2025. The population included an Undergraduate English Culture cohort ($N = 81$), focusing on cultural literacy and critical analysis, and two Graduate Citizenship and Intercultural Education cohorts ($N = 40$ total), focusing on advanced critical citizenship and pedagogical preparation for future primary school teachers.

Figure 1 presents the sequential explanatory mixed-methods design in chronological order, showing how quantitative and qualitative data collection occurred concurrently during implementation (2023–2025), followed by sequential analysis where qualitative findings were used to explain and contextualize quantitative performance patterns. The design integrates naturalistic implementation data across diverse academic levels and student populations, allowing examination of both "what happened" (performance patterns) and "how and why it happened" (developmental processes). Portfolio assessment served as the primary course evaluation method in all cohorts, ensuring authentic stakes and maximizing ecological validity. Attrition ($N = 38$) occurred primarily due to course withdrawal or non-completion of portfolio requirements, with no differential attrition patterns observed across cohorts.

Portfolio assessment was implemented as the primary evaluation method across all three cohorts, with structured rubrics emphasizing process-oriented criteria including reflective writing, critical thinking, and pedagogical application. This naturalistic implementation context allows for documentation of both performance patterns and developmental processes as they emerged in authentic educational settings, providing insights into the practical realities of shifting from traditional to process-oriented assessment paradigms.

This study adopts a descriptive, exploratory approach to examine the implementation and impacts of portfolio assessment in teacher education, addressing the following research questions through a sequential explanatory mixed-methods design:

RQ1: What patterns of assessment performance emerge when implementing structured portfolio assessment across different academic levels (undergraduate and graduate) and cohorts?

This question examines whether consistent grading patterns can be observed when portfolios are evaluated using detailed,

Figure 1
Sequential explanatory mixed-methods research design: participant flow and analytical procedures

PHASE 1: IMPLEMENTATION AND DATA COLLECTION (2023-2025)

2023/2024 Academic Year

Graduate Cohort 1 (MA1)

Course: Citizenship & Intercultural Education

Program: MA in English Teaching

Sample: N = 20 initial → 19 final

Focus: Pedagogical preparation, initial rubric implementation

2024/2025 Academic Year

Undergraduate Cohort (UG)

Course: English Culture

Program: Foreign Languages & Cultures

Sample: N = 81 initial → 51 final

Focus: Cultural literacy, critical analysis

Graduate Cohort 2 (MA2)

Course: Citizenship & Intercultural Education

Program: MA in English Teaching

Sample: N = 20 initial → 13 final

Focus: Refined rubric with increased metacognitive weighting

QUANTITATIVE DATA
- Final portfolio grades
- Performance distributions
- Descriptive statistics
- Comparative analyses

QUALITATIVE DATA
- Student reflective essays (N=25)
- Mid-term reflections
- Instructor journal entries
- Implementation observations

↔ *Concurrent data collection across all cohorts* ↔

↓

PHASE 2: SEQUENTIAL EXPLANATORY ANALYSIS

1 **Quantitative Analysis**
Descriptive statistics, t-tests, ANOVA, ICC calculations to document performance patterns and grading consistency across cohorts (UG M=14.10; MA1 M=14.11; MA2 M=17.15)

2 **Qualitative Analysis**
Thematic analysis (Braun & Clarke, 2006) of reflective essays and instructor journal using NVivo 12; identification of metacognitive development, self-regulation, and implementation barriers

3 **Integration & Interpretation**
Qualitative findings used to explain and contextualize quantitative patterns; exploration of how process-oriented rubrics fostered metacognitive awareness and what performance data represents developmentally

FINAL ANALYTICAL SAMPLE

Total Initial: N = 121 students across three cohorts
Total Final: N = 83 portfolio submissions analyzed
Attrition: N = 38 (primarily due to course withdrawal or non-completion)

Note: UG = undergraduate cohort; MA = master's level cohorts; M = mean score; SD = standard deviation.

process-oriented rubrics across disparate student populations. Rather than claiming to establish psychometric reliability, this analysis documents observed trends in performance distribution and grading consistency across two years of implementation.

RQ2: How does reflective portfolio assessment foster metacognitive development and SRL in language teacher education contexts?

This question explores the qualitative evidence of student growth, focusing specifically on the development of metacognitive awareness (planning, monitoring, and evaluating one's own learning) and SRL behaviors as documented through student reflections and instructor observations throughout the portfolio process.

RQ3: What are the pedagogical benefits and practical challenges of implementing process-oriented portfolio assessment in teacher training programs?

This question investigates the implementation realities of portfolio assessment, documenting both the observed developmental benefits for future teachers and the systemic barriers (workload, scaffolding needs, student resistance) encountered during authentic classroom implementation.

This study contributes to the empirical literature on portfolio assessment by providing detailed documentation of implementation processes across multiple cohorts and academic levels, an area identified as under-researched in recent systematic reviews. Rather than making strong causal claims about assessment effectiveness, this research offers rich descriptive data and qualitative evidence that can inform future controlled studies and support practitioners considering portfolio adoption in their own contexts.

The research posits that portfolio assessment, when thoughtfully designed and rigorously scaffolded, offers a pedagogically valuable mechanism for fostering and documenting metacognitive growth in teacher education. However, successful implementation requires explicit institutional support, structured student scaffolding, and realistic acknowledgment of faculty workload implications. The study documents both the transformative potential and the practical challenges of this assessment approach, providing a balanced, evidence-based perspective for educators and administrators.

This study employs a sequential explanatory mixed-methods design [35], which first utilizes descriptive quantitative data to document performance patterns and grading consistency, followed by qualitative data to provide contextual depth and explain the processes underlying student development. This design is particularly appropriate for exploratory research in naturalistic settings where the goal is to understand both "what happened" (quantitative patterns) and "how and why it happened" (qualitative processes).

The quantitative component provides descriptive statistics documenting grade distributions, central tendencies, and variance across cohorts. These analyses offer insights into the consistency of grading when structured rubrics are applied, though they do not claim to establish the portfolio as a psychometrically validated measurement instrument. The qualitative component, drawing from student reflections and instructor observations, examines the developmental processes fostered by portfolio work, with particular attention to metacognitive growth and SRL behaviors.

The research was conducted over a two-year period (2023/2024 to 2024/2025) at a Portuguese university within the Language and Cultures program. The portfolio served as the main instrument for final course grading in all cohorts, ensuring authentic stakes and maximizing student engagement with the assessment process. The author served as the sole lecturer and assessor for all cohorts, ensuring consistency in pedagogical approach, scaffolding protocols, and evaluation criteria, though this also represents a limitation discussed in Section 4.1.

