BETWEEN ALGORITHM AND AGENCY: DESIGNING INCLUSION WITH UNIVERSAL DESIGN FOR LEARNING IN THE AGE OF GENERATIVE ARTIFICIAL INTELLIGENCE

Lia Daniela Sasanelli [1]
Pegaso University
liadaniela.sasanelli@unipegaso.it

Alfonso Filippone
Pegaso University
alfonso.filippone@unipegaso.it

Umberto Barbieri
Pegaso University
umberto.barbieri@unipegaso.it

Double Blind Peer Review

Citation
Sasanelli, L. D., Filippone, A., & Barbieri, U. (2026). Between algorithm and agency: Designing inclusion with universal design for learning in the age of generative artificial intelligence. *Italian Journal of Health Education, Sport and Inclusive Didactics*, *10*(4). https://doi.org/10.32043/gsd.v10i4.1908

Doi: https://doi.org/10.32043/gsd.v10i4.1908

Copyright notice:
© 2026 The Author(s). This is an open access, peer-reviewed article published by Cares S.r.l. – REN Press and distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

renpress.it/index.php/gsdj
ISSN: 3103-5744

ABSTRACT
In the transition from digital competence to Digital Affective Literacy, instructional design emerges as a space for negotiation between the human and the artificial. This paper presents a quasi-experimental study with repeated measures and a crossover design involving 100 TFA special education trainees who were tasked with designing lessons based on Universal Design for Learning under three conditions: no AI, use of a generalist chatbot, and use of an MCP server designed according to inclusive principles. The lesson plans were evaluated by experts using a structured UDL rubric, while reflective logbooks were analysed qualitatively to investigate perceived agency, uncertainty, delegation and awareness during the design process. The results show that the UDL-MCP server supports usability and provides clearer guidance in translating theoretical knowledge of UDL into concrete inclusive design choices. The integration of assessment data and reflective narratives suggests that AI can act as a symbolic mediator when it preserves teacher agency and strengthens inclusive intentionality in hybrid educational contexts.

KEYWORDS
Universal Design for Learning; Artificial intelligence; Teacher training.

Received 2026-07-11
Accepted 2026-10-27
Published 2026-10-30

1. Introduction

The introduction of generative Artificial Intelligence (AI) systems into educational settings raises a question that exceeds the plane of instrumental efficiency and bears upon the relational dimension of teaching. The teaching profession has historically been constituted in the tension between a *technical component*, the planning, sequencing and assessment of learning – and a *relational component*, grounded in the recognition and accompaniment of the learner. It is in this second component, above all, that the most demanding challenge of inclusive teaching is concentrated: designing pathways capable of anticipating, rather than merely compensating for, the variability and heterogeneity of learners (CAST, 2024; Meyer et al., 2014). The hypothesis that a generative system might intervene in the design phase of inclusion therefore concerns not only the teacher's technical competences, but the very nature of the design act as a relational act.

Within this frame, *Universal Design for Learning (UDL) I*s not a mere repertoire of devices for accessibility, but an approach that reconfigures design as an exercise of inclusive awareness: rather than locating the cause of difficulty in the learner, it traces the barrier back to the rigidity of the learning environment and requires the teacher to anticipate it at the point of conception, picturing in advance the plurality of minds that will learn. The most recent revision of the Guidelines, version 3.0 (CAST, 2024) heightens this scope: by moving affect from the margin to the centre of the engagement principle, it extends inclusive awareness beyond the accessibility of content, so as to encompass the affirmation of identities, the sense of belonging and the recognition of mechanisms of exclusion. Fostering such awareness in trainee teachers therefore constitutes a far from secondary concern, one that the advent of a

generative system able to return ready-made designs may equally support or undermine.

Affective neuroscience has documented that the processes traditionally regarded as the core of learning are functionally integrated with emotion (Immordino-Yang & Damasio, 2007). This insight also bears upon trainee teachers who learn to design with, or in opposition to, a generative system: the affective dimension is thus doubly implicated, constituting at once the object that inclusive design must take into account and the *medium* through which prospective teachers develop trust in, resistance to, or improper reliance upon the tool. This twofold implication reveals a limitation in the models through which teacher education has framed the relationship with the digital – models centred, as in the case of the DigCompEdu framework (Redecker, 2017), on what the educator must be able to do with technology, yet silent on the affective and regulatory components of its use. We propose to define this level as *Digital Affective Literacy*: the capacity to sustain a relational and inclusive intentionality when designing through an artificial system, without delegating one's judgement to generative fluency, nor forgoing its contribution out of apprehension.

The theoretical urgency of the construct stems from a documented cognitive mechanism: generative systems facilitate the offloading not only of executive effort, but of the processes of thought themselves (Risko & Gilbert, 2016), such that the teacher, on receiving a polished design product, may in effect suspend the activity of designing. Whether generative AI amplifies or erodes inclusive intentionality thus depends less on the teacher's competence as such than on the design characteristics of the tool, that is, on whether it operates as a device that merely supplies answers or as a semiotic mediator (Vygotsky, 1978) that reorganises thought rather than replacing it.

Proceeding from these premises, the present contribution takes the question as the object of empirical enquiry. Its general *aim* is to examine the extent to which, and the processes through which, recourse to a generative AI system bears upon the quality of the inclusive teaching design produced by trainee support teachers and upon the design intentionality that sustains it. Three *specific objectives* contribute to this aim: (a) to compare the quality of the designs produced without AI, with a generalist chatbot, and with a chatbot aligned to UDL principles, taking the 3.0 Guidelines as the criterion of reference; (b) to explore whether, and in what terms, the nature of the tool – a device that returns finished answers or a mediator that prompts the design decision – modulates not only the outcome of the product, but the degree of delegation or cognitive elaboration on the part of the subject; (c) to elicit the affective and reflective traces that accompany the design act, so as to submit *Digital Affective Literacy* to empirical enquiry as a dimension distinct from technical competence.

