

Impact of an Unplugged Intelligent Tutoring System on Flow and Learning Gains in Physics: A Quasi-Experimental Study in Low-Infrastructure Brazilian Public Schools

João Carlos Herculano da Silva Queiroz¹, Alex Almeida do Amaral¹,
Huliane Medeiros da Silva², Glaydson Francisco Barros de Oliveira²,
Geiser Chalco Chalco²

¹Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Universidade Federal Rural do Semi-Árido (UFERSA)
Universidade do Estado do Rio Grande do Norte (UERN) – Mossoró – RN – Brasil

²Universidade Federal Rural do Semi-Árido (UFERSA) – RN – Brasil

{joao.queiroz72216, alex.amaral}@alunos.ufersa.edu.br

{huliane, glaydson.barros, geiser}@ufersa.edu.br

Abstract. *This paper presents the design and quasi-experimental evaluation of STI-Física, an unplugged Intelligent Tutoring System (ITS) for Physics education in low-infrastructure Brazilian public schools. The system integrates local Bayesian Knowledge Tracing (BKT) for adaptive student modeling, generative-AI step-level hints precomputed for offline use, QR-based session access without student registration, and deferred synchronization with a teacher analytics dashboard. All of these components operate on students' personal smartphones without internet dependency. We evaluated the system in a public high school in the semi-arid Northeast region of Brazil with 31 second-year students for flow analysis (control group [CG]: $n = 18$; experimental group [EG]: $n = 13$) and 29 for learning analysis (CG: $n = 13$; EG: $n = 16$) on a thermodynamics unit. The EG used STI-Física in 90-minute sessions while the CG performed equivalent paper-based exercises. The 10-item learning instrument was built from a BNCC-aligned content matrix with teacher validation. The EG maintained significantly higher flow states than the CG (ANCOVA, $F(1, 28) = 4.417$, $p = .045$, $\eta^2 = .127$) and showed greater learning gains (Mann-Whitney, $U = 57.0$, $p = .041$, $r_{rb} = .452$). These findings provide preliminary, proof-of-concept evidence that unplugged ITS architectures combining BKT with generative-AI hints can support science education in resource-constrained settings.*

Keywords: Intelligent Tutoring Systems · Unplugged AIED · Bayesian Knowledge Tracing · Physics Education · Flow · Low-Infrastructure Schools

Resumo. *Este artigo apresenta o projeto e a avaliação quasi-experimental do STI-Física, um Sistema Tutor Inteligente (STI) unplugged para o ensino de Física em escolas públicas brasileiras de baixa infraestrutura. O sistema integra Bayesian Knowledge Tracing (BKT) local para modelagem adaptativa do estudante, hints de nível de passo gerados por IA generativa e pré-computados para uso offline, acesso a sessões via QR sem cadastro e sincronização diferida*

com um dashboard analítico para o professor. Todos esses componentes operam em smartphones pessoais dos estudantes sem dependência de Internet. Avaliamos o sistema em uma escola pública de ensino médio no semiárido nordestino com 31 estudantes do 2º ano para análise de fluxo (grupo controle [GC]: $n = 18$; experimental [GE]: $n = 13$) e 29 para análise de aprendizagem (GC: $n = 13$; GE: $n = 16$) em uma unidade de termodinâmica. O GE manteve estados de fluxo significativamente mais elevados que o GC (ANCOVA, $F(1, 28) = 4,417$, $p = .045$, $\eta^2 = .127$) e apresentou maiores ganhos de aprendizagem (Mann-Whitney, $U = 57,0$, $p = .041$, $r_{rb} = .452$). Estes resultados fornecem evidência empírica preliminar, de prova de conceito, de que arquiteturas unplugged de STI combinando BKT com hints de IA generativa podem apoiar o ensino de ciências em contextos de recursos restritos.

1. Introduction

Physics education in Brazilian public schools faces persistent structural and pedagogical challenges, reflected in low scientific proficiency and pronounced educational inequality. According to the OECD, only 45% of Brazilian 15-year-olds reached at least Level 2 proficiency in science, compared with 76% across OECD countries [OECD 2023]. Data from the Brazilian School Census [INEP 2023] reveal that fewer than 30% of public schools have science laboratories in usable condition, and these disparities are exacerbated in underserved areas such as the semi-arid Northeast, where infrastructure is limited, laboratory resources are scarce, and student profiles are highly heterogeneous.