3.2. Participants

The study involved three distinct cohorts enrolled in two different programs, totaling an initial $N = 121$ students, with a final submission N of 83 for quantitative analysis. Participants were drawn from:

Undergraduate Cohort (UG): Students in the English Culture unit within the Foreign Languages and Cultures program ($N = 51$ final submissions). This cohort provided a baseline for testing the reliability of the portfolio rubrics across a larger, non-teaching-focused academic level.

Graduate Cohort 1 (MA 1): Students in the Citizenship and Intercultural Education unit within the MA in English Teaching program in the first year of implementation (2023/2024, $N = 19$ final submissions). This cohort provided the initial graduate data point for testing stability.

Graduate Cohort 2 (MA 2): Students in the same MA unit in the second year of implementation (2024/2025, $N = 13$ final submissions). This cohort served as the intervention group, receiving the refined assessment criteria with increased weighting on metacognitive elements.

The comparison across these cohorts allowed for robust analysis of the portfolio system's stability (UG vs MA 1) and its capacity to drive improved performance through criteria refinement (MA 1 vs MA 2).

3.2.1. Instruments

The primary instrument for data collection and pedagogical intervention was the Developmental and Assessment Portfolio. This instrument was rigorously designed to serve a dual function: to capture both the final mastery of content (summative) and the trajectory of growth and metacognitive development (formative).

A. Portfolio design and function

The portfolio was structured as a living document, modeled after constructivist learning theories, which emphasize the documentation of process over final product. The portfolio's developmental function and formative role were enforced by requiring continuous submission of dated artifacts, drafts, and revisions, along with the mandatory reflective essays. This structure compelled students to maintain an ongoing record of their learning process, providing the necessary evidence for the metacognitive loop (planning, monitoring, evaluating).

The assessment function, namely, the summative evaluation, was holistic, integrating the quality of the final professional artifacts with the rigor of the documented process. The final score (out of 20 points) was a comprehensive measure of the student's capacity to mobilize knowledge and apply theory to practice, rather than simply measuring knowledge recall.

B. Process-oriented rubric criteria

The rubrics were systematically designed to explicitly weight and reward process-oriented criteria. The emphasis was placed on high-order cognitive skills. The evolution of rubric criteria across the two graduate cohorts demonstrates the study's commitment to continuous pedagogical refinement. Table 1 presents the specific weighting changes implemented between MA1 and MA2, documenting how assessment criteria were adjusted to increase emphasis on metacognitive development while maintaining balanced evaluation across all portfolio dimensions.

Table 1
Rubric criteria comparison across graduate cohorts: evidence of increased emphasis on metacognitive development

Assessment Criterion	MA1 (2023/2024) Weight	MA2 (2024/2025) Weight	Change	Pedagogical Rationale
Reflective Writing and Critical Thinking	4 points (20%)	6 points (30%)	+2 points (+50%)	Increased to explicitly require students to document metacognitive processes (planning, monitoring, evaluating) and connect learning trajectories to teaching practice
Creativity and Pedagogical Application	5 points (25%)	5 points (25%)	No change	Maintained emphasis on authentic artifact creation (lesson plans, materials) demonstrating cultural competence application
Autonomy, Research, and Theoretical Mobilization	5 points (25%)	4 points (20%)	–1 point (– 20%)	Slightly reduced to accommodate increased metacognitive emphasis while maintaining research expectations
Organization and Professional Presentation	6 points (30%)	5 points (25%)	–1 point (–17%)	Reduced to prevent overemphasis on surface formatting relative to reflective depth
Total	20 points	20 points	Reweighted	Structural shift toward valuing process over product

Note: This table documents the intentional refinement of assessment criteria between the first and second graduate cohorts. The primary modification increased the weighting of metacognitive reflection by 50%, operationalizing the study's theoretical commitment to fostering self-regulated learning through explicit assessment of the planning–monitoring–evaluating cycle. This reweighting correlated with observed performance improvements (MA2 M = 17.15, SD = 1.2) compared to MA1 (M = 14.11, SD = 3.3), though multiple factors may have contributed to this change as discussed in Section 4.1.

Critical reflection: Criteria required progression from descriptive summarization to critical interpretation and self-analysis (e.g., reflexive writing and critical thinking). In the MA 2 cohort, this criterion was increased to 6 points, requiring explicit articulation of the link between learning and teaching practice.

Professional application: Criteria such as Creativity and Pedagogical Application required the creation of real-world artifacts (e.g., lesson plans, unit designs) that demonstrated the application of theoretical concepts, thus validating the instrument as a measure of teacher readiness.

C. Scaffolding and continuous feedback cycle

To ensure that the continuous nature of the assessment did not compromise student autonomy, a structured, three-part scaffolding protocol was implemented throughout the semester:

Initial plan and artifact submission (formative): Students submitted early-stage work and project plans, which received detailed, written corrective feedback on structure and direction but carried no grading penalty, encouraging early engagement and risk-taking.

Mid-term reflection check (formative): This mandatory checkpoint required students to submit a self-assessment essay. This protocol was critical for enforcing the metacognitive loop by forcing students to analyze learning gaps and propose specific corrective strategies before the final submission.

Final portfolio submission (summative): The complete collection of artifacts, revisions, and the comprehensive final reflective essay was submitted for the final grade.

The rigorous application of this protocol ensured consistency across all cohorts, providing the foundation for the descriptive statistical analyses.

3.2.2. Qualitative data sources and analysis protocol

Qualitative data was collected systematically throughout the implementation period from two complementary sources, designed to capture both student perspectives on their learning processes and instructor observations of portfolio implementation realities.

Student reflective essays: All students across the three cohorts were required to submit structured reflective essays at two critical junctures: mid-term (formative checkpoint) and as part of their final portfolio submission (summative assessment). These reflections were not free-form narratives but followed structured prompts designed to elicit evidence of metacognitive awareness and SRL. Specifically, students were asked to (1) analyze their learning trajectory by identifying moments of confusion, breakthrough, or conceptual reorganization, (2) document specific challenges encountered and the self-correction strategies they employed, (3) explicitly connect theoretical concepts studied in the course to their pedagogical practice or cultural understanding (particularly for MA students preparing lesson plans), and (4) evaluate their own growth in metacognitive awareness by comparing their current reflection to earlier work and identifying specific evidence of development.

For qualitative analysis, a purposive sample of 25 reflective essays was selected to ensure representativeness across all three cohorts and to capture varying levels of reflective depth. Selection criteria prioritized essays demonstrating substantive engagement with the prompts (minimum 800 words), evidence of genuine self-analysis rather than mere description, and presence of both initial mid-term and final reflections to enable analysis of developmental trajectories. The sample included 10 essays from the UG cohort, 8

from MA 1, and 7 from MA 2, proportionally representing each group's contribution to the total population.