In keeping with these objectives, the research is guided by two *questions*. The first, predominantly quantitative, concerns the product: *in what way does the use of a generative system, and specifically its design architecture, affect the inclusive quality of the artefacts produced by trainee teachers*? The second, qualitative, concerns the process: *which affective experiences and reflective movements does such use activate in the designing subject, and with what implications for the consolidation of inclusive intentionality as a professional competence*? The distinction between the two planes, the artefact and the process that generates it, is not merely methodological, but reflects the theoretical assumption that a formally inclusive product may coexist with an impoverishment of the teacher's design agency.

Regards the design, the study is framed as an exploratory pilot and adopts quasi-experimental, repeated-measures arrangement with a cross-over structure: one hundred trainees on a special-needs support teaching programme (TFA) produced designs informed by UDL principles across the three conditions referred to above, following a balanced sequence intended to control for order and learning effects. Consistent with the twofold nature of the questions, the enquiry adopts a mixed-methods approach, integrating the quantitative assessment of the artefacts, entrusted to a panel of experts, with the qualitative analysis of the participants' reflective journals on the process side. The assessment of the products draws on a structured rubric (in the appendix), adapted from the *UDL Progression Rubric* (Novak & Rodriguez, 2018) and recalibrated against the 3.0 Guidelines (CAST, 2024): articulated into the three principles, the nine guidelines and the thirty-six considerations, it expresses for each indicator a judgement across four progressive levels – emerging, developing, proficient and expert – and, by virtue of its focus on the engagement principle, allows the affective dimension of inclusion, decisive within the hypothesis advanced here, to be appraised as well. Far from serving as a mere measurement instrument, the rubric performs a pivotal function in this study: it translates the UDL approach into an operational, shared criterion by which a panel of experts assesses the inclusive quality of the design products, rendering the three conditions comparable and anchoring the evaluative judgement to the principles of the 3.0 Guidelines rather than to subjective impressions. Before the research design is presented, the theoretical framework is organised around four nuclei: the evolution of UDL from accessibility towards the affective dimension of inclusion; the so-called affective turn in the learning sciences; the shift from digital competence to *Digital Affective Literacy*; and the still-open question of the effects of generative systems on the teacher's design processes.

2. Background

2.1 Universal Design for Learning and the affective dimension

UDL rests on a conceptual inversion of traditional compensatory models. Rather than asking how to adapt the learner with a disability to a curriculum conceived for an "average" student, the approach developed by CAST proceeds from the need to design the curriculum, from its very inception, in accordance with the foreseeable variability of learning minds (Meyer et al., 2014).

The neuroscientific root of the model is not a novelty of the most recent revision. From its very formulation, UDL drew its principles from the identification of three neural networks involved in learning, recognition, strategic and affective, to which the means of representation, of action and expression, and of engagement correspond (Compagno et al., 2025); from the outset, therefore, engagement is tied to the system that governs motivation and emotional regulation. What version 3.0 introduces (CAST, 2024) is thus not the entry of affect into the model, but its displacement from the margin to the centre: whereas the emotional dimension was previously confined to the eliciting of interest, the update places at the heart of engagement the affirmation of identities, the sense of belonging, joy and empathy, together with the overcoming of bias and of mechanisms of exclusion. Inclusion, on this view, is not exhausted in the accessibility of content, but bears upon the quality of the emotional experience of the one who learns: on the design plane, the model thus converges towards the very insight reached by affective neuroscience, namely the impossibility of conceiving of learning in abstraction from its emotional component.

2.2 The affective turn in inclusive design

The construct of emotional thought (Immordino-Yang & Damasio, 2007) shows how emotion does not interfere with rational processes from without, but constitutes an intrinsic component of them; within the Italian pedagogical tradition, Lucangeli (2019) has condensed the import of this insight in the notion of *hot cognition*, according to which all learning is fixed together with the emotional experience that accompanies it, with facilitating or inhibiting effects upon its subsequent retrieval.

The extension of this principle from learners to trainee teachers represents a less customary, yet particularly significant, move for educational research. The design of an inclusive pathway is not an emotionally neutral technical operation: it unfolds under conditions of uncertainty and entails an investment of identity, heightened by the responsibility towards the most vulnerable learners. The introduction of a generative system is therefore grafted onto an already affectively charged space, activating heterogeneous reactions: from relief at the availability of a starting point, to anxiety bound up with a possible deskilling of one's role, through to an excessive reliance that diminishes the tolerance of uncertainty. In the light of the hot-cognition paradigm, such reactions are not epiphenomena to be neutralised, but constitutive elements of professional learning, liable to be reactivated in subsequent uses of the tool. It follows that research focused exclusively on the quality of the design artefact, and indifferent to the affective reworking of the subject who produces it, would grasp the phenomenon only in part.

2.3 From digital competence to Digital Affective Literacy

The principal conceptual instruments concerning the relationship between teachers and technologies were defined before the diffusion of generative systems capable of complex linguistic interaction. The DigCompEdu framework (Redecker, 2017), still the most authoritative European reference for teachers'

digital competence, retains an operational centre of gravity, focused on the operations of selecting, producing, managing and evaluating resources. It was not conceived for a device that does not merely transmit content, but proposes and argues for it in dialogical form, exerting upon its interlocutor the relational pressure proper to natural language.

