Balochistan Winter School Programme

Research Context Brief

For Georgetown University / World Bank Research Partnership

Programme: FLN (Foundational Literacy & Numeracy) Winter School

Location: Balochistan Province, Pakistan

Report Date: February 2026 (Updated)

Data Period: January 5 – February 4, 2026

Prepared by: Taleemabad (Programme Partner)

Executive Summary

This brief provides research context for the Balochistan Winter School Programme, combining field observations with quantitative analysis of classroom practices. The dataset has grown substantially since the initial January 2026 brief, now incorporating 2,217 AI observations and 283 coach observations across 5 weeks of programme implementation.

The findings offer critical context for interpreting baseline and endline student assessments, revealing both the strengths and limitations of classroom instruction in a low-resource, post-conflict educational setting.

Programme Scale at a Glance

Metric	Value
Schools	17
Teachers	87
Students	2,247
Coaches	19
AI Observations (completed)	1,773
Coach Observations	283
School Visits	83
Lesson Plans Analyzed	191
Teachers with Both AI + Coach	68
Programme Duration	5 weeks (Jan 5 – Feb 4, 2026)
Subjects	English Literacy, Urdu Literacy, Numeracy
Grades	2, 3, 4, 5

Key Classroom Metrics

Metric	Value
Teacher Talk Time	85.3%
Student Talk Time	7.1%
Open-Ended Questions	14.8%
Closed-Ended Questions	85.2%
Student-Generated Questions (AI-assessed)	4%
Avg Questions per Lesson	17.5
Avg Class Size (Coach-assessed)	6.5 students

1. Programme Design

The Winter School Model

Element	Detail
Programme	FLN (Foundational Literacy & Numeracy) Winter School
Location	Balochistan province, Pakistan
Duration	Winter vacation period (5 weeks)
Training	One-week intensive teacher onboarding
Support	Hybrid AI + human coaching model
Avg Class Size	6.5 students (coach-reported)

Why Winter School Context Matters

Teachers reported that FLN programmes are more effective during vacation periods due to:

— Reduced curriculum pressure: Protected time for foundational skills

— Smaller class sizes: More individualized attention possible (avg 6.5 vs 50+ in regular sessions)

— Focused instruction: FLN content without competing demands

This context is important for interpreting student outcomes, as main academic sessions present different constraints.

Observation Methodology

AI-Powered Observations (n = 1,773 completed):

— Analyze audio recordings from classroom instruction

— Provide scores across 30 indicators (Sections B, C, D)

— Capture quantitative metrics (talk time, question types, group work)

— Enable high-frequency feedback (every lesson)

Human Coach Observations (n = 283):

— In-person classroom observations by trained coaches

— Focus on FLN programme fidelity indicators (~15 items)

— Capture student level distribution (beginner/intermediate/advanced)

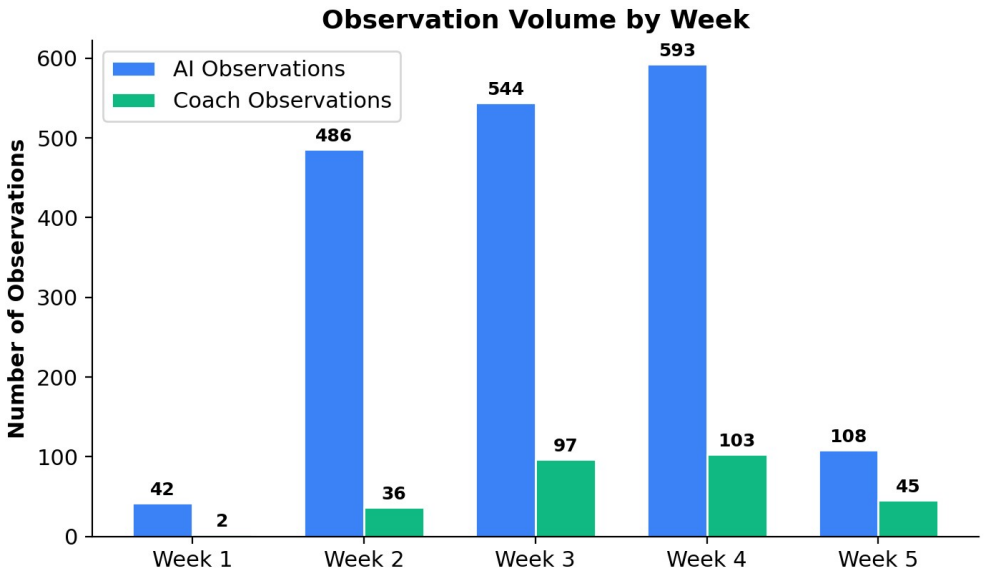
— Provide contextual judgment and relationship support