Intelligent Tutoring Systems (ITS) are a promising approach to address learning gaps in Science and Mathematics. Several decades of research, consolidated in the meta-analysis by Ma et al. [Ma et al. 2014], indicate that ITS improve learning outcomes by providing one-to-one tutoring, immediate feedback, step-by-step guidance, progress monitoring, and adaptive content sequencing. ITS therefore offer a path to mitigate persistent instructional gaps in Physics education in Brazil, particularly where students would benefit from continuous practice and personalized scaffolding.

However, ITS have historically been developed for well-resourced educational settings, and barriers such as shared hardware, unstable electricity, limited connectivity, and uneven digital skills constrain their adoption in underserved communities [Nye 2015]. Empirical evidence on ITS impact in low-infrastructure schools remains scarce, and the few existing systems either operate at the modeling layer in isolation or lack contextualized evaluations in real classrooms. This gap motivates the present study: we design and implement an unplugged ITS for Physics instruction and evaluate its impact in a low-resource school. Specifically, we test two hypotheses: (H1) the use of an unplugged ITS promotes greater engagement (operationalized as flow) than traditional instruction; and (H2) students using the unplugged ITS demonstrate superior learning gains.

The contributions of this paper are threefold: (i) the design of STI-Física, an offline-first ITS architecture that integrates local Bayesian Knowledge Tracing with generative-AI step-level hints precomputed for offline use, frictionless QR-based session access, and deferred synchronization with a teacher analytics dashboard; (ii) empirical evidence on flow and learning outcomes from a quasi-experiment in a public high school in the Brazilian semi-arid Northeast; and (iii) a discussion of design and methodological

implications for AIED Unplugged research.

2. Background and Related Work

2.1. Intelligent Tutoring Systems and the Tutoring Loop

Technology-mediated education has progressively moved beyond static instructional software toward systems that adapt to the individual learner. Intelligent Tutoring Systems (ITS) are computational environments that apply Artificial Intelligence to approximate the guidance of a human tutor, diagnosing learner needs in real time and adjusting content, difficulty, and support accordingly. The classical account by Nwana [Nwana 1990] structures this adaptivity around three knowledge modules (domain, student, and pedagogical knowledge), whose integration distinguishes an ITS from a conventional digital exercise list that offers only binary, static feedback.

Operationally, ITS behavior is commonly described through two cyclical processes [VanLehn 2006]: the outer loop, which selects the next problem or topic from the learner's overall performance, and the inner loop, which manages interaction step by step within a single problem. The inner loop is the engine of immediate tutoring: at each student action it interprets the action against domain knowledge, updates the student model, and selects a tutorial response. VanLehn [VanLehn 2006] organizes inner-loop responses along a gradient of increasing support: validation feedback (confirming step correctness), specific corrective feedback (identifying what is wrong, often via error models or constraints [Mitrovic 2003]), hint-based tutoring (progressive guidance that balances guided discovery against frustration [Aleven et al. 2016]), probing and student-model updating (probabilistically refreshing estimated mastery at each step [Corbett and Anderson 1995]), and metacognitive review (structured reflection after problem completion [Roll et al. 2011]). Personalization across both loops is driven by Knowledge Tracing, the probabilistic algorithm introduced by Corbett and Anderson [Corbett and Anderson 1995] that continuously updates the estimated mastery of each skill or Knowledge Component (KC).

Realizing the inner loop offline. Established Cognitive Tutors realize the inner loop through real-time tracing against a hand-authored production-rule model [Anderson et al. 1995], which presupposes the computational budget and authoring tooling of well-resourced settings. STI-Física instead realizes the same functions under an offline-first constraint by pairing a local Bayesian Knowledge Tracing (BKT) student model [Corbett and Anderson 1995] that updates per-KC mastery on the device with generative-AI step-level hints precomputed before class and cached locally—reproducing immediate feedback, progressive hints, and continuous mastery estimation without runtime rule tracing, server connectivity, or institutional logins, consistent with data-driven ITS development [Koedinger et al. 2013].

2.2. AIED Unplugged and ITS in Low-Infrastructure Contexts

The notion of AIED Unplugged [Isotani et al. 2023] has emerged to make AI in Education accessible where infrastructure and Internet access are limited. Within this paradigm, Rodrigues et al. [Rodrigues et al. 2024] validated Knowledge Tracing models for disconnected, low-cost mobile devices in numeracy education (grades K-2 to K-5), showing

that both classical BKT and lightweight Deep Knowledge Tracing variants retain acceptable predictive performance offline. That work establishes the technical feasibility of on-device student modeling, but it targets numeracy with elementary students and validates KT models in isolation, without embedding them within a complete tutoring system that delivers real-time adaptive feedback and step-level guidance.