Instructor reflective journal: Throughout both academic years (2023/2024 and 2024/2025), the instructor-researcher maintained a systematic reflective journal documenting observations about portfolio implementation. Journal entries, recorded weekly during the semester and after each major assessment checkpoint, focused on recurring patterns in student work and common misconceptions requiring intervention, scaffolding adjustments made in response to student struggles, critical incidents that illuminated the portfolio process (e.g., students' resistance to formative feedback, breakthroughs in metacognitive awareness), and the practical realities of portfolio assessment workload and time management. The journal served both as a reflexive tool for improving pedagogical practice and as qualitative data capturing the instructor's perspective on implementation challenges and student development patterns not visible in student reflections alone.

Qualitative analysis followed Braun and Clarke's [36] six-phase approach to thematic analysis, chosen for its systematic yet flexible framework appropriate for exploratory research. The lead researcher conducted initial open coding of all 25 selected reflective essays and relevant journal entries, identifying recurring patterns, phrases, and concepts related to metacognitive development, self-regulation, professional identity formation, and affective responses to the portfolio process. Codes were then grouped into preliminary themes through an iterative process of comparison and consolidation. These preliminary themes were reviewed and refined by returning to the original data to verify that coded segments genuinely supported the thematic interpretation and that themes captured meaningful patterns rather than superficial similarities.

To enhance trustworthiness and reduce single-coder bias, a subset of 10 reflections (4 UG, 3 MA 1, 3 MA 2) was independently coded by a colleague with expertise in teacher education and qualitative methodology. The two coders met to compare their coding schemes, discuss discrepancies, and refine code definitions. Inter-coder agreement, calculated using percentage agreement on the presence/absence of major codes within each essay, reached 87%, exceeding the conventional threshold of 80% for exploratory qualitative research. Remaining discrepancies were resolved through discussion and consensus, with refined code definitions applied consistently to the full dataset.

Thematic saturation (the point at which no new codes or themes emerged from the data) was assessed by analyzing reflections in chronological order of submission. Saturation was considered achieved when the final five essays analyzed yielded no new codes and all content could be categorized within existing themes. NVivo 12 qualitative data analysis software facilitated the organization, coding, and retrieval of coded segments, enabling systematic tracking of theme development and providing audit trails for transparency.

All students provided informed consent for their reflections to be used anonymously for research purposes. Student quotes included in this paper are presented with cohort identifiers (UG, MA 1, MA 2) but without individual identifying information.

3.3. Results

The results are presented in two integrated parts: descriptive quantitative data documenting performance patterns across cohorts, followed by qualitative evidence exploring the developmental processes underlying these patterns.

3.3.1. Descriptive analysis: performance patterns across cohorts

Before presenting the quantitative data, it is essential to clarify the scope and limitations of the comparative analysis. This study examines portfolio performance across three cohorts that differ substantively in academic level (undergraduate vs graduate), course objectives (cultural literacy vs pedagogical preparation), and rubric weighting (particularly the increased emphasis on metacognitive criteria in MA 2). These differences were intentional from a pedagogical perspective but create inherent limitations for cross-cohort comparison.

The quantitative analysis presented below therefore serves primarily to document observed grading patterns when structured, process-oriented rubrics are applied consistently across diverse student populations, rather than to establish the portfolio as a psychometrically validated measurement instrument in a traditional sense. The statistical tests employed examine whether meaningful differences or consistencies emerge despite contextual variations, offering insights into both the flexibility and the constraints of portfolio assessment as an evaluation method. Future research employing controlled experimental designs with standardized portfolios across comparable groups would be needed to make stronger claims about measurement reliability in a psychometric sense.

Performance distribution across three cohorts

The descriptive analysis of final portfolio grades, summarized in Table 2, documents the performance patterns observed across two academic years and three distinct student populations.

Performance patterns across academic levels

A notable finding emerged when comparing the undergraduate cohort (UG) and the first graduate cohort (MA 1): despite differences in academic level, course content, and student populations, both groups demonstrated remarkably similar mean performance (MUG = 14.10, MMA1 = 14.11). An independent-samples t-test confirmed that this similarity was not due to

Table 2
Descriptive analysis of portfolio performance across three cohorts

Cohort	Course	Year	N	M	SD
UG	English Culture	2024/2025	51	14.10	3.5
MA1	Citizenship and Intercultural Ed.	2023/2024	19	14.11	3.3
MA2	Citizenship and Intercultural Ed.	2024/2025	13	17.15	1.2

Note: UG = undergraduate cohort; MA = master's level cohorts; M = mean score (out of 20 points); SD = standard deviation. All scores represent final portfolio grades based on structured rubrics emphasizing process-oriented criteria including reflective writing, critical thinking, and pedagogical application (for MA cohorts).

chance, revealing no statistically significant difference between the mean scores of these two groups ($t(68) = 0.012$, $p = 0.991$).

While this similarity might initially appear to suggest that the portfolio assessment system produces consistent outcomes across academic levels, this interpretation requires substantial qualification. The UG and MA 1 cohorts, though achieving similar means, were evaluated on different rubrics aligned with their respective learning objectives: the UG rubric emphasized cultural analysis and critical reading, while the MA 1 rubric weighted pedagogical application and theoretical integration. The observed similarity in means therefore suggests that when rubrics are thoughtfully calibrated to appropriate expectations for each level, comparable overall performance distributions can emerge—but this does not constitute evidence of cross-level reliability in a traditional psychometric sense.

What this pattern does suggest is that the structured rubric approach maintained internally consistent grading standards within each context, a finding that offers practical insight for educators considering portfolio adoption: detailed, transparent rubrics can support consistent evaluation even when the specific criteria adapt to different course objectives. However, the comparison across fundamentally different courses and levels cannot support strong claims about the portfolio as a standardized measurement tool.

Observed effects of criteria refinement

Following the initial year of graduate program implementation (MA 1), portfolio criteria were intentionally refined for the second cohort (MA 2) to increase emphasis on metacognitive reflection. Specifically, the "Reflective Writing and Critical Thinking" criterion was increased from 4 to 6 points (out of 20 total), requiring more explicit articulation of connections between learning processes and teaching practice. This pedagogical intervention allows for examination of whether increased weighting of process-oriented criteria correlates with observable changes in performance.

A one-way analysis of variance comparing all three cohorts indicated a statistically significant difference in mean performance ($F(2, 80) = 15.68$, $p < 0.001$). Post hoc analysis (Tukey HSD) confirmed that the MA 2 cohort ($M = 17.15$, $SD = 1.2$) performed significantly higher than both the UG ($M = 14.10$, $SD = 3.5$) and MA 1 cohorts ($M = 14.11$, $SD = 3.3$), with $p < 0.01$ for both comparisons.