2. Data Overview

Observation Sample

Type	Count	Coverage
AI Observations (total)	2,217	Every lesson uploaded
AI Observations (completed)	1,773	With full scoring & metrics
AI Observations (pending/processing)	444	Awaiting processing
Coach Observations	283	Periodic in-person visits
School Visits	83	Administrative oversight
Overlap (both AI + coach)	68 teachers	Enables comparison

Observation Volume by Week



Distribution by Subject

Subject	AI Observations	% of Total
English Literacy	787	35.5%
Numeracy	718	32.4%
Urdu Literacy	712	32.1%

Distribution by Grade

Grade	AI Observations	% of Total
Grade 2	437	19.7%
Grade 3	603	27.2%

Grade 4	562	25.4%
Grade 5	615	27.7%

Lesson Plan Analysis

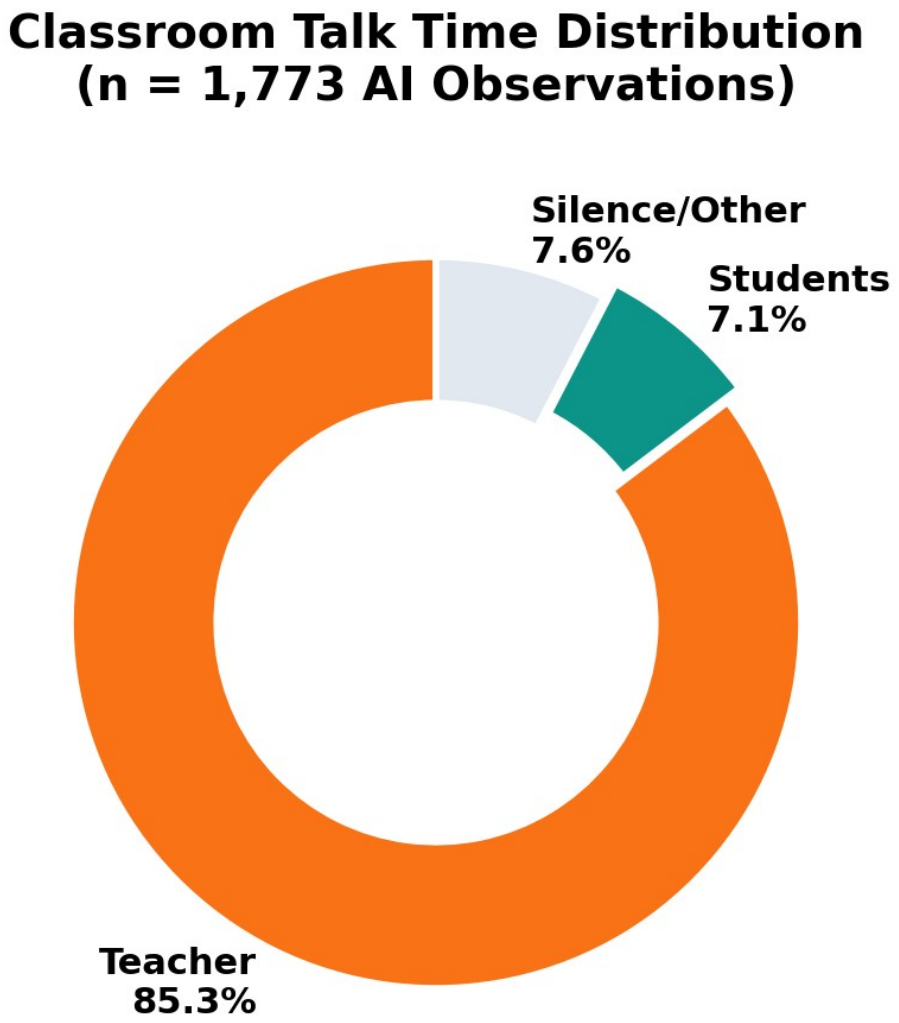
191 lesson plans were analyzed for prescribed practices:

Component	Presence in Plans
Clear learning objectives	100%
Open-ended questions	100%
Collaboration instructions	97%
Real-life connections	100%
Concrete manipulatives	100%
Differentiation by level	Yes

3. Key Findings

Finding 1: Teacher-Centered Instruction Dominates

Classrooms remain heavily teacher-dominated, with teachers occupying 85.3% of talk time.