Established ITS for Physics such as Andes [VanLehn et al. 2005] and Cognitive Tutors [Anderson et al. 1995] have produced significant learning gains but were designed for well-resourced environments dependent on desktop computers and stable networks. In Brazil, Bittencourt et al. [Bittencourt et al. 2011] proposed a generic ITS architecture emphasizing accessibility and adaptivity, and Jaques et al. [Jaques et al. 2013] developed PAT2Math, a step-by-step ITS for algebraic equation solving; however, these systems were designed for conventional connected settings rather than the offline, low-infrastructure conditions targeted here. More broadly, Nye [Nye 2015] reviewed ITS in developing-world contexts and documented hardware, electricity, and connectivity barriers as recurring obstacles. Research on digital technologies in challenging Brazilian contexts similarly stresses critical pedagogical integration over mere access [Kenski 2012]; studies of low-cost tools such as PhET simulators in Brazilian basic education report gains in interest and conceptual understanding [Araújo et al. 2021], but these and other educational-software approaches [Fialho and Matos 2010] remain generic and non-adaptive rather than systems that personalize instruction to the heterogeneous prior knowledge typical of Brazilian classrooms.

2.3. Positioning of STI-Física

Table 1 positions our work relative to representative prior systems along five dimensions of the AIED Unplugged paradigm. To the best of our knowledge, STI-Física is the first complete ITS for Physics that combines on-device BKT with precomputed generative-AI step-level hints, frictionless QR-based session access, and an offline-synced teacher analytics dashboard, all operating without internet dependency and validated in a Brazilian public school in an underserved region.

Table 1. Positioning of STI-Física relative to representative prior ITS work along five dimensions of the AIED Unplugged paradigm.

System	Local BKT	GenAI Hints Offline	No Login	Offline Dashboard	Validated in Underserved K-12
Andes	–	–	–	–	–
Cognitive Tutor	–	–	–	–	–
Bittencourt et al.	–	–	–	–	–
PAT2Math	–	–	–	–	–
Rodrigues et al.	✓	–	n/a	n/a	numeracy only
STI-Física (this work)	✓	✓	✓	✓	✓ (Physics)

3. STI-Física: Unplugged Intelligent Tutoring System

We developed STI-Física as a native mobile application with an offline-first architecture explicitly designed to address the intermittent connectivity typical of rural and peripheral Brazilian schools. Following the AIED Unplugged paradigm [Isotani et al. 2023],

the system operationalizes four main features, whose deployed interfaces are shown in Figure 1.

(a) Offline content provisioning. Tutor packages, including lessons, exercises, and all required assets, are downloaded and cached locally on student and teacher devices prior to classroom use. The platform predicts the most likely tutor packages and exercise sequences a student will encounter and pre-downloads these resources, ensuring uninterrupted offline access during class.

(b) Frictionless session access. Students join learning sessions without account registration or online authentication. Teachers generate a session-specific QR code (or short numeric token); students scan the QR code to access the session offline, removing the common barriers of login credentials, password recovery, and internet dependency. Each session establishes an ephemeral local identifier that links subsequent interactions to a student record on the device, which is later reconciled with the teacher's dashboard during deferred synchronization.

(c) Offline adaptive support and student modeling. Students receive step-level hints generated in advance using a generative-AI pipeline, tailored to their profile, prior errors, and estimated knowledge states. Before each session, the system simulates plausible interaction paths and precomputes hint sequences stored locally. In parallel, a local Bayesian Knowledge Tracing (BKT) implementation [Corbett and Anderson 1995] continuously updates mastery probabilities for each KC, reflected in a real-time skill meter that supports metacognitive awareness. The combination of BKT-driven adaptive sequencing and precomputed hints allows the inner loop of tutoring [VanLehn 2006] to operate fully offline.

Because generative models can introduce conceptual errors, every precomputed hint underwent a manual curation protocol before being packaged into a tutor unit. Two practicing Physics teachers and the authors reviewed each hint against four criteria: (i) *physical and conceptual correctness*, screening for generative-AI hallucinations such as incorrect formulas, invalid unit reasoning, or unphysical intermediate steps; (ii) *pedagogical and linguistic adequacy*, ensuring vocabulary and framing calibrated to the high-school (BNCC) level; (iii) *alignment with the target Knowledge Component and with the progressive, Socratic hint structure*, so that each hint scaffolds reasoning without prematurely revealing the final answer; and (iv) *terminological and notational consistency* with the lesson content. Hints failing any criterion were rewritten or discarded, and only approved hints were cached for offline delivery.