Two notable patterns emerge from this data. First, the substantial increase in mean performance suggests that when process-oriented criteria are explicitly weighted more heavily and described more precisely in rubrics, students appear to allocate greater effort and attention to those dimensions. However, this interpretation is complicated by potential confounding factors: the MA 2 cohort was smaller ($N = 13$), may have differed in composition from MA 1, and received the benefit of instructor experience from the previous year's implementation, making it impossible to attribute performance differences solely to rubric refinement.

Second, and perhaps more interesting, the dramatic reduction in score variance, with standard deviation dropping from approximately 3.4 (average of UG and MA 1) to just 1.2 in MA 2, suggests that more prescriptive, heavily weighted process criteria may produce more homogeneous performance distributions. This could indicate either that students more uniformly met clearer expectations or that the increased prescription reduced the range of acceptable approaches. The qualitative data (presented in Section 3.3.2) provides evidence supporting the former

interpretation, showing enhanced metacognitive engagement across most MA 2 students.

Grading consistency over time

To examine whether the instructor's application of portfolio rubrics remained consistent across the two-year implementation period, intra-rater reliability was calculated using the intraclass correlation coefficient (ICC). The ICC assesses the degree of agreement in ratings made by the same evaluator across time, with values above 0.75 generally considered indicative of good reliability.

The analysis yielded an ICC of 0.91 (95% CI [0.86, 0.94]), suggesting that the instructor's scoring patterns remained highly consistent across the 83 portfolios graded over two academic years. This finding indicates that the structured rubrics, combined with the instructor's developed expertise in applying them, supported stable evaluation practices over time, though it is important to note that this represents single-rater consistency rather than inter-rater reliability, which would require multiple independent evaluators.

This consistency is noteworthy given that portfolio grading is substantially more complex and time-intensive than traditional testing. The high ICC value suggests that detailed analytical rubrics with clearly defined performance descriptors can support consistent application even for multidimensional, process-oriented assessment. However, generalizability to other instructors or contexts would require additional research examining inter-rater reliability across multiple trained evaluators.

Interpreting the quantitative patterns

The quantitative data presented above documents three primary patterns: (1) comparable mean performance across initially different contexts when rubrics are appropriately calibrated, (2) observable performance increases correlated with increased weighting of process criteria, and (3) high consistency in grading application by a single trained evaluator over time.

These patterns offer practical insights for portfolio implementation, suggesting that structured rubrics can support consistent evaluation and that explicit emphasis on process criteria appears to focus student effort effectively, but they do not constitute psychometric validation of portfolio assessment as a standardized measurement instrument. The cross-level comparisons involve fundamentally different learning objectives and rubrics, the intervention analysis lacks experimental controls, and the consistency data reflects only single-rater reliability.

The value of this quantitative component lies in documenting that portfolio assessment, when rigorously structured, can produce interpretable and apparently consistent grading patterns across diverse implementations. These patterns provide a foundation for the qualitative analysis that follows (Figure 2), which explores the developmental processes underlying observed performance and illuminates what these quantitative patterns may represent in terms of student learning and metacognitive growth.

### 3.3.2. Qualitative findings: documenting growth and process

Thematic analysis of the student reflections and instructor journal data (described in Section 3.2.2) revealed compelling qualitative evidence that the shift to portfolio assessment successfully compelled students to engage in higher-order thinking and SRL.

A. Theme 1: metacognitive development and self-correction

Metacognitive growth was the clearest indicator of the "Grades to Growth" shift. Students were required to document and analyze their learning journey, requiring them to monitor their own performance and initiate self-correction, rather

Figure 2
Metacognitive development framework in portfolio assessment: the planning–monitoring–evaluating cycle with scaffolded checkpoints
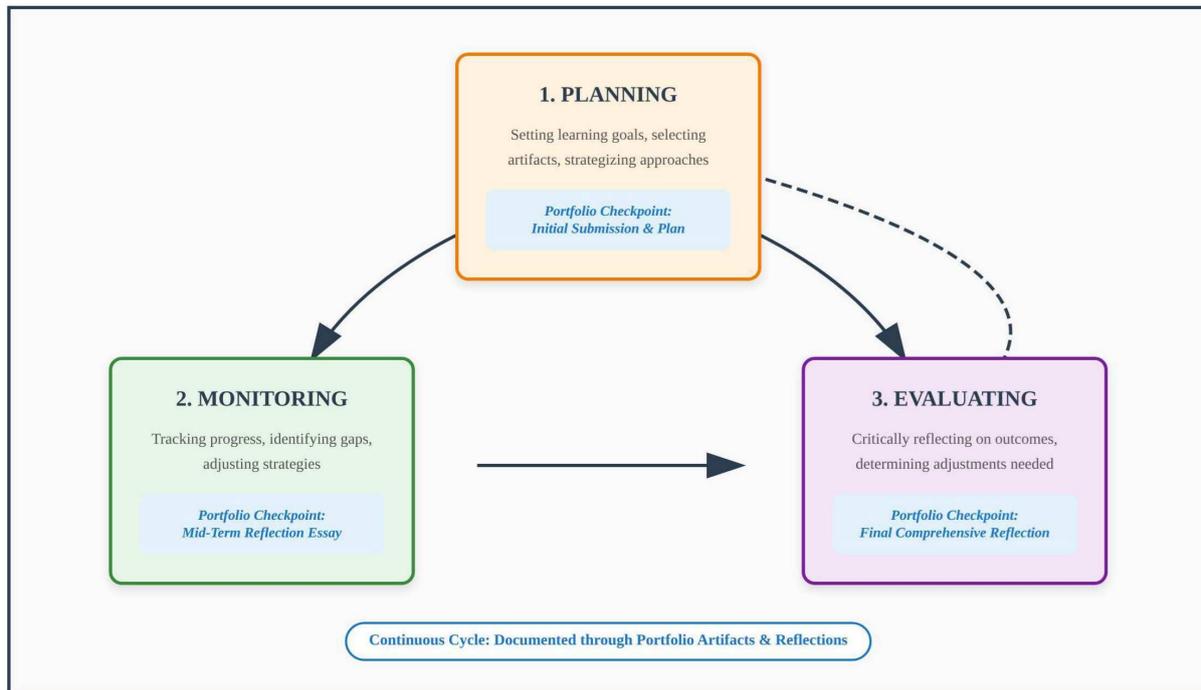
Note: This figure illustrates the metacognitive cycle operationalized through the portfolio assessment system documented in this study. The three-stage process—planning, monitoring, and evaluating—represents the core components of self-regulated learning [9] as measured through portfolio criteria. Each stage corresponds to a mandatory portfolio checkpoint: (1) initial submission requiring articulation of learning goals and artifact selection rationale; (2) mid-term reflection essay requiring explicit analysis of progress, identification of gaps, and proposed corrective strategies; and (3) final comprehensive reflection requiring critical evaluation of growth trajectory and explicit connections between theoretical learning and professional application. The dashed feedback arrow from evaluating to planning represents the cyclical nature of metacognitive development, where final reflections inform subsequent learning goals. This framework was weighted more heavily in the MA2 cohort (6 points vs 4 points in MA1) through increased emphasis on the "Reflective Writing and Critical Thinking" criterion, corresponding to observed improvements in metacognitive articulation documented in the qualitative findings (Section 3.3.2).

than passively waiting for external validation. One MA student captured this cognitive shift perfectly:

"My first reflection was very descriptive, just summarizing the course readings. Now, I see the reflection must be a dialogue with my own learning. I realized I was confusing what I knew with what I was able to apply pedagogically." (Student Comment, MA)

Instructor reflection: The mid-term reflection check was pivotal. It required several students to completely reorganize their portfolio structure because they realized their initial artifact selection did not actually support their thesis of growth, demonstrating genuine self-correction before the final summative grade.