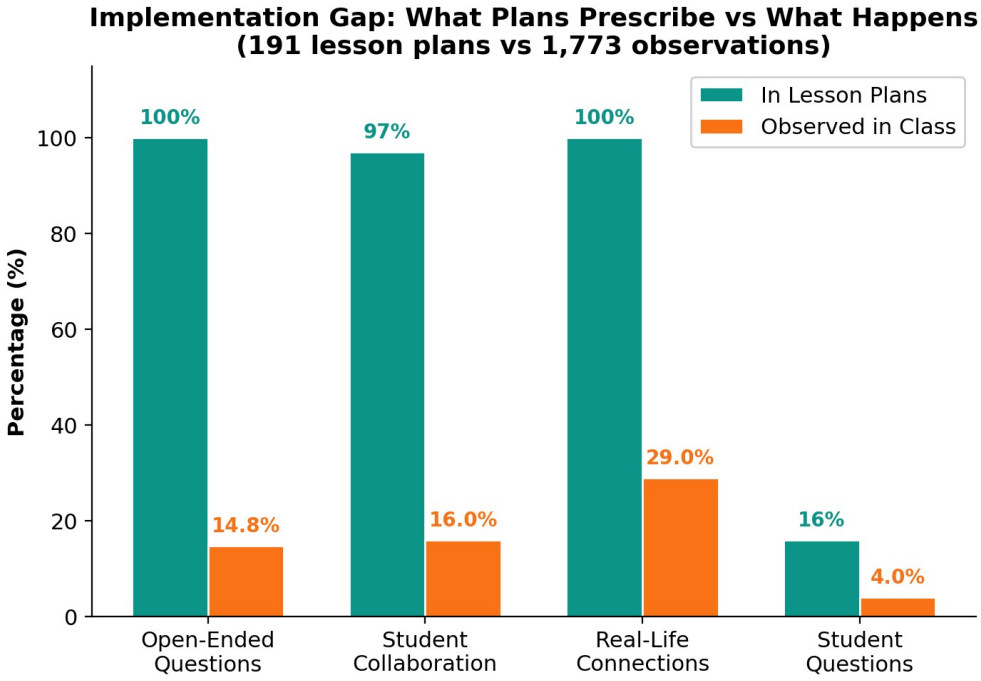


Speaker	% of Class Time	Minutes (40-min lesson)
Teacher	85.3%	34 minutes
Students	7.1%	3 minutes
Silence/Other	7.6%	3 minutes

Implication for research: In a class of 6-7 students (winter school average), each student speaks for approximately 25 seconds per lesson. In a regular session class of 50+, this would drop to ~3 seconds per student.

Finding 2: The Implementation Gap

The most significant finding is the persistent gap between what lesson plans prescribe and what actually happens in classrooms.



Practice	In Lesson Plans	Observed in Class	Gap
Open-Ended Questions	100%	14.8%	-85.2 pp
Student Collaboration	97%	16%	-81 pp
Real-Life Connections	100%	29%	-71 pp
Student Questions	16%	4%	-12 pp

Key insight: The lesson plans are well-designed. The challenge is implementation, not curriculum design.