(d) Deferred synchronization and teacher analytics. After each session, a queue-based synchronization service transmits interaction data to the server when connectivity is restored (e.g., when the device reaches a Wi-Fi point at the teacher's home). The teacher dashboard provides class-level and individual diagnostic reports, including completion rates, step-level analytics, and KC mastery summaries, enabling formative assessment and targeted intervention.

All instructional intelligence (BKT, hint selection, mastery estimation) executes on the student's device during class. The network is only required later, when the device reaches a connectivity point to upload its queued interaction logs.

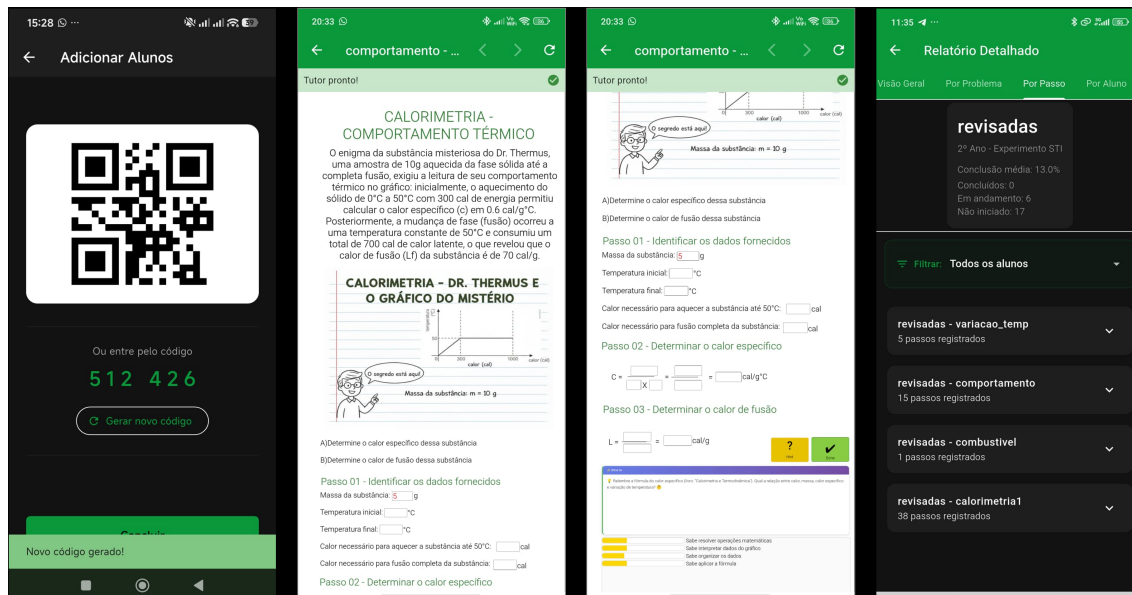


Figure 1. STI-Física interfaces during the unplugged deployment (thermodynamics unit). From left to right: (a) frictionless session access through a teacher-generated QR code or short numeric token, with no student registration; (b) cached offline lesson content on calorimetry; (c) step-by-step problem solving with immediate feedback and precomputed generative-AI hints; and (d) the teacher analytics dashboard (Relatório Detalhado) summarizing per-student, per-Knowledge-Component progress after deferred synchronization.

4. Methodology

4.1. Design and Participants

We adopted a quasi-experimental design with non-equivalent, non-randomized control (CG) and experimental (EG) groups, assigned by convenience based on pre-existing classes. The study was conducted in a public state high school in a small municipality in the semi-arid Northeast of Brazil, characterized by limited school infrastructure (no functional science laboratory; shared computer room with intermittent connectivity).

Participants were 2nd-year high school students (ages 16–18). The final analytical sample comprised 31 students for the flow analysis (CG: $n = 18$; EG: $n = 13$) and 29 students for the learning analysis (CG: $n = 13$; EG: $n = 16$). The asymmetry between analyses reflects per-instrument missingness (students absent on the day a specific instrument was administered), a source of potential selection bias that we revisit in Section 8. Group comparability was verified for gender ($\chi^2(1) = 0.267$, $p = .605$, Cramér's $V = 0.093$) and age (CG: $M = 16.83$; EG: $M = 16.92$).

4.2. Instructional Domain

The intervention targeted a thermodynamics unit on calorimetry and thermal behavior, covering specific heat capacity, latent heat, and phase changes. Knowledge Components

(KCs) were identified jointly with two practicing Physics teachers and aligned with the Brazilian National Common Curricular Base (BNCC) descriptors for high school Physics.