B. Theme 2: application, autonomy, and teacher readiness

In the graduate cohorts, the portfolio's design transitioned it from an academic exercise into a direct measure of professional readiness. The criteria for creativity and pedagogical application (5 points) explicitly enforced this transfer. As one MA student noted:

"Instead of just writing about 'intercultural competence,' I had to design a 4th-grade lesson plan that modelled intercultural competence. This connection was much harder than writing a final exam, but it made the theory real for my future classroom." (Student Comment, MA)

This theme confirms that the portfolio effectively bridges the gap between academic study and professional practice, directly impacting the quality of pre-service teacher preparation. The

criterion of Autonomy, Research, and Theoretical Mobilization demanded sustained evidence of self-directed learning (SRL), successfully enforcing the professional behavior of seeking out new knowledge.

C. Theme 3: student perception of workload and the feedback paradox

While positive about the long-term impact on learning, students consistently registered the high demands inherent in the continuous contribution model, which introduced significant time management challenges.

"It was really demanding. Every week, I felt like I had to choose between perfecting a portfolio artifact and creating new entries. The continuous process is good for learning, but it's a lot of work." (Student Comment, UG)

Furthermore, analysis of end-of-year feedback revealed a significant feedback paradox. Although students received continuous, low-stakes formative feedback, they often requested *more* personalized feedback in their final, summative comments. A master's student's comment exemplified this deep-seated insecurity:

"Even though we had been given a comprehensive set of instructions from the beginning and we also had the mid-term check, I felt like I was working in the dark before the final grade. As a future teacher, I know

I should be autonomous, but I needed more continuous feedback, inter-mediate marks that showed me I was building the portfolio correctly." (Student Comment, MA)

This paradox suggests a learned dependency on high-stakes, summative feedback.

4. Discussion

This section interprets the quantitative and qualitative findings presented in Section 3.3, situating them within the broader literature on portfolio assessment and reflective practice in teacher education. We begin by acknowledging the methodological constraints that shape appropriate interpretation of our results, then explore the pedagogical implications for teacher training curricula, and conclude by proposing systemic solutions to the practical barriers documented throughout implementation. The discussion maintains focus on the study's central contribution: providing integrated empirical documentation of how process-oriented portfolio assessment can foster and document metacognitive growth when appropriately structured and scaffolded.

4.1. Validating reliability and acknowledging methodological constraints

The findings from this two-year implementation study provide complementary quantitative and qualitative evidence supporting the pedagogical value of process-oriented portfolio assessment in language teacher education. The quantitative patterns document that structured, criterion-referenced rubrics can support consistent grading practices across diverse student populations, while the qualitative data reveals meaningful metacognitive development and professional identity formation. However, before discussing these contributions in detail, it is essential to acknowledge the methodological constraints that shape the appropriate interpretation of the findings.

Methodological limitations and interpretive boundaries

This study's design involves several limitations that must be explicitly recognized. First, the comparative quantitative analysis examines three cohorts that differ fundamentally in academic level (undergraduate vs graduate), disciplinary focus (cultural analysis vs pedagogical preparation), and assessment criteria (particularly the increased weighting of metacognitive reflection in MA 2). While these differences were pedagogically intentional and the rubrics were appropriately calibrated to each context, they preclude strong claims about cross-cohort reliability in a psychometric sense. The observed similarities and differences in performance patterns offer practical insights into portfolio implementation but cannot establish the portfolio as a standardized measurement instrument comparable to validated psychometric tools.

Second, the study reflects a single instructor-researcher's implementation across all cohorts, ensuring consistency in pedagogical approach and evaluation criteria but simultaneously limiting generalizability. The high intra-rater reliability (ICC = 0.91) documents that this instructor applied rubrics consistently over time, but the critical question of inter-rater reliability, whether multiple trained evaluators would apply the same rubrics with similar consistency, remains unanswered. The subjective nature of evaluating complex, multidimensional work like portfolios means that instructor expertise, interpretation of criteria, and implicit standards may significantly influence grading patterns. Future research employing multiple independent raters would be essential to examine whether the observed consistency extends beyond a single evaluator's practice.

Third, the intervention analysis comparing MA 1 and MA 2 performance, while revealing intriguing patterns, lacks the experimental controls necessary for causal claims. The MA 2 cohort's higher performance and reduced variance could reflect the refined rubric criteria, but alternative explanations are equally plausible: smaller cohort size (N = 13 vs 19), potential differences in cohort composition or motivation, accumulated instructor expertise from the previous year's implementation, or the compound effect of multiple factors. Without random assignment and controlled conditions, attributing performance differences solely to rubric refinement would be methodologically unjustified.

Fourth, the qualitative data, while rich and systematically analyzed, reflects student self-reports of metacognitive development rather than independent behavioral measures of metacognitive functioning. Students may have learned to articulate metacognitive language without necessarily demonstrating substantive changes in self-regulatory behavior. The purposive sample of 25 reflective essays, though representative across cohorts, may not capture the full range of student experiences, particularly those of students who struggled with or resisted the reflective process. Additionally, the instructor journal data, while providing valuable contextualization, represents a single perspective that may emphasize positive developments while underweighting challenges or failures not documented through formal reflections.

Finally, the naturalistic implementation context, with portfolios serving as the sole course assessment rather than an experimental addition, precludes comparison with alternative assessment methods within the same student populations. While this design ensures authentic stakes and maximizes ecological validity, it means we cannot definitively claim that portfolio assessment produces superior outcomes compared to well-designed traditional assessments in these specific contexts.

Contributions within methodological boundaries

Despite these limitations, the study makes several valuable contributions to the empirical literature on portfolio assessment. The longitudinal, multi-cohort design provides rare documentation of implementation processes across two academic years and diverse student populations, a research approach identified as critically needed in recent systematic reviews. The integration of quantitative performance patterns with rich qualitative developmental data offers a more complete picture than studies examining either dimension in isolation. Most importantly, by documenting both the transformative potential and the substantial practical challenges of portfolio implementation, this study provides evidence-based insights that can inform practitioners and administrators considering portfolio adoption in their own contexts.