Finding 3: Question Quality

Teachers ask an average of 17.5 questions per lesson, but the vast majority are closed-ended:

Question Type	Per Lesson	Percentage
Closed-ended	14.9	85.2%
Open-ended	2.6	14.8%

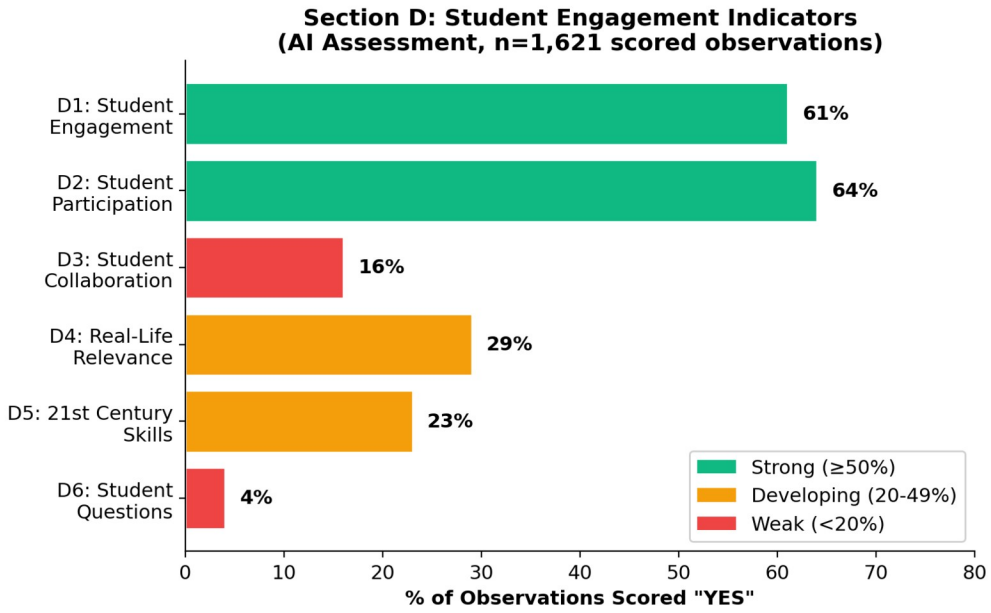
Examples observed:

— Closed: "Is this correct?" / "What is 2+3?" / "Do you understand?" / "All good?"

— Open: Rare instances of "Why?" or "How?" questions

Finding 4: Student Engagement Indicators (Section D)

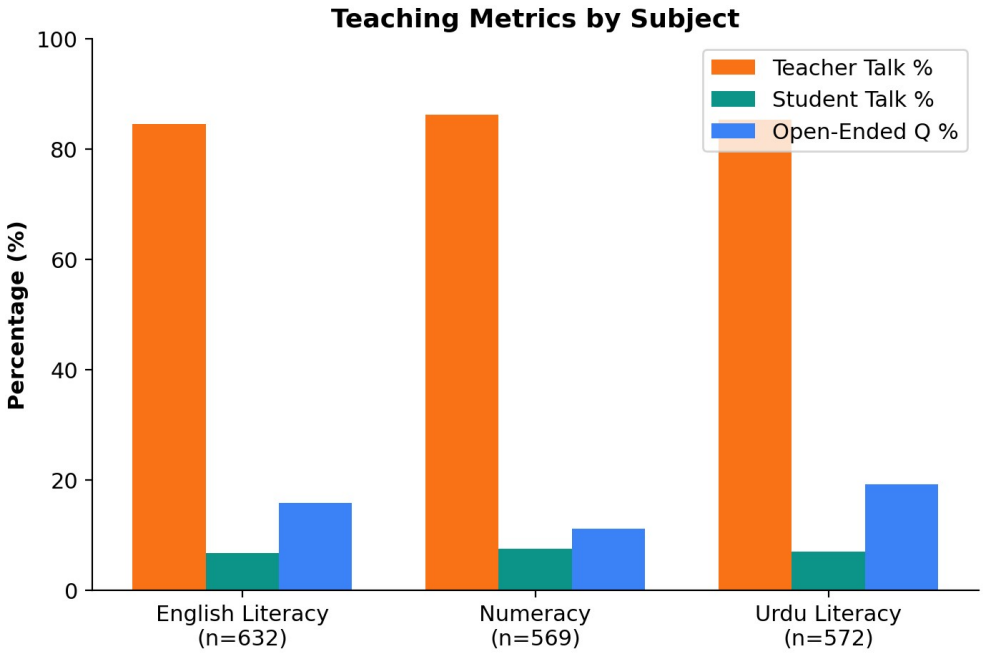
The AI assessment's Section D provides granular insight into student-centered practices:



Only 4% of observations show students generating their own questions — the weakest indicator across the entire framework. Collaboration (16%) and 21st century skills (23%) are also areas requiring significant attention.