That technical mastery alone does not exhaust the question is, moreover, borne out by empirical research. In a study conducted with a sample of trainee teachers, Lecce and Di Tore (2025) find that declared familiarity with digital tools does not automatically translate into the capacity to integrate them knowingly into disciplinary content, and that it is rather the direct experimentation with AI systems, more than the possession of technical competences considered in the abstract, that correlates with an increased perception of design competence and with a more reflective and critical disposition towards the tool. Read within the TPACK model, the finding suggests that the crux lies not in technological knowledge as such, but in its articulation with the pedagogical dimension and with the processes that govern its use; it is precisely in this interstice, left in shadow by operational frameworks, that the construct proposed here is situated. Consistently with this reading, a large-scale study of prospective educators shows that the disposition towards the use of artificial intelligence in inclusive settings is inseparable from the sphere of the subjects' knowledge, appraisals and emotions (Zanazzi et al., 2025). The training need that emerges from it cannot be resolved on the technical plane, but requires a pedagogical grounding and the capacity for critical analysis as a crucial competence: a finding that corroborates the need to thematise the affective-regulatory component under discussion here.

We therefore propose the notion of *Digital Affective Literacy* as a complement to, and not a substitute for, digital competence. Whereas the latter concerns the capacity to use the tool, Digital Affective Literacy pertains to that of preserving an inclusive intentionality during its use: recognising the pull exerted by a confidently formulated answer, keeping the space of the pedagogical decision open, and steering the interaction towards a process of growth rather than of dependence. In the terms of hot cognition, it is the disposition that determines whether the affective trace associated with the use of AI will be encoded as an enhancement or as a replacement of professional agency: two outcomes that may prove indistinguishable at the level of the product and profoundly different at the level of the processes that generated it.

2.4 Generative artificial intelligence between delegation and elaboration

The potential of generative systems for inclusive design is well documented. A device able to reformulate a text across different levels of complexity, to propose alternative expressive modalities, or to suggest representations not envisaged by the designer may constitute a support consistent with the UDL imperative of the multiplicity of options (Barbieri et al., 2025). The same mechanism, however, carries a mirror-image risk. The ease with which a fluent system returns a finished product favours cognitive offloading (Risko & Gilbert, 2016), that is, the transfer of the mental load to an external resource. This process is not in itself dysfunctional – the use of calculation tools, for instance, frees cognitive resources for solving the problem – but it becomes so when what is offloaded is the specific cognitive operation that the activity was meant to develop. In inclusive design, that operation consists in anticipating the barrier encountered by a particular learner and in intentionally deliberating the strategies to remove it (Filippone et al., 2025a). Its delegation entails the production of a formally inclusive artefact alongside the failure to develop

inclusive intentionality, that is, the very component that, in the terms of hot cognition, would consolidate as a professional competence. There thus emerges the distinction, highlighted by recent studies, between a use of AI directed at *elaborating* the subject's thought and one aimed at *delegating* it (Filippone et al., 2025b).

From this follows a precise design question: whether it is possible to conceive of a generative system capable of countering its own persuasive efficacy, behaving not as a device that concludes the process of thought, but as a mediator that sustains it. The question is inscribed within a broader reflection that figures artificial intelligence as a didactic and relational *third way*, valid only on condition that the centrality of the educational relationship and of the teacher as a mediator of meaning be preserved (Bellantonio, 2025). The Vygotskian distinction between tools oriented towards the execution of a task and *psychological tools* that reorganise the subject's mental processes (Vygotsky, 1978) offers an interpretive frame in this respect: a generalist assistant, optimised for the return of a finished answer, and a chatbot structured on UDL principles, designed to prompt the justification of choices and the maintenance of an open decision, do not represent variants of the same tool, but embody divergent conceptions of the supporting function. It is this contrast that the present study sets out to operationalise, proceeding from an exploratory pilot study in which a chatbot based on the UDL 3.0 Guidelines was used to support pre-service teachers in inclusive design (Sasanelli et al., 2025). From this standpoint, the research compares its effects not only on the quality of the design product, but on the affective and reflective traces discernible in the designing subjects, taking artificial intelligence not as an executor, but as a potential symbolic mediator of inclusive intentionality.

3. Method

3.1 UDL-aligned generative mediator

UDL-MCP gives any generative chatbot expertise in Universal Design for Learning. It runs as a Model Context Protocol (MCP) server, an open and model-agnostic standard for connecting AI applications to external context (Anthropic, 2024). Any MCP-compatible client can connect to it. UDL-MCP exposes the UDL 3.0 framework (CAST, 2024) through the three components of the protocol. Resources hold the guidelines, the 36 considerations, and worked examples of inclusive practice. Tools run design operations on a draft lesson. Prompts ask the user to state the inclusive rationale behind each choice. UDL-MCP works as a semiotic mediator. It returns criteria, examples, and questions, and the design team writes the lesson. The scaffolds act on the justification of each choice, so the tool reaches the reflective process that Digital Affective Literacy describes.

3.2 Pilot Study

An exploratory pilot study tested the feasibility of the method and provided preliminary estimates rather than a confirmatory test of condition effects. The design was a within-subjects cross-over arranged as a 3 × 3 Latin square. The condition had three levels: NoAI (no generative tool), generalist (each group chose ChatGPT, Claude, Gemini, or NotebookLM), and UDL-MCP. One hundred trainees on a special-needs support teaching programme (TFA) took part. They worked in 11 class-council (*consiglio di classe*) groups. Each group produced one lesson per condition across three sessions. Participants had previously encountered UDL at a theoretical level, although this knowledge was not yet fully consolidated as an autonomous design competence. The Latin-square arrangement was used to distribute the three conditions across periods and sequences, thereby reducing systematic order and learning effects in the

repeated-measures design. The rubric scores were analysed at the group level (*n* = 11 groups, 33 lessons), and trainees completed the SUS individually (*n* = 92).