4.3. Instruments

Flow. Flow was measured via the Dispositional Flow Scale (DFS, administered pre-activity) and the Flow State Scale (FSS, administered post-activity), adapted from Csikszentmihalyi [Csikszentmihalyi 1990]. Each instrument contains 9 Likert items (1–5). Scores were computed as item means. The derived variable $\text{DiffFlow} = \text{FSS} - \text{DFS}$ captures the change in flow between baseline disposition and the in-task experience.

Learning. Learning was assessed via a 10-item objective pre/post knowledge test on thermodynamics, administered on paper to both groups. The instrument was designed jointly with two Physics teachers using a content-by-objective specification matrix derived from BNCC descriptors. Face validity was reviewed by two additional teachers external to the test construction team. The gain variable was computed as $\text{Gain} = \text{Post} - \text{Pre}$.

4.4. Procedures

The intervention lasted four weeks. In the first session, both groups completed the knowledge pre-test and the DFS questionnaire. In subsequent sessions, the CG performed traditional paper-based exercises on equivalent thermodynamics problems, while the EG accessed STI-Física on personal mobile devices in 90-minute sessions to solve tutorial problems with step-by-step BKT-guided support and precomputed hints. Both groups received the same total instructional time. At the end of the intervention, both groups completed the knowledge post-test and the FSS questionnaire.

4.5. Data Analysis

We followed three stages of analysis. First, group comparability was verified using χ^2 tests (categorical) and independent-samples t -tests or Mann-Whitney U tests (continuous) on baseline measures. Second, parametric assumptions were evaluated: normality via Shapiro-Wilk and homogeneity of variances via Brown-Forsythe (Levene's test for ANCOVA). Third, the two hypotheses were tested with Student's t -tests, Mann-Whitney U tests, and ANCOVAs (controlling DFS for flow and the pre-test for learning). To complement the frequentist analyses, we additionally conducted Bayesian independent-samples t -tests (default Cauchy prior, $r = 0.707$) and report Bayes Factors (BF_{10}) interpreted under the Jeffreys evidence classification. All analyses were conducted in JASP 0.95.4 with $\alpha = .05$. Throughout, descriptive statistics are reported as M (mean), SD (standard deviation), and Mdn (median), with N denoting the number of participants.

5. Validity of the Learning Instrument

To support the interpretation of learning-gain results, we document the validity basis of the 10-item thermodynamics test. As described in Section 4, content validity was maximized through a BNCC-derived content-by-objective specification matrix, joint item construction with two practicing Physics teachers, and a face-validity review by two additional teachers not involved in item writing. Each item was mapped to a single target Knowledge Component, and the pre- and post-test forms were matched in content and difficulty.

We note a measurement limitation and clarify its cause. The pre- and post-tests were administered on paper to both groups—not through the STI-Física application—and, when the data were entered for analysis, only each student’s total score was recorded, with the item-by-item responses neither digitized nor retained. The limitation is therefore one of data collection and record-keeping, not of the system’s architecture. As a consequence, internal-consistency reliability (e.g., KR-20) and item-level difficulty and discrimination indices could not be computed, and the learning results should be read in light of this constraint. Future administrations will preserve item-level responses to enable full psychometric reporting.

6. Results

Results are organized along two axes: (i) flow during the intervention; (ii) conceptual learning gain.

6.1. Flow Analysis

Table 2 presents descriptive statistics for the flow variables. Both groups showed similar baseline flow disposition (DFS), reinforcing pre-intervention comparability. However, the in-task flow experience (FSS) was higher in the EG ($M = 3.513$) than in the CG ($M = 3.222$). DifFlow was negative in both groups, indicating a general decrease in flow during the activity; this decrease was considerably smaller in the EG ($M = -0.453$) than in the CG ($M = -0.914$), suggesting that STI-Física supported flow maintenance.

Table 2. Descriptive statistics for flow variables by group.

Variable	Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	Range
DFS	CG	18	4.136	0.363	4.111	3.56–4.89
	EG	13	3.966	0.583	4.111	3.11–4.89
FSS	CG	18	3.222	0.464	3.444	2.11–4.00
	EG	13	3.513	0.405	3.556	2.78–4.11
DifFlow	CG	18	−0.914	0.509	−0.833	−1.67 to −0.11
	EG	13	−0.453	0.594	−0.444	−1.44 to 0.44

Note. *N* = number of participants; *M* = mean; *SD* = standard deviation; *Mdn* = median; Range = minimum–maximum.