The quantitative patterns, while not establishing psychometric properties, demonstrate that thoughtfully designed portfolio rubrics can support interpretable and apparently consistent evaluation across varied implementations, a practical concern for educators skeptical that authentic assessment inevitably produces arbitrary grading. The qualitative evidence, triangulated across student reflections and instructor observations, documents substantive metacognitive growth that aligns with theoretical predictions about reflective practice's developmental impact, even as we acknowledge that self-reported growth requires corroboration through additional measures in future research.

Within the methodological boundaries acknowledged above, the findings support and extend recent research on reflective assessment and intercultural learning [5, 6, 37] and the shift to process-oriented assessment [38]. The quantitative stability documented in Table 1, with comparable mean performance across

initially different contexts (UG M = 14.10, MA 1 M = 14.11), offers practical evidence that authentic assessments, when structured through detailed rubrics, can support consistent evaluation within a single rater's practice. The transparent rubrics and criteria structure effectively standardized the grading, providing support for the assessment's consistency. The qualitative evidence provides substantive support for the system's developmental impact, documenting that the reflective components fostered meaningful metacognitive engagement. The documented shift in student language from descriptive summarization to deep, critical self-analysis provides compelling evidence that the portfolio assessment successfully fostered metacognition and SRL [9].

4.2. Implications for educational change and teacher training

The evidence from this implementation study, combined with broader empirical literature [1, 2, 13, 39], suggests the value of integrating process-oriented assessment more centrally into teacher education curricula. Teacher preparation programs would benefit from greater emphasis on assessment literacy alongside content knowledge and traditional methodology. Future teachers who are trained primarily through content-focused exams may lack the experiential foundation to confidently implement complex, process-oriented assessment methods in their own classrooms. Their preparation must include the practical skills of designing, scaffolding, and reliably evaluating authentic, process-oriented instruments. Preparing future teachers to use tools that foster and measure complex skills, such as cultural competence and critical reflection, better positions them to support primary students in navigating an increasingly globalized world. This new focus on assessment literacy should be structured around three critical pillars:

Assessment design: ensuring validity and mapping complexity
The first pillar is assessment design, requiring explicit training on how to align portfolio rubrics directly with complex competencies to ensure the validity of the assessments. Future teachers should move beyond binary or numeric grading scales to understand the architecture of an analytical rubric. This training should focus on developing rubrics where performance descriptors precisely map the progression of higher-order skills, for instance, distinguishing between a *descriptive* reflection (Level 1) and a *metacognitive and self-correcting* reflection (Level 4). By learning to construct criteria that evaluate process *over* product (e.g., assessing the quality of evidence of self-correction rather than just the final draft score), educators can better ensure their assessment tools meaningfully capture the abstract, yet essential, learning outcomes required for the 21st-century citizen, directly supporting the development of critical thinking and creativity.

Scaffolding: guiding growth without compromising autonomy
The second pillar emphasizes the essential skill of scaffolding, which is crucial to ensure that authentic assessment remains rigorous without compromising student autonomy. Teacher training programs should emphasize how to strategically structure formative checkpoints such as the Mid-Term Reflection Check used in this study to guide SRL effectively. Scaffolding is the art of providing just enough support to enable growth. Teachers benefit from learning to implement low-stakes, mandatory process checks that enforce the metacognitive loop, ensuring students receive corrective guidance when they need it most, thereby mitigating the anxiety related to workload and maximizing the students' capacity to internalize the assessment criteria before the final

submission. This technique directly addresses the students' complaints about "working in the dark" by integrating continuous, low-stakes accountability that empowers the learner's autonomy and fosters long-term self-management skills.

Reliability training: standardizing judgment and defending rigor
The third and most critical practical pillar for overcoming peer skepticism is reliability training [40, 41]. The most significant practical hurdle for authentic assessment is the perception of subjective grading. Therefore, future teachers should engage in experiential practice in standardizing subjective criteria. This component should include mock portfolio grading sessions, utilizing tools like peer-grading and inter-rater reliability exercises, to normalize judgment among different assessors. This training is important for ensuring that the observed consistency in this study (SD = 1.2) can be replicated in their own future classrooms. By actively practicing the consistent application of complex rubrics, educators can build confidence in defending the rigor and objectivity of the portfolio system against the persistent criticism that authentic assessments are inherently too subjective. Ultimately, this training transforms the reflective practitioner into a reflective assessor, guaranteeing that their assessment choices are both pedagogically sound and defensible.

4.3. Overcoming systemic barriers: prescriptive models for scalability

While the pedagogical benefits are substantial, the findings illuminate critical, interconnected practical challenges that require institutional acknowledgment and systemic solutions.

A. Policy solution for instructor workload: the assessment intensity factor (AIF)
The assessment of process-oriented portfolios requires significant intellectual labor. As noted in the instructor journal, grading took "roughly three times the time of traditional assessment." This level of individualized assessment is unsustainable without adequate administrative recognition. This study proposes the implementation of an assessment intensity factor (AIF) within faculty workload models as one mechanism for institutional recognition. The AIF mathematically recognizes the 3× complexity associated with continuous assessment, helping to ensure that portfolio implementation does not depend on unsustainable faculty over-commitment.

B. Pedagogical solution for the feedback paradox: mandated feedback loops
To address the observed feedback paradox, one promising approach involves training future teachers to implement structured feedback loops. This requires students to dedicate a specific, graded section of the e-portfolio to a "Feedback Response Log." In this log, students are required to formally quote the formative feedback received, analyze its implications, and document the specific changes made to their subsequent artifacts or reflections. This intervention eliminates the possibility of passive compliance and shifts the locus of control: the student is actively and verifiably processing the assessment input.

C. Technological solution for scalability: the e-portfolio as the engine
E-portfolio platforms offer promising solutions for addressing both administrative workload and organizational challenges. E-portfolios streamline the collection of diverse artifacts and the systematic tagging of evidence to specific rubric criteria [42]. Furthermore, the e-portfolio platform is the ideal environment

for implementing structured peer-review protocols. By training students to apply portions of the rubric to their peers' work, a practice validated for enhancing reliability, we effectively distribute the cognitive load of continuous assessment without compromising the quality of the final instructor evaluation.

5. Conclusion

This investigation provides evidence that portfolio assessment, as documented in this study and supported by broader literature [2, 19, 20], can serve as a valuable mechanism for assessing the complex competencies required of 21st-century professionals in language teacher education contexts. The study's central contribution is providing integrated quantitative and qualitative documentation of portfolio implementation across diverse contexts. The quantitative analysis documented consistent performance patterns when structured rubrics were applied across undergraduate and graduate levels (UG M = 14.10, MA 1 M = 14.11), offering practical evidence that authentic assessment, when carefully structured, can support consistent evaluation practices.