Finding 5: Subject-Level Differences

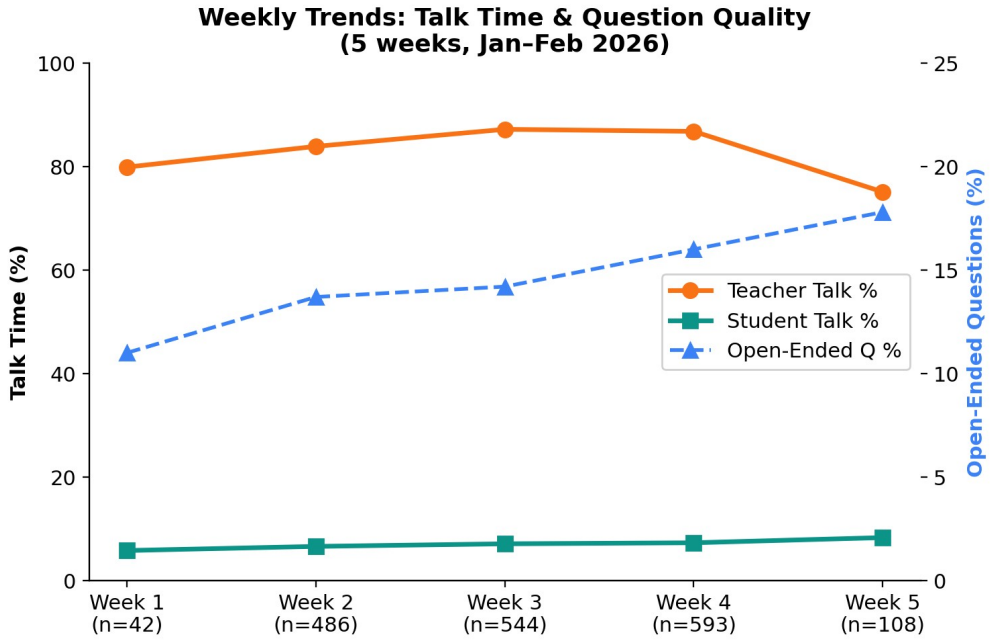
Teaching practices vary meaningfully by subject:



Urdu Literacy shows the highest rate of open-ended questioning (19.2%), while Numeracy has the lowest (11.2%). This may reflect the nature of the subjects — language instruction naturally invites more open-ended discussion.

Finding 6: Weekly Trends Show Modest Improvement

Over the 5-week programme, there are signs of gradual improvement in student-centered practices:

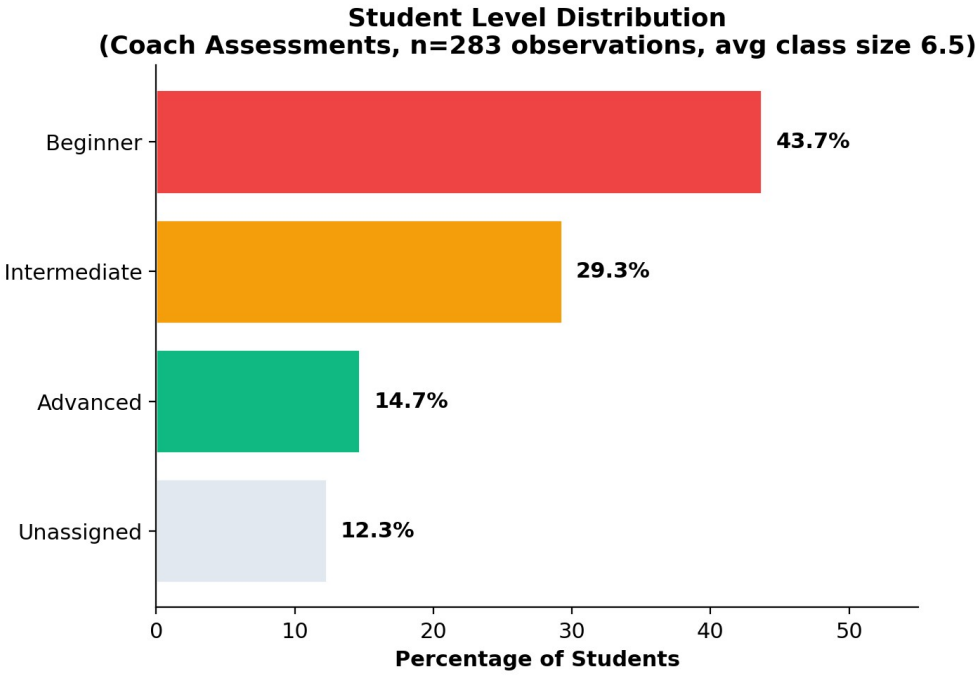


Metric	Week 1	Week 5	Change
Teacher Talk	79.9%	75.1%	-4.8 pp ↓
Student Talk	5.8%	8.3%	+2.5 pp ↑
Open-Ended Questions	11.0%	17.8%	+6.8 pp ↑

By Week 5, open-ended questioning increased by 62% relative to Week 1, and teacher talk time decreased. These trends, while modest in absolute terms, suggest the coaching model is producing measurable shifts over time.

Finding 7: Student Level Distribution

Coach assessments reveal the majority of students are still at the beginner level:



With 43.7% of students at beginner level and only 14.7% at advanced, there is a clear need for continued differentiated instruction. The lesson plans prescribe differentiation, but implementation data (Finding 2) suggests this remains a gap.

Finding 8: Teacher Strengths

Teachers demonstrate strong performance in teacher-centered practices:

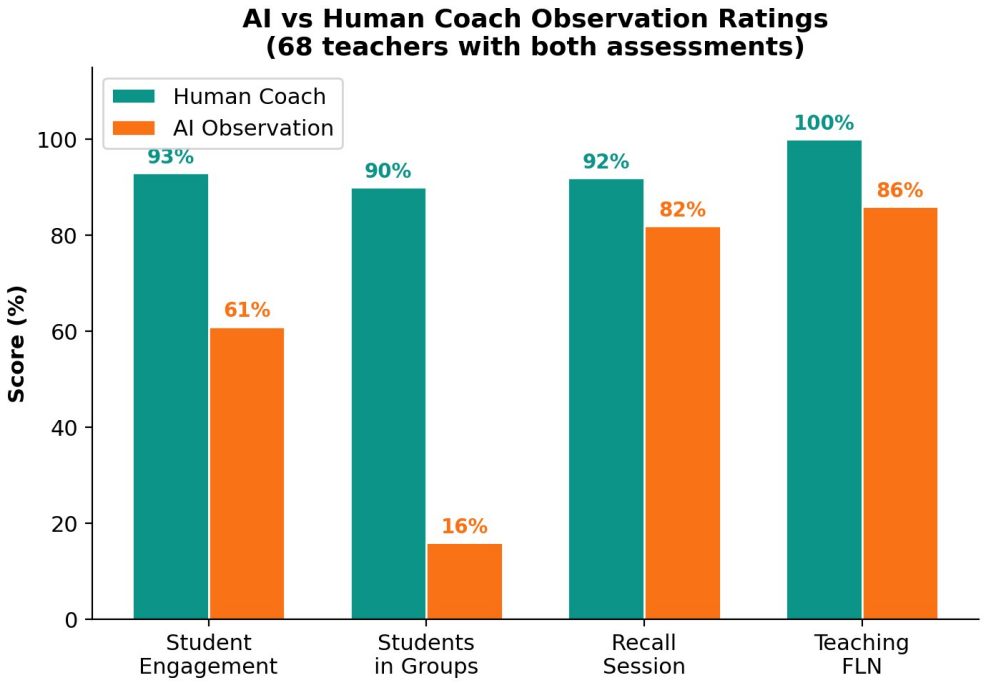
Practice	AI Score
Classroom management	~90%
Stating learning objectives	86%
Following content sequence	82%
Maintaining safe environment	~90%
Asking questions (mostly closed)	85%
Using physical resources	Observed consistently

Teachers follow the concrete parts of lesson plans (materials, sequence) but struggle with pedagogical parts (questioning strategies, collaboration facilitation).