Normality (Shapiro-Wilk, $p > .05$) and homogeneity of variances (Brown-Forsythe, $F(1, 29) = 0.035$, $p = .853$) assumptions were met. The independent-samples *t*-test revealed a statistically significant difference between groups in DifFlow ($t(29) = -2.318$, $p = .028$), corroborated by the Mann-Whitney test ($U = 64.5$, $p = .037$).

To verify that the in-task flow difference was independent of initial disposition, we conducted an ANCOVA with FSS as the dependent variable, Group as factor, and DFS as covariate (Levene, $F(1, 29) = 0.684$, $p = .415$, satisfied). The ANCOVA confirmed a significant Group effect on FSS ($F(1, 28) = 4.417$, $p = .045$, $\eta^2 = .127$), indicating that approximately 13% of the variance in in-task flow is explained by group membership after controlling for initial disposition. The DFS covariate was non-significant

($p = .139$), confirming that the EG advantage reflects the intervention rather than prior dispositional differences. As a Bayesian counterpart, an independent-samples t -test on DiffFlow (JZS prior, $r = 0.707$) yielded $BF_{10} = 2.40$, which constitutes anecdotal evidence for the group difference under the Jeffreys classification; it is directionally consistent with, though weaker than, the significant frequentist t -test and ANCOVA.

6.2. Learning Gain Analysis

Table 3 summarizes test performance. Both groups started from similarly low pre-test scores, reinforcing initial equivalence. On the post-test, the EG showed a mean nearly twice that of the CG ($M_{EG} = 2.031$ vs. $M_{CG} = 1.115$). The mean learning gain of the EG ($M = 1.063$) was substantially greater than that of the CG ($M = 0.031$), whose near-zero mean indicates effectively no improvement.

Table 3. Descriptive statistics of Pre-Test, Post-Test, and Gain by group.

Variable	Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>Mdn</i>	Range
Pre-Test	CG	13	1.085	1.415	0.300	0–4.5
	EG	16	0.969	1.161	0.500	0–4.0
Post-Test	CG	13	1.115	2.152	0.500	0–8.0
	EG	16	2.031	1.396	2.000	0–5.0
Gain	CG	13	0.031	1.464	0.000	–2.9 to 3.5
	EG	16	1.063	1.430	1.125	–2.0 to 3.0

Note. *N* = number of participants; *M* = mean; *SD* = standard deviation; *Mdn* = median; Range = minimum–maximum.

The gain variable showed approximately normal distribution (Shapiro-Wilk, $p > .05$) and homogeneous variances (Brown-Forsythe, $F(1, 27) = 0.252$, $p = .620$). The t -test indicated a marginally significant difference ($t(27) = -1.912$, $p = .067$, $d = 0.714$, 95% CI $[-1.46, 0.05]$), while the Mann-Whitney test, less sensitive to normality violations, indicated a significant difference ($U = 57.0$, $p = .041$, $r_{rb} = 0.452$, 95% CI $[0.06, 0.72]$).

To isolate the group effect controlling for initial performance, we conducted an ANCOVA with Post-Test as dependent variable, Group as factor, and Pre-Test as covariate (Levene, $p = .633$, satisfied). The Pre-Test was a significant predictor of Post-Test ($F(1, 26) = 14.102$, $p < .001$, $\eta^2 = .324$). The Group effect showed a trend toward significance ($F(1, 26) = 3.459$, $p = .074$, $\eta^2 = .079$), indicating an 8% group-attributable variance in post-test performance after controlling for prior knowledge. As a Bayesian counterpart, an independent-samples t -test on the learning gain (JZS prior, $r = 0.707$) yielded $BF_{10} = 1.32$, that is, only anecdotal and essentially inconclusive evidence. We therefore anchor the learning claim on the significant Mann-Whitney result and the moderate-to-large effect size ($d = 0.714$), interpreting the marginal ANCOVA p -value and the weak Bayes factor as symptomatic of the limited statistical power of an $N = 29$ sample rather than evidence for a null effect; replication with larger samples is required to settle H2.

7. Discussion

This study provides preliminary, proof-of-concept evidence that an offline-capable, mobile ITS based on local BKT and precomputed generative-AI hints can positively impact both engagement (flow) and learning outcomes in Physics in a low-infrastructure Brazilian public school.