Furthermore, the observed performance patterns in the MA 2 cohort, following refinement of assessment criteria and resulting in reduced score variance (SD = 1.2), suggest that when process-oriented criteria are explicitly weighted and carefully scaffolded, more focused student performance may emerge. While alternative explanations exist, this pattern aligns with theoretical predictions about the impact of transparent, detailed assessment criteria.

The qualitative findings document substantial developmental impacts consistent with the portfolio's intended learning objectives. The portfolio process fostered engagement with the metacognitive loop, transforming student engagement from passive consumption of content into active SRL, a critical component of professional growth. For the future educators in the graduate cohort, this process was transformative, effectively merging theory with practice and validating the portfolio as an ecologically valid measure of teacher readiness. The evidence overwhelmingly supports the move "from grades to growth" by prioritizing the process competencies, the intellectual scaffolding of the "4 C's", over merely the final product.

While the implementation requires substantial instructor time investment and strategic student scaffolding, addressing these challenges through institutional policy, rather than relying on individual faculty commitment, is essential for sustainable adoption. The findings suggest that sustainable portfolio implementation depends on addressing systemic barriers through institutional support, including mechanisms such as the proposed AIF to recognize faculty labor and structured feedback protocols to support students' productive engagement with formative critique. The documented evidence from this and related studies suggests that diversifying assessment approaches to include portfolio-based methods offers valuable opportunities for fostering and documenting student growth, critical reflection, and professional application. Overreliance on summative testing alone may limit opportunities for developing the metacognitive and reflective capacities increasingly recognized as essential for professional practice.

The successful application of this model within the teacher training curriculum provides a clear call to action: Higher education institutions must integrate the theory and practical application of process-oriented assessment into teacher preparation curricula. By preparing future educators with assessment approaches that emphasize growth, reflection, and application,

teacher education programs can better support the development of complex competencies required for professional effectiveness in diverse educational contexts. This study's findings contribute to a growing body of evidence suggesting that process-oriented portfolios represent a valuable tool for documenting meaningful, developmentally focused learning experiences in higher education.

Recommendations

The current study establishes foundational evidence regarding the implementation and developmental impacts of process-oriented portfolio assessment in language teacher education, while explicitly acknowledging methodological limitations. To advance this research agenda and establish broader generalizability, several critical directions for future work merit attention.

A. Longitudinal impact and generalizability
Future studies should move beyond initial implementation to rigorously test the long-term professional impact of portfolio-based teacher preparation. A longitudinal follow-up study of MA graduates could track students from this program as they enter professional practice, comparing their assessment practices, reflective capacity, and documented efficacy in fostering 21st-century competencies in their own students against peers trained via traditional assessment models. The key inquiry is whether the portfolio's emphasis on metacognition and SRL translates into more effective, process-oriented teaching and assessment methodologies in primary classrooms.

To establish generalizability and external validity, comparative multi-institutional analyses should test the portfolio criteria and rubrics across different universities and national contexts, tracking quantitative measures (mean scores, SD) to determine whether observed consistency reflects this specific instructor and institutional context or represents more broadly replicable patterns. Testing portfolio efficacy in non-language/culture disciplines would further establish whether the model supports generalized critical thinking and SRL skills, positioning portfolios as versatile assessment tools.

B. Scalability, technology, and institutional cost
Future research must address the practical barriers documented in this study. Dedicated investigation should explore how e-portfolio platforms can be optimized for workload reduction through AI-assisted tools for artifact management, while maintaining human judgment for final evaluation. Concurrently, a formal cost–benefit analysis of the AIF should quantify trade-offs between increased faculty labor and student competency gains, potentially including employer surveys or professional trajectory tracking to establish return-on-investment metrics crucial for institutional policy decisions.

Research should empirically test the effectiveness of structured feedback loops in addressing students' learned dependency on summative grades, comparing cohorts using mandatory feedback response logs against control groups to validate this implementation strategy.

Funding Support

Ethical Statement

This study involving human participants was conducted in accordance with the ethical guidelines of the Polytechnic Institute of Porto. According to institutional regulations, formal Ethics Committee approval was not required for this low-risk educational research. All participants provided their informed consent prior to participation.

Conflicts of Interest

The author declares that she has no conflicts of interest to this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

Author Contribution Statement

Patrícia Ferreira: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

References

[1] Lambriex-Schmitz, P., Van der Klink, M. R., Beausaert, S., Bijker, M., & Segers, M. (2020). Towards successful innovations in education: Development and validation of a multi-dimensional innovative work behaviour instrument. *Vocations and Learning*, *13*(2), 313–340. https://doi.org/10.1007/s12186-020-09242-4

[2] Kukulska-Hulme, A., Friend Wise, A., Coughlan, T., Biswas, G., Bossu, C., Burriss, S. K., & Whitelock, D. (2024). *Innovating pedagogy 2024: Open University innovation report 11*. The Open University.

[3] Biesta, G. (2020). *Educational research: An unorthodox introduction*. UK: Bloomsbury Publishing.

[4] Pellegrino, J. W., & Hilton, M. L (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. USA: National Academies Press. https://doi.org/10.17226/13398

[5] Arasaratnam-Smith, L. A., & Deardorff, D. K. (2022). *Developing intercultural competence in higher education: International students' stories and self-reflection*. US: Routledge. https://doi.org/10.4324/9781003229551

[6] Dervin, F (Ed.). (2024). *Critical intercultural perspectives on higher education: Characterizing, critiquing and unsettling internationalization*. US: Routledge. https://doi.org/10.4324/9781003567950

[7] Shavelson, R. J. (2007). On the measure of achievement. In G. Schraw & M. T. McCrudden (Eds.), *Toward a new era of educational assessment* (pp. 53–78). Springer. https://doi.org/10.25656/01:5235

[8] Yang, H., & Wong, R. (2024). An in-depth literature review of e-portfolio implementation in higher education: Processes, barriers, and strategies. *Issues and Trends in Learning Technologies*, *12*(1), 65–101. https://doi.org/10.2458/itlt.5809

[9] Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

[10] 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

[11] Dinsmore, D. L., Rakita, G. F., & Kulp, A. M. (2024). Promises and challenges in developing self-regulatory assessment and feedback practices in diverse higher education contexts. In C. Evans & M. Waring (Eds.), *Research Handbook on Innovations in Assessment and Feedback in Higher Education* (pp. 16–37). Edward Elgar Publishing. https://doi.org/10.4337/9781800881600.00009

[12] Panadero, E., & Alonso-Tapia, J. (2014). Contrasting self-regulation educational theories: A theoretical review. *Psicologia Educativa*, *20*(1), 11–22. https://doi.org/10.1016/j.pse.2014.05.002

[13] Teng, M. F. (2025). *Metacognition in Language Teaching*. UK: Cambridge University Press.