Each session set a civic-education brief: diet and body image in Period 1, relationships and social media in Periods 2 and 3. One group skipped one lesson, so 32 of the 33 lessons were scored. Each lesson was scored with a rubric adapted from the UDL Progression Rubric (Novak & Rodriguez, 2018) and recalibrated against the 3.0 Guidelines (CAST, 2024). The rubric covers the three UDL principles, nine guidelines, and 36 considerations: Engagement (13), Representation (12), and Action and Expression (11). Each consideration runs from 1 (*emerging*) to 4 (*expert*). The adaptation was intended to align the original progression structure with the 36 considerations of the UDL 3.0 Guidelines while preserving the four-level developmental scale. Given the exploratory nature of the pilot, the adapted rubric should be regarded as an operational assessment tool rather than as a fully psychometrically validated instrument. The lesson score is the mean of its 36 considerations, and each subscore is a principle-level mean.

Table 1: *Cyclic 3 × 3 Latin-square cross-over design: condition by sequence and period*

Sequence	Groups	Period 1	Period 2	Period 3
S1	1-3	NoAI	UDL-MCP	Generalist
S2	4-7	Generalist	NoAI	UDL-MCP
S3	8-11	UDL-MCP	Generalist	NoAI

Note. Generalist = group-selected generalist generative AI (ChatGPT, Claude, Gemini, or NotebookLM). Each condition appears once per period and once per sequence (column-balanced). The square is cyclic, not Williams.

After the UDL-MCP sessions, trainees rated its usability on the 10-item System Usability Scale (Brooke, 1996), in Italian, scored 0 to 100, with Usability and Learnability subscales. Published benchmarks framed the result (Bangor et al., 2009; Lewis & Sauro, 2018). Expert judges scored the lessons blind to condition, period, group, and sequence. One lesson was scored by more than one judge,

and its scores ranged from 1.25 to 2.61, so judge severity entered the model as a covariate. The overlap covered only that lesson, so inter-rater agreement could not be estimated.

A linear mixed-effects model fitted each lesson score y:

$$y = \beta_0 + \beta_{condition} + \beta_{period} + \beta_{judge} + u_{group} + \varepsilon,$$

where u_{group} is a random intercept for group and ε the residual. Two one-sided tests checked equivalence between conditions against a bound of half a rubric level (±0.5). Descriptive statistics and a 95% confidence interval summarised the SUS. Trainees also kept reflective journals; that strand is reported separately.

3.3 Reflective logbooks and qualitative analysis

To complement the quantitative assessment of the lesson plans, the study included a qualitative strand based on structured reflective logbooks completed by the trainees during the design process. This methodological choice is consistent with mixed-methods approaches in educational research, where qualitative data are used to illuminate the meanings, perceptions and processes underlying quantitative outcomes (Creswell & Plano Clark, 2018; Flick, 2022). In particular, reflective logbooks were adopted as metacognitive devices capable of making visible the reasoning, uncertainty, delegation and self-regulation processes that often remain implicit during AI-mediated learning and design activities (Flavell, 1979; Schön, 2017).

The use of logbooks was also coherent with previous research on generative artificial intelligence and cognitive processes in teacher education, in which written reflections and focus groups were used to analyse whether interaction with AI promoted cognitive delegation or autonomous elaboration (Filippone et al., 2025a). In that study, the qualitative strand made it possible to

distinguish between AI systems that provide ready-made answers and systems that scaffold reasoning through prompts, questions and metacognitive guidance. The present study extends this methodological logic to the field of inclusive instructional design, with specific reference to Universal Design for Learning.

After each design activity, trainees were invited to document their experience through a structured reflective logbook. The prompts focused on four main areas: perceived difficulty in applying the UDL framework; perceived usefulness of the technological support, when present; degree of orientation or disorientation during the design process; and perception of agency, delegation or guidance in relation to the artificial system. The logbooks were therefore not conceived as simple satisfaction reports, but as process-oriented artefacts aimed at capturing how trainees moved from theoretical knowledge of UDL to its concrete application in lesson planning. The qualitative corpus consisted of 32 group reflective logbooks collected across the three experimental conditions. Each of the 11 class-council groups was invited to complete one logbook after each design activity, yielding an expected corpus of 33 logbooks. One group did not complete one of the scheduled activities; consequently, the corresponding logbook was not available for analysis. The final corpus therefore comprised 32 logbooks, each associated with one group, one experimental condition and one design session. All logbooks were originally written in Italian. Before analysis, the logbooks were pseudonymised through alphanumeric codes identifying the group, the experimental condition and, where relevant, the study period. The codes did not contain personal identifiers. The abbreviations NoAI, GAI and MCP were used to indicate, respectively, the no-AI condition, the generalist-AI condition and the UDL-MCP

condition. For example, G04-GAI-P2 identifies the reflective logbook produced by Group 4 during the generalist-AI condition in Period 2.

The excerpts reported in the Results section were translated from Italian into English by the author who conducted the qualitative analysis. The translations aimed to preserve the semantic content, tone and reflective meaning of the original entries. Only minimal syntactic adjustments were made where necessary to ensure readability in English.

The main dimensions addressed by the reflective prompts are summarised in Table 2. These dimensions guided data collection but were not treated as a fixed or exhaustive coding framework. The themes presented in the Results section were developed recursively through engagement with the complete corpus.

This aspect was particularly relevant because, although the participants had already encountered Universal Design for Learning at a theoretical level, many of them reported that this knowledge was not sufficiently consolidated to support the autonomous development of a complete UDL-based lesson plan. The qualitative data therefore made it possible to investigate the gap between declarative knowledge and situated design competence. In this sense, the logbooks provided evidence on the ways in which AI systems helped trainees orient themselves within a framework that was known conceptually but not yet fully internalised as a design grammar.

The qualitative corpus was analysed through reflexive thematic analysis, following the six recursive phases proposed by Braun and Clarke (2006, 2021): familiarisation with the data, generation of initial codes, development of candidate themes, review and refinement of themes, definition and naming of themes, and production of the analytic account. The analysis combined

inductive attention to recurrent meanings in the logbooks with a theoretically informed interpretation guided by the concepts of cognitive offloading, semiotic mediation, reflective practice and inclusive intentionality.