The flow effect is consistent across analyses. The three frequentist tests (Student's t , Mann-Whitney, and ANCOVA) converged, and the Bayesian counterpart pointed the same way, albeit only anecdotally ($BF_{10} = 2.40$). The ANCOVA confirmed that the EG advantage persists after controlling for initial dispositional flow ($\eta^2 = .127$), supporting H1 and aligning with predictions that adaptive feedback and step-level scaffolding sustain the challenge–skill balance central to flow [Csikszentmihalyi 1990, VanLehn 2006]. Its immediate feedback, progressive hints, and real-time skill meter appear to keep students within the productive zone of difficulty.

Learning evidence points in the same direction but is more nuanced. As detailed in Section 6.2, it is the significant non-parametric test and the moderate-to-large effect size, rather than the inconclusive Bayes factor, that carry the claim. The near-zero CG gain ($M = 0.031$) versus the meaningful EG gain ($M = 1.063$) suggests a real benefit whose marginal frequentist significance reflects the limited power of the sample rather than the absence of an effect. We therefore read the learning result as preliminary, proof-of-concept evidence pending replication, consistent with the content validity of the test documented in Section 5.

Regarding the possible novelty effect, some contextualization is warranted. Although personal smartphone ownership is widespread among these students, the devices were not routinely used for structured pedagogical activities in this school prior to the intervention; classroom use was typically informal or restricted rather than part of an established didactic routine. The intervention therefore introduced a relatively novel *didactic* use of an otherwise familiar device, so an initial engagement boost driven by novelty cannot be fully excluded and may have contributed to the observed flow advantage. Future studies will measure students' prior exposure to technology-mediated instruction explicitly, so that novelty can be modeled rather than merely acknowledged.

Architectural and equity implications. The offline-first architecture proved viable on students' personal smartphones, showing that adaptive AI-driven instruction need not depend on desktop computers, stable internet, or institutional logins—a relevant result for Brazilian public schools, where missing laboratories and unreliable connectivity are recurring barriers to experimental science education. By integrating local BKT with precomputed generative-AI hints, STI-Física extends the AIED Unplugged paradigm [Isotani et al. 2023, Rodrigues et al. 2024] from isolated student modeling to a complete tutoring loop running fully offline.

8. Limitations

We highlight six limitations that should temper the interpretation of these results. First, the quasi-experimental design with non-random group assignment limits causal inferences. Although the groups were comparable on gender, age, and pre-test performance, and both

received the same total instructional time, unmeasured factors such as teacher effects and student motivation were not directly controlled; the differential per-instrument attrition reported in Section 4 is a further potential source of selection bias. Second, the sample size ($N = 29\text{--}31$) restricts statistical power and generalizability; larger replications are needed, though our Bayesian estimates partly mitigate this by quantifying evidence rather than relying on dichotomous significance. Third, the four-week intervention does not allow assessing long-term retention or sustained behavioral effects. Fourth, we cannot fully disentangle the adaptive components (BKT, hints) from a possible novelty effect of the mobile app; an additional non-adaptive-app control arm, planned for future work, would isolate this contribution. Fifth, in this deployment we did not obtain a sufficiently granular interaction log from the application to analyze hint-usage patterns, per-KC mastery trajectories, or time-on-task; the logging-and-synchronization layer needed to make these data reliably available is still under implementation. Sixth, because only total pre/post scores were retained, internal-consistency reliability (KR-20) and item-level statistics could not be computed; we relied on the content-validity procedures of Section 5 and will preserve item-level responses in future administrations.

9. Conclusions and Future Work

We presented STI-Física, an unplugged ITS for Physics education in low-infrastructure Brazilian public schools, and reported quasi-experimental evidence that its use is associated with significantly higher flow states and with greater learning gains on a thermodynamics unit, compared to traditional paper-based instruction. Given the design and sample constraints, we frame these results as preliminary, proof-of-concept evidence rather than confirmatory causal findings. The architectural contribution shows that the complete ITS inner loop can run offline on students' personal smartphones.

Future work will: (i) replicate the study with larger, randomized samples and an additional non-adaptive mobile control arm to isolate the contribution of adaptivity; (ii) integrate the planned interaction-logging layer to enable BKT mastery-trajectory analysis, hint-usage analytics, and time-on-task investigation; (iii) extend the system to other Physics topics and conduct longitudinal retention assessments; and (iv) investigate teacher dashboard usability and its effect on formative pedagogical decision-making.

Ethical Considerations

This research was conducted in accordance with the guidelines of Brazilian National Health Council Resolution CNS No. 466/2012 and was approved by the Research Ethics Committee (CEP) of the Universidade do Estado do Rio Grande do Norte (UERN), under opinion No. 7.748.570 (CAAE No. 90618025.9.0000.5294). All participants—and the legal guardians of those under 18—signed the Free and Informed Consent Form (TCLE) before data collection began, ensuring the anonymity and confidentiality of the information.