[14] Liddicoat, Anthony J. (2024). Developing a critical intercultural perspective in language learning. *Electronic Journal of Foreign Language Teaching*, *21*, 97–108. https://doi.org/10.56040/atld2121

[15] Zeichner, K. M., Darling-Hammond, L., Berman, A. I., Dong, D., & Sykes, G. (2024). Evaluating and Improving Teacher Preparation Programs. *Consensus Report. National Academy of Education*. https://doi.org/10.31094/2024/2

[16] Schön, D. A. (2017). *The reflective practitioner: How professionals think in action*. UK: Routledge.

[17] Sá-Chaves, I. (2007). *Portfolios reflexivos: estratégia de formação e de supervisão*. UA Editora. https://doi.org/10.48528/5mypba36

[18] Boyd, V., & Feeting, C. (2024). Blending belonging: Using an e-portfolio as a locus for the convergence of the foundations of belonging in tertiary education. *Journal of Perspectives in Applied Academic Practice*, *12*(2), 24–34. https://doi.org/10.56433/jpaap.v12i2.603

[19] Stein, S., Andreotti, V., Restoule, J. P., Vukovic, R., McGregor, C., Pelton, L. F., . . . , & Mohajeri, A. (2025). Deepening relational capacity to confront the polycrisis in higher education and beyond. *Higher Education Research & Development*, *44*(6), 1572–1587. https://doi.org/10.1080/07294360.2025.2525109

[20] Simmonds, S. (2024). Thinking decolonisation within teacher education: Curriculum as a relational space of becoming. In L. Ramrathan, S. Maistry, & S. Blignaut (Eds.), *Critical reflections on teacher education in South Africa* (pp. 17–37). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-58090-1_2

[21] Villas Boas, B. M. de F. (2005). *Portfólio, avaliação e trabalho pedagógico*. Brazil: Papirus Editora.

[22] Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

[23] Walland, E., & Shaw, S. (2022). E-portfolios in teaching, learning and assessment: Tensions in theory and praxis. *Technology, Pedagogy and Education*, *31*(3), 363–379. https://doi.org/10.1080/1475939X.2022.2074087

[24] Colomer, J., Serra, T., Cañabate, D., & Bubnys, R. (2020). Reflective learning in higher education: Active methodologies for transformative practices. *Sustainability*, *12*(9), 3827. https://doi.org/10.3390/su12093827

[25] Korthagen, F., & Nuijten, E. (2022). *The power of reflection in teacher education and professional development: Strategies for in-depth teacher learning*. US: Routledge.

[26] Caixia, L., Bakar, Z. A., & Qianqian, X. (2025). Self-regulated learning and academic achievement in higher education: A decade systematic review. *International Journal of Research and Innovation in Social Science*, *9*(3), 4488–4504. https://dx.doi.org/10.47772/IJRISS.2025.90300358

[27] Guntur, M., & Purnomo, Y. W. (2024). A meta-analysis of self-regulated learning interventions studies on learning outcomes in online and blended environments. *Online Learning*, *28*(3), 563–584. https://doi.org/10.24059/olj.v28i3.4025

[28] Xu, X., Qiao, L., Cheng, N., Liu, H., & Zhao, W. (2025). Enhancing self-regulated learning and learning experience in generative AI environments: The critical role of metacognitive support. *British Journal of Educational Technology*, *56*(5), 1842–1863. https://doi.org/10.1111/bjet.13599

[29] Arianto, F., & Hanif, M. (2024). Evaluating metacognitive strategies and self-regulated learning to predict primary school students' self-efficacy and problem-solving skills in science learning. *Journal of Pedagogical Research*, *8*(3), 301–319. https://doi.org/10.33902/JPR.202428575

[30] Muijs, D., & Bokhove, C. (2020). *Metacognition and self regulation: Evidence review*. Education Endowment Foundation. https://files.eric.ed.gov/fulltext/ED612286.pdf

[31] European Journal of Education and Pedagogy. (2024). The use of e-portfolios as a teaching, learning and assessment tool in higher education: Differing opinions among Ghanaian pre-service teachers and nurses. *European Journal of Education and Pedagogy*, *5*(6), 35–45. https://doi.org/10.24018/ejedu.2024.5.6.858

[32] Zaabalawi, R. S., & Zaabalawi, J. (2024). Portfolios versus exams: A study to gauge the better student assessment tool. *Language Testing in Asia*, *14*(1), 28. https://doi.org/10.1186/s40468-024-00296-y

[33] Doğan, Y., Yıldırım, N. T., & Batdı, V. (2024). Effectiveness of portfolio assessment in primary education: A multi-complementary research approach. *Evaluation and Program Planning*, *106*, 102461. https://doi.org/10.1016/j.evalprogplan.2024.102461

[34] Farrell, T. S. (2025). 'Bending Back' to move forward: Contemplating reflective practice in TESOL. *Teaching English as a Second Language Electronic Journal*, *29*(2). https://doi.org/10.55593/ej.29114s1

[35] Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). US: SAGE Publications.

[36] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[37] Ferreira, A. P. (2025). A educação intercultural na formação inicial dos professores de inglês do primeiro ciclo do ensino básico. *PRATICA - Revista Multimédia de Investigação em Inovação Pedagógica e Práticas de E-Learning*, *8*(1), 1–11. https://doi.org/10.34630/pel.v8i1.5976

[38] Boud, D., & Falchikov, N. (2007). *Rethinking assessment in higher education: Learning for the longer term*. US: Routledge.

[39] Adachi, C., Ajjawi, R., Apps, T., Bakharia, A., Bearman, M., Beckman, K., . . ., & Winstone, N. (2020). *Re-imagining university assessment in a digital world*. Switzerland: Springer. https://doi.org/10.1007/978-3-030-41956-1

[40] Sridharan, B., McKay, J., & Boud, D. (2023). The four pillars of peer assessment for collaborative teamwork in higher education. In O. Noroozi & B. De Wever (Eds.), *The Power of Peer Learning. Social Interaction in Learning and Development*. Springer. https://doi.org/10.1007/978-3-031-29411-2_1

[41] Khasawneh, M. A. S., Aladini, A., Assi, S. A., & Ajanil, B. (2025). Portfolio assessment in AI-enhanced learning environments: a pathway to emotion regulation, mindfulness, and language learning attitudes. *Language Testing in Asia*, *15*(1), 5. https://doi.org/10.1186/s40468-025-00345-0

[42] Barrett, H. C. (2007). Researching electronic portfolios and learner engagement: The REFLECT initiative. *Journal of Adolescent & Adult Literacy*, *50*(6), 436–449. https://doi.org/10.1598/JAAL.50.6.2