The qualitative analysis was conducted by one author, who became familiar with the complete corpus through repeated reading of the logbooks. Initial codes were generated around expressions of uncertainty, orientation, perceived support, delegation, justification of design choices, learner variability and inclusive awareness. These codes were progressively compared, reorganised and interpreted in relation to the research questions and the theoretical framework, leading to the development and refinement of the five themes.

Interpretative credibility was supported through repeated engagement with the complete corpus, recursive checking of candidate themes against the original logbooks, and the use of representative excerpts to make the relationship between data and interpretation transparent. Tables 2 and 3 document the progression from the initial analytical dimensions to the final thematic structure, providing a concise audit trail of the analytical process.

Consistent with reflexive thematic analysis, the themes were not treated as objective categories mechanically extracted from the data, but as interpretative patterns developed through the researcher's sustained engagement with the corpus. The analysis therefore did not aim to calculate statistical agreement between independent coders. The final thematic structure was illustrated and supported through the selection of representative verbatim excerpts.

Table 2: Qualitative dimensions explored through reflective logbooks

Dimension	Analytical focus	Theoretical connection
Prior familiarity with UDL	Perceived gap between theoretical knowledge of UDL and ability to use it in lesson planning	Transition from declarative knowledge to situated design competence
Orientation support	Perception of AI as a tool for approaching a complex or only partially mastered framework	AI as cognitive support and reduction of initial disorientation
Delegation risk	Tendency to rely on AI-generated outputs without fully reworking or justifying design choices	Cognitive offloading and possible weakening of teacher agency
Reflective guidance	Perception of UDL-MCP as a guide that clarifies steps, criteria and inclusive rationale	Scaffolding, mediation and reflective practice
Inclusive intentionality	Ability to connect design decisions with learner variability, barriers and UDL principles	Digital Affective Literacy and preservation of pedagogical agency

The interpretative process was informed by three theoretical references. First, the notion of cognitive offloading, according to which external tools may reduce mental effort but may also weaken the very cognitive processes that the task is intended to develop (Risko & Gilbert, 2016). Second, the Vygotskian idea of mediation, which allows a distinction between tools that merely execute operations and psychological tools that reorganise thought (Vygotsky, 1978). Third, the construct of reflective practice, according to which professional learning emerges when practitioners are supported in making their own decisions, assumptions and uncertainties explicit (Schön, 2017).

In the overall mixed-methods design, the logbooks played a complementary role to the UDL rubric. Whereas the rubric assessed the inclusive quality of the

final lesson plans, the qualitative analysis explored the subjective and processual dimensions of design: perceived agency, uncertainty, guidance, delegation and reflective awareness. This distinction was essential for the purposes of the study, since a formally inclusive product may coexist with different degrees of teacher authorship, cognitive elaboration or reliance on the artificial system.

4. Results

4.1 Preliminary Findings

The pilot covered 11 groups and 100 trainees. Each lesson received a score from 1 to 4 on a 36-item UDL rubric, under three conditions: no AI, a generalist assistant, and UDL-MCP.

The three conditions reached similar inclusive quality. Mean rubric scores were 1.86 for no AI, 2.17 for the generalist, and 1.97 for UDL-MCP (Figure 1, panel A), and the omnibus test was null, $\chi^2(2) = 1.92$, $p = 0.38$. The pilot detects only large effects: with one rating per lesson it reaches $d_z \approx 1.0$, while the observed condition differences span $d_z = 0.17$ to 0.43. Group identity and judge severity each explained more variance than condition (ICC = 0.70). One equivalence test reached the ±0.5 bound: UDL-MCP against the generalist tool (TOST $p = 0.009$). The two comparisons against no AI stayed inconclusive ($p = 0.064$ and 0.068; their 90% confidence intervals crossed the bound).

The similar product scores point past the product. A lesson can read as inclusive whether or not the teacher led the inclusive choices. The rubric measures the lesson; the reflective journals measure the design process. UDL-MCP received a mean SUS score of 72.0 (n = 92, SD = 18.8, 95% CI [68.1, 75.8]; Figure 1, panel B), slightly above the conventional benchmark of 68 for acceptable perceived usability. Since no comparative usability measure was collected in the other

conditions, this result should not be interpreted as evidence that UDL-MCP improved usability relative to another tool (Bangor et al., 2009; Lewis & Sauro, 2018).

One limit bounds the rubric results. Only one lesson was scored by more than one judge, and its scores diverged sharply (1.25 to 2.61), so inter-rater agreement could not be estimated. The near-perfect internal consistency (α = 0.983) reflects single scoring and a halo effect. A reliable estimate requires several scores per lesson.

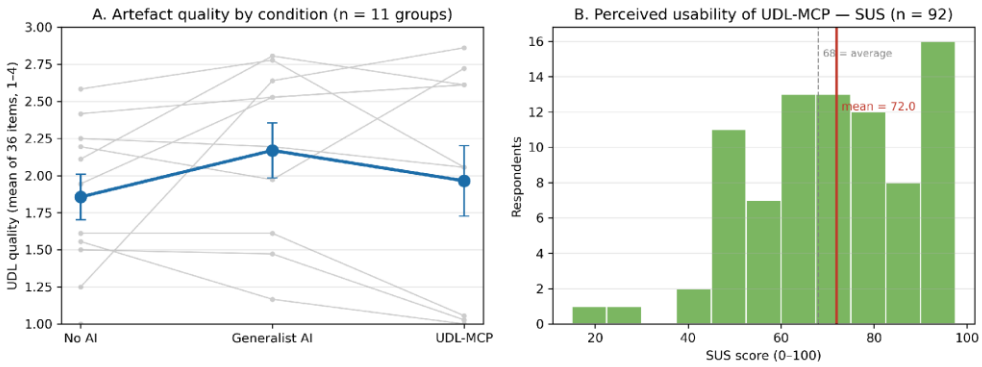


Figure 1: Inclusive-design quality across conditions and perceived usability of UDL-MCP

Note. Panel A: mean UDL rubric score (1–4) per condition, with group trajectories in grey and condition means in blue (±1 *SE*); *n* = 11 groups. Panel B: System Usability Scale scores for UDL-MCP (*n* = 92); dashed line = 68-point benchmark, solid line = mean (72.0).