Declaration on the Use of Generative AI

In accordance with the SBC Code of Conduct for authors, the authors declare two uses of generative AI in this work. First, as a technical component of the system: STI-Física uses a generative-AI pipeline to precompute step-level hints offline (Section 3), which the

authors review and curate before packaging into tutor units. Second, as a writing-support tool for language editing, structural review, and translation assistance during manuscript preparation. Generative AI was not used to produce empirical data, statistical analyses, or scientific claims. All content was reviewed and validated by the authors, who take full responsibility for the integrity of the work, including any potential plagiarism.

Artifact Availability

To support reproducibility, the anonymized dataset (DFS/FSS responses, pre/post-test scores, group assignment, gender, age), the statistical analysis files (JASP), the experimental materials (knowledge test, flow questionnaires), and the STI-Física source code (BKT module and precomputed-hints pipeline) are available from the authors upon reasonable request; a public, openly licensed repository will accompany the camera-ready version.

Acknowledgments

This work was carried out with financial support from Fundação Itaú, under the 2024 *IA para Educação* (AI for Education) call, through the project “Sistema Tutor Inteligente Unplugged com IA Generativa para o Ensino de Física”.

References

- Aleven, V., Roll, I., McLaren, B. M., and Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 26(1):205–223.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2):167–207.
- Araújo, E. S., Nascimento, J. L. B., Silva, J. C., and Andrade, C. F. (2021). O uso de simuladores virtuais educacionais e as possibilidades do PhET para a aprendizagem de física no ensino fundamental. *Revista de Ensino de Ciências e Matemática (RENCIMA)*, 12(3):1–25.
- Bittencourt, I. I., Costa, E., Silva, M., and Soares, E. (2011). Uma arquitetura para sistemas tutores inteligentes com foco na acessibilidade e na adaptatividade. *Revista Brasileira de Informática na Educação*, 19(1):45–58.
- Corbett, A. T. and Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.
- Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. Harper & Row, New York.
- Fialho, N. N. and Matos, E. L. M. (2010). A arte de envolver o aluno na aprendizagem de ciências utilizando softwares educacionais. *Educar em Revista (n. esp. 2)*, pages 121–136.
- INEP (2023). Censo escolar da educação básica 2023. Technical report, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

- Isotani, S., Bittencourt, I. I., Chalco, G. C., Dermeval, D., and Mello, R. F. (2023). AIED unplugged: Leapfrogging the digital divide to reach the underserved. In *Artificial Intelligence in Education (AIED 2023)*, volume 1831 of *CCIS*, pages 772–779. Springer.
- Jaques, P. A., Seffrin, H., Rubi, G., de Moraes, F., Ghilardi, C., Bittencourt, I. I., and Isotani, S. (2013). Rule-based expert systems to support step-by-step guidance in algebraic problem solving: The case of the tutor PAT2Math. *Expert Systems with Applications*, 40(14):5456–5465.
- Kenski, V. M. (2012). *Educação e Tecnologias: o Novo Ritmo da Informação*. Papirus, Campinas, 8 edition.
- Koedinger, K. R., Brunskill, E., Baker, R. S. J. d., McLaughlin, E. A., and Stamper, J. (2013). New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41.
- Ma, W., Adesope, O. O., Nesbit, J. C., and Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4):901–918.
- Mitrovic, A. (2003). An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education*, 13(2-4):173–197.
- Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review*, 4(4):251–277.
- Nye, B. D. (2015). Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context. *International Journal of Artificial Intelligence in Education*, 25(2):177–203.
- OECD (2023). PISA 2022 results: Factsheets – brazil. Technical report, OECD. Available at [oecd.org](https://www.oecd.org).
- Rodrigues, L., Avila-Santos, A. P., Silva, T. E., da Penha, R. S., Neto, C., Chalco, G., dos Santos, E. L., Souza, E., Guerino, G., Vieira, T., Marinho, M., Macario, V., Bittencourt, I. I., Dermeval, D., and Isotani, S. (2024). Knowledge tracing unplugged: From data collection to model deployment. In *Artificial Intelligence in Education (AIED 2024)*, volume 14829 of *LNAI*, pages 91–104. Springer.
- Roll, I., Aleven, V., McLaren, B. M., and Koedinger, K. R. (2011). Improving students’ help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2):267–280.
- VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3):227–265.
- VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., and Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, 15(3):147–204.