Self-Assessed and Observed Instructional Competence of Teachers:

Basis for Targeted Technical Assistance

A Basic Research

[Copy for blind review: the researchers' names, positions, school, and schools division are removed. The final copy lists them here.]

Conducted February to July 2026

Abstract

This basic research compared the self-assessment of the Teachers I to VII of one large public junior high school with their classroom observations, read the observers' notes for the practices behind the ratings, and asked which development needs the school's technical assistance should address first. With the consent of 104 of 114 teachers, the school's School Year (SY) 2025–2026 records were linked by code: the Electronic Self-Assessment Tool (e-SAT), the first full-period observation's ratings on six indicators of the Philippine Professional Standards for Teachers (PPST), and its notes, coded by two coders; 96 teachers rated the technical assistance they had received. Teachers rated their capability High or Very High in 82.1% of e-SAT answers. Observation found a need (Satisfactory or lower) in 7.7% to 54.5% of teachers per indicator, most often in differentiated instruction and higher-order thinking; on these and on classroom structure, observation flagged more teachers than self-assessment did (McNemar's exact tests, Holm-adjusted $p <$ .01), and agreement was slight (κ = .02 to .39). The e-SAT did not flag 79.7% of the needs the observations found, a share that fell with years in teaching (*OR* = 0.41 per band, adjusted for career stage and indicator). Lessons whose notes recorded why or how questions, learners' explanations, tiered tasks, or group roles were rated higher. Post-observation conferences held within three days satisfied 92.6% of teachers; later ones, 55.1%. The resulting plan for SY 2026–2027 sets school-wide sessions on two indicators, small-group coaching on two, and individual coaching on two.

Keywords: technical assistance, teacher self-assessment, classroom observation, professional development needs, instructional supervision

Acknowledgment

The researchers thank the 104 teachers who allowed their records to be read for this study, and the 96 who also answered the questionnaire; the school's e-SAT focal person, who served as the data custodian and kept every name away from the researchers; the head teachers and master teachers whose observation notes made the third research question possible; the two master teachers of a neighboring school who coded the notes; the research ethics committee that reviewed the protocol; and the members of the Schools Division Research Committee. The study was funded by the Basic Education Research Fund (BERF) for 2026.

Self-Assessed and Observed Instructional Competence of Teachers: Basis for Targeted Technical Assistance

Introduction of the Research

Every public school teacher in the Philippines now leaves two records of their teaching. The first is their own: under the multi-year Performance Management and Evaluation System (PMES), a teacher rates their capability and their priority for development on each of the 37 indicators of the Philippine Professional Standards for Teachers (PPST), in a full self-assessment done "once every three (3) years" (Department of Education [DepEd], 2025a, item 25). The second is their rater's: full-period classroom observations scored with the Classroom Observation Tool (COT), whose rating "shall be solely based on the actual classroom observation as recorded in the Observation Notes Form" (item 55). The memorandum expects the two to meet in the rater's Instructional Supervisory (IS) Plan, which manages professional development "identified based on their IDPs, SAT result, and observed performance gaps throughout the school year" (item 30), and it asks for "regular monitoring and targeted provision of technical assistance" (item 38).

The school head is accountable for that assistance. The Philippine Professional Standards for School Heads (PPSSH) define technical assistance as "coaching, guiding and empowering school personnel" (DepEd, 2020, Glossary), and its Career Stage 2 indicator 3.2.2 asks school heads to "provide technical assistance to teachers on teaching standards and pedagogies within and across learning areas to improve their teaching practice" (DepEd, 2020, Indicator 3.2.2). The school-based professional development guidelines of the

National Educator's Academy of the Philippines (NEAP) tell the school head to identify the needs of teachers from the self-assessment tool, the COT, the individual development plan, and the IS Plan, among other data (DepEd, 2024, Encl. No. 1, Sec. I.A).

Whether those data agree is an open question in this school, and nationally the signs are poor. The Second Congressional Commission on Education (EDCOM II) found that "despite the presence of formal Training Needs Assessments (TNAs), there is substantial misalignment between training content and teachers' identified needs" (EDCOM II, 2026, p. 288). A decade earlier, a national survey set teachers' self-ratings beside tests of their subject knowledge and concluded that "self-assessments of teachers' strengths and weaknesses may not be a good basis on which to plan professional development activities" (World Bank, 2016, p. 43). If a school plans its assistance from the e-SAT alone, it may teach what teachers already do and miss what they do not see.

This study used the school's own records of SY 2025–2026 to compare the two views indicator by indicator, to read what the observers actually saw, and to set the school's technical assistance for SY 2026–2027 from both. It addresses the Human Resource Development theme of the Basic Education Research Agenda, under Career Development: "What kind of capacity-building activities are necessary and most effective in addressing development needs and improving the work performance of teachers and other DepEd personnel?" (DepEd, 2016, p. 12).

Literature Review

The review follows the study's questions: how accurate teachers' self-assessments are, how much one classroom observation can be trusted, and how a school turns measured needs into professional development and technical assistance.

Teachers' Self-Assessments and What Observers See

People's self-views hold "only a tenuous to modest relationship with their actual behavior and performance," and people tend to "overrate themselves" (Dunning et al., 2004, p. 69). Across studies, self-ratings correlated with objective performance at about .29 (p. 71). Kruger and Dunning (1999) found that participants who scored lowest overestimated themselves most