4.2 Qualitative findings from reflective logbooks

The reflective logbooks added a process-oriented dimension to the interpretation of the pilot findings. Reflexive thematic analysis generated five themes concerning the operationalisation of UDL, the ambivalent role of generalist AI and the specific contribution of UDL-MCP. All excerpts were originally written in Italian and translated into English by the author who conducted the qualitative analysis.

The first theme concerned the gap between theoretical familiarity with UDL and its practical application: “Knowing the UDL principles in theory was not enough. The most difficult part was deciding how to translate them into concrete choices for the lesson” *(G05-NoAI-P2, author’s translation)*. This reflects the distance between declarative knowledge and situated design competence.

Generalist AI primarily functioned as orientation support. As one group reported, “The AI helped us overcome the initial difficulty of starting. It gave us some ideas and a possible structure from which we could begin to develop the lesson” *(G04-GAI-P1, author’s translation)*. The tool therefore reduced initial uncertainty and facilitated entry into the design task.

However, its fluent outputs also created a risk of delegation: “The response generated by the AI was already very complete and well written, so we were initially inclined to accept it without making substantial changes” *(G06-GAI-P1, author’s translation)*. This suggests that ease of use may encourage cognitive offloading when outputs are insufficiently revised or contextualised.

UDL-MCP was instead perceived as a clearer design guide: “The tool did not simply give us a ready-made lesson. It helped us justify our decisions by linking them to specific UDL criteria” *(G07-MCP-P3, author’s translation)*. Its prompts made the connection between design choices and UDL principles more explicit.

Finally, UDL-MCP appeared to strengthen inclusive intentionality: “The guided questions encouraged us to reflect on the reasons behind our choices rather than simply adding inclusive strategies to the lesson plan” *(G05-MCP-P3, author’s translation)*. Overall, generalist AI mainly supported orientation but could also encourage partial delegation, whereas UDL-MCP more clearly

promoted justification, attention to learner variability and inclusive design agency. The complete set of representative excerpts is reported in Table 3.

Table 3: Main qualitative themes emerging from reflective logbooks

Theme	Evidence from logbooks	Interpretation
UDL theoretically known but difficult to operationalise	"We were already familiar with the general principles of UDL, but when we had to turn them into a complete lesson plan, we found it difficult to understand how to connect each principle with specific activities, materials and assessment methods." *(G01-NoAI-P1, author's translation)* "Knowing the UDL principles in theory was not enough. The most difficult part was deciding how to translate them into concrete choices for the lesson." *(G05-NoAI-P2, author's translation)* "We understood the importance of offering different ways of engaging and learning, but we were uncertain about how to organise all these elements coherently within the same lesson plan." *(G09-NoAI-P3, author's translation)*	The excerpts reveal a gap between theoretical familiarity with UDL and its translation into coherent design decisions.
Generalist AI as orientation support	"The AI helped us overcome the initial difficulty of starting. It gave us some ideas and a possible structure from which we could begin to develop the lesson." *(G04-GAI-P1, author's translation)* "At the beginning, we did not know how to organise the activity. The suggestions generated by the AI helped us put the different parts of the lesson in order." *(G08-GAI-P2, author's translation)* "Using the AI made the task seem less complex because it provided an initial outline and helped us identify possible activities and materials." *(G02-GAI-P3, author's translation)*	Generalist AI reduced initial uncertainty and provided ideas and structures that facilitated entry into the task.
Risk of accepting ready-made outputs	"The response generated by the AI was already very complete and well written, so we were initially inclined to accept it without making substantial changes." *(G06-GAI-P1, author's translation)* "Because the proposal appeared convincing, we did not immediately question whether all the suggested activities were really appropriate for our class." *(G10-GAI-P2, author's translation)* "In some cases, we relied on the AI-generated lesson structure without discussing every pedagogical choice in depth." *(G03-GAI-P3, author's translation)*	The fluency and completeness of the outputs could reduce critical revision and encourage cognitive offloading or partial delegation.
UDL-MCP as clearer design guide	"Compared with the generalist AI, UDL-MCP guided us more clearly through the different stages of the design process and helped us understand which UDL principle was connected to each choice." *(G08-MCP-P1, author's translation)* "The sequence proposed by UDL-MCP made the work more organised because it asked us to reflect step by step on objectives, barriers, activities and assessment." *(G01-MCP-P2, author's translation)* "The tool did not simply give us a ready-made lesson. It helped us justify our decisions by linking them to specific UDL criteria." *(G07-MCP-P3, author's translation)*	UDL-MCP made the design sequence and its relationship with UDL principles more explicit.
Strengthening of inclusive intentionality	"Using UDL-MCP led us to consider more carefully the different needs of the learners and the barriers that some students might encounter during the lesson." *(G11-MCP-P1, author's translation)* "We paid greater attention to learner variability and tried to explain why each option could make the activity more accessible and inclusive." *(G03-MCP-P2, author's translation)* "The guided questions encouraged us to reflect on the reasons behind our choices rather than simply adding inclusive strategies to the lesson plan." *(G05-MCP-P3, author's translation)*	UDL-MCP strengthened attention to learner variability, barriers and the justification of inclusive choices.

Note. The excerpts were originally written in Italian and translated into English by the author who conducted the qualitative analysis. Codes identify the group, experimental condition and study period. NoAI = no artificial intelligence; GAI = generalist artificial intelligence; MCP = UDL-MCP; P = period.