4. Measurement Validity Notes

AI vs Human Coach Ratings

Significant discrepancies persist between AI and human coach ratings, with human coaches rating substantially higher across all indicators:



Indicator	AI Rating	Human Rating	Gap
Student Engagement	61%	93%	-32 pp
Students in Groups	16%	90%	-74 pp
Recall Session	82%	92%	-10 pp
Teaching FLN	86%	100%	-14 pp

Why These Gaps Exist

Factor	Human Assessment	AI Assessment
Definition of engagement	"Students appear attentive"	"Students ask questions, discuss"
Evidence standard	General impression	Specific transcript analysis
Threshold	Did it happen at all?	Did it happen meaningfully?
Focus	Programme compliance	Teaching quality depth

Implications for Research

Both measurement approaches capture different aspects of classroom reality:

— Human observations assess programme fidelity and general compliance

— AI observations assess depth of pedagogical practice

Neither is "wrong," but they measure different constructs. This has implications for how student outcomes are interpreted relative to teacher practice metrics.

5. Research Considerations

Winter School vs Main Academic Session

Factor	Winter School	Main Session
Curriculum pressure	None	High
Class size	~6.5 students	Typically 50+
Teacher focus	FLN only	Full curriculum
Time available	Protected	Competing demands

Teacher feedback:

"During the regular session, we are mostly worried about completing the curriculum. The attendance — there are too many students in the classroom for us to focus on foundational skills."

Questions for Future Research

— Context Transfer: How do teaching practices and student outcomes differ between winter school and main academic sessions?

— Class Size Effects: How does student-to-teacher ratio affect implementation of student-centered practices?

— Measurement Alignment: What is the relationship between AI-measured teaching quality and student learning outcomes?

— Cultural Adaptation: How should student engagement metrics be interpreted in contexts where student questioning is not culturally normative?

— Sustainability: What programme elements predict sustained practice change over time?

— Dosage Effects: The Week 5 improvements in student talk and open-ended questioning — do they persist, plateau, or require continued coaching support?

6. Summary Statistics

Observation Data

Total Observations: 2,217
├── AI Observations: 2,217
│ ├── Completed with scores: 1,773
│ ├── Pending/processing: 444
│ ├── With talk time data: 1,773
│ └── With question counts: 1,773
├── Coach Observations: 283
│ ├── FLN checklist items: ~15
│ └── Rating scales: ~5
└── School Visit Observations: 83

Unique Teachers: 87
├── With AI observations: 72
├── With coach observations: 69
└── With both AI + coach: 68

Lesson Plans Analyzed: 191
Students Enrolled: 2,247

Implementation Gap Summary

Practice	In Plans	Observed	Gap
Open-Ended Questions	100%	14.8%	-85.2 pp
Student Collaboration	97%	16%	-81 pp
Real-Life Connections	100%	29%	-71 pp
Student Questions	16%	4%	-12 pp

Talk Time Distribution

Speaker	% of Time	Minutes (40-min)
Teacher	85.3%	34 min
Students	7.1%	3 min
Silence/Other	7.6%	3 min

Appendix: Observation Frameworks

AI Observation Framework (30 Indicators)

Section B — Lesson Planning (13 indicators)

Learning objectives, resources, timing, assessment strategies, differentiation, transitions, closure, vocabulary, higher-order thinking, engagement, safe environment

Section C — High-Leverage Practices (11 indicators)

Modeling, guided practice, checking understanding, wait time, feedback, questioning, discussion, scaffolding, metacognition, classroom management

Section D — Student Engagement (6 indicators)

Student engagement, participation, collaboration, real-life relevance, 21st century skills, student questions

Human Observation Framework (~15 Indicators)

FLN-Specific Checklist:

Learning objectives clear (yes/no), Student engagement (yes/no), Students in groups (yes/no), Review activity done (yes/no), FLN drills completed (blending, segmenting, etc.)

Rating Scales:

Follows sequence (satisfactory/excellent), Differentiates instruction (satisfactory/excellent), Corrects mistakes (satisfactory/excellent)

Contact

Programme Partner: Taleemabad

Research Partners: Georgetown University, World Bank

This brief provides programme context for research purposes. For questions about methodology or data access, please contact the programme team.

Balochistan Winter School Programme, February 2026