5. Discussion

5.1 The contribution

UDL-MCP is a UDL-aligned Model Context Protocol server that works as a semiotic mediator (Vygotsky, 1978). Its resources, tools, and prompts scaffold the justification of inclusive choices. The tool comes with a construct and a study design. The construct, Digital Affective Literacy, is the teacher's capacity to hold a relational, inclusive intentionality alongside digital competence (Immordino-Yang & Damasio, 2007; Lucangeli, 2019; Redecker, 2017). The design separates the product of teaching from its process. The product is the formal inclusiveness of the lesson, scored on a UDL 3.0 rubric (CAST, 2024; Meyer et al., 2014; Novak & Rodriguez, 2018). The process is design agency: who makes the inclusive choice and who can justify it. A lesson can read as inclusive while the teacher's agency stays weak. The product score captures the lesson and leaves the agency unmeasured. When the conditions produce similar lessons, the relevant difference sits in the process (Filippone et al., 2025a).

5.2 What the pilot shows

The quantitative findings provide no statistically detectable evidence that condition affected the inclusive quality of the lesson plans, $\chi^2(2) = 1.92$, $p = 0.38$. This null result does not establish equivalence among all three conditions. Under the exploratory ±0.5 bound, equivalence was supported only between UDL-MCP and the generalist condition, while both comparisons with NoAI remained inconclusive. The wide confidence intervals around the paired effect sizes, together with a minimum detectable effect of approximately $dz = 1.00$, show that small-to-moderate differences cannot be excluded. Between-group variation and the difference between judges were both larger than the

estimated condition contrasts. The quantitative results should therefore be read as preliminary and imprecise estimates from an underpowered pilot rather than as proof of no effect. Perceived usability constitutes a separate descriptive result: UDL-MCP obtained a mean SUS score of 72.0 (95% CI [68.1, 75.8]), but no comparative usability measure was collected for the other conditions.

5.3 From usability to reflective guidance: the role of UDL-MCP

The qualitative findings help interpret the apparent stability of the rubric scores across conditions. If only the final lesson plans are considered, the three conditions seem to produce broadly comparable levels of inclusive quality. However, the reflective logbooks suggest that similar products may correspond to different design processes. This distinction is central to the argument of the study: the pedagogical value of AI in inclusive design does not lie merely in improving the formal quality of the artefact, but in the extent to which it supports the teacher in remaining the author of the inclusive decision.

The generalist AI condition was experienced by trainees as useful, especially because it helped them approach a framework that they knew in theoretical terms but did not yet master operationally. This finding is important because it prevents a simplistic interpretation of cognitive delegation as an exclusively negative phenomenon. In the initial phases of a complex design task, AI can function as an orientation device, reducing uncertainty and providing access to examples, vocabulary and possible structures. This use of external support is not necessarily detrimental; as cognitive offloading theory suggests, external tools can free cognitive resources when they are used to support rather than replace the core learning operation (Risko & Gilbert, 2016).

At the same time, the logbooks show that the usefulness of AI depends on the architecture of the interaction. A generalist system tends to provide fluent and complete responses, which may reassure the user but also reduce the need to justify the pedagogical reasoning behind design choices. UDL-MCP, by contrast, appeared to guide the design process more explicitly. Its value was not simply that it helped trainees produce a lesson plan, but that it made the UDL logic of the lesson more visible. By connecting design decisions to guidelines, barriers, learner variability and inclusive options, the tool functioned as a semiotic mediator rather than as a substitute designer (Vygotsky, 1978).

This point strengthens the construct of Digital Affective Literacy. The affective dimension of AI-mediated design emerged in the trainees' reports of uncertainty, relief, clarity and guidance. The issue is not only whether teachers can use AI, but whether they can regulate the affective and cognitive pressure generated by a fluent artificial system. This point is consistent with research showing that emotion is not external to cognition, but deeply involved in learning, judgement and decision-making (Immordino-Yang & Damasio, 2007). When AI produces a plausible lesson plan, the temptation to accept the answer may be strong, especially for trainees who do not feel fully competent in the framework. Digital Affective Literacy therefore concerns the ability to use the tool without surrendering judgement, to transform the AI output into an object of reflection, and to preserve the inclusive intentionality of the design act.

The qualitative strand also clarifies why the UDL-MCP condition should not be evaluated only through immediate product scores. Its main contribution may lie in process quality: greater clarity, stronger procedural guidance, and a more explicit connection between design actions and inclusive principles. In this sense, UDL-MCP does not necessarily replace the teacher's effort; rather, it reorganises it. It reduces disorientation while maintaining the need for

pedagogical decision-making. This is precisely the difference between AI as a provider of answers and AI as a mediator of professional learning.

6. Conclusion: limits and validation

Two main limits bound these numbers. Power comes first: eleven paired groups detect only very large effects, so the product comparison cannot yet test small-to-medium differences. Measurement comes second, and it weighs more. Judges scored separate sets of lessons, only one lesson was scored by more than one judge, and its scores diverged sharply (1.25 to 2.61), so inter-rater agreement could not be estimated. The near-perfect internal consistency (α = 0.983) reflects single scoring and a halo effect. A reliable instrument requires several scores per lesson. The agenda follows. Future studies should score every lesson with several judges in order to estimate inter-rater reliability and reduce the rater confound. The number of groups should also be increased so that the design can detect small-to-medium effects comparable to those observed in the pilot.

Generalisability is further limited by the convenience sampling strategy and by recruitment within a single institutional context. The findings should therefore be interpreted as context-dependent exploratory evidence rather than as estimates that can be directly generalised to the wider population of pre-service teachers.

The qualitative strand should be extended and deepened through a broader corpus and additional forms of data triangulation. The reflective logbooks provide an initial operationalisation of Digital Affective Literacy (Immordino-Yang & Damasio, 2007; Lucangeli, 2019) within current research on generative AI in inclusive education and teacher formation (Barbieri et al., 2025; Sasanelli et al., 2025; Lecce & Di Tore, 2025). The contribution lies precisely in the

distinction between the formal quality of the lesson and the authorship of the inclusive process. Two lessons may appear equally inclusive even when, in one case, the teacher actively constructed and justified the inclusive choices, whereas in the other those choices were largely received from the artificial system. The product score treats these outcomes as equivalent; the qualitative strand examines whether the teacher remains the author of inclusion.

The qualitative strand also has limitations. The reflective logbooks constituted a situated and exploratory corpus of group-generated reflections, and the themes were developed through reflexive thematic analysis. The findings therefore provide interpretative evidence rather than generalisable estimates of thematic frequency. Moreover, because the logbooks were jointly produced by the groups, they represent negotiated collective accounts and do not allow the individual positions of the 100 trainees to be distinguished.

A further limitation concerns the analytic process, which was conducted by a single author. Within reflexive thematic analysis, researcher subjectivity is understood as an analytic resource rather than as an error to be eliminated. Nevertheless, the interpretation necessarily reflects the situated perspective of one researcher. Future studies could strengthen analytic reflexivity through peer debriefing, collaborative discussion of candidate themes, and a more explicit audit trail documenting the development of codes, themes and interpretations.

Future research should also broaden the qualitative corpus and integrate group logbooks with individual interviews, focus groups or stimulated-recall activities. Such triangulation would make it possible to distinguish more clearly between collectively negotiated accounts and individual experiences of uncertainty, delegation, guidance and inclusive agency. It would also support a more fine-grained examination of the relationship between textual indicators of

delegation and elaboration and the quantitative dimensions of the UDL rubric and SUS scores.

A further development should concern the integration of logbook analysis with more fine-grained indicators of delegation and elaboration. Previous research on AI-mediated problem-solving has shown that the distinction between autonomous elaboration and cognitive delegation can be explored through textual traces, such as explicit justification, acknowledgement of difficulty, reference to AI as support, or acceptance of AI as executor of the task (Filippone et al., 2025a). Examining similar textual patterns in UDL-based lesson planning would allow future studies to investigate not only whether the final lesson is inclusive, but also whether trainee teachers have actively constructed, justified and appropriated the inclusive rationale underlying the design.

References

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies, 4*(3), 114–123.

Barbieri, U., Carruba, M. C., Limone, P., & Di Fuccio, R. (2025). Is a chatbot enough? A pilot study on integrating UDL principles into intelligent tutoring systems. *QWERTY – Interdisciplinary Journal of Technology, Culture and Education, 21*(1), 148–168. https://doi.org/10.30557/QW000097

Bellantonio, S. (2025). L'intelligenza artificiale (IA) come «terza via» didattica e relazionale: Uno scenario per l'apprendimento. *Educrazia, 1*(2), 20–28.

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

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE Publications.

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

CAST. (2024). *Universal Design for Learning guidelines version 3.0*.

Compagno, G., Maniscalco, L., & Scolaro, I. (2025). UDL in multifaceted learning environments: Rethinking teachers' inclusive role and action. *Italian Journal of Health Education, Sport and Inclusive Didactics, 9*(1). https://doi.org/10.32043/gsd.v9i1.1405

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

Filippone, A., Barbieri, U., Marsico, E., Bevilacqua, A., De Carlo, M. E., & Di Fuccio, R. (2025). Can 3D virtual worlds be used as intelligent tutoring systems to innovate teaching and learning methods? Future challenges and possible scenarios for metaverse and artificial intelligence in education. *Engineering Proceedings, 87*(1), Article 110. https://doi.org/10.3390/engproc2025087110

Filippone, A., Di Fuccio, R., & Barbieri, U. (2025). Delegate or elaborate? A pilot study on the impact of generative AI on cognitive processes in problem-solving. *IUL Research, 6*(12), 57–81. https://doi.org/10.57568/iulresearch.v6i12.760

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Flick, U. (2022). *An introduction to qualitative research* (7th ed.). SAGE Publications.

Immordino-Yang, M. H., & Damasio, A. R. (2007). We feel, therefore we learn: The relevance of affective and social neuroscience to education. *Mind, Brain, and Education, 1*(1), 3–10. https://doi.org/10.1111/j.1751-228X.2007.00004.x

Lecce, A., & Di Tore, S. (2025). Digital skills of teachers in the age of artificial intelligence. *Italian Journal of Health Education, Sport and Inclusive Didactics, 9*(Suppl. 1). https://doi.org/10.32043/gsd.v9i2_Sup.1530

Lewis, J. R., & Sauro, J. (2018). Item benchmarks for the System Usability Scale. *Journal of Usability Studies, 13*(3), 158–167.

Lucangeli, D. (2019). *Cinque lezioni leggere sull'emozione di apprendere*. Erickson.

Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal design for learning: Theory and practice*. CAST Professional Publishing.

Novak, K., & Rodriguez, K. (2018). *UDL progression rubric*. CAST.

Redecker, C. (2017). *European framework for the digital competence of educators: DigCompEdu* (Y. Punie, Ed.). Publications Office of the European Union. https://doi.org/10.2760/178382

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Sasanelli, L. D., Barbieri, U., & Di Fuccio, R. (2025). Universal design for learning, accessible learning design and artificial intelligence: An exploratory study on pre-service teachers. *Italian Journal of Health Education, Sport and Inclusive Didactics, 9*(2). https://doi.org/10.32043/gsd.v9i2.1465

Schön, D. A. (2017). *The reflective practitioner: How professionals think in action*. Routledge. (Original work published 1983).

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

Zanazzi, S., Marcuccio, M., & Tassinari, M. E. (2025). L'intelligenza artificiale per l'inclusione: La prospettiva degli educatori in formazione. *Educrazia, 1*(2), 239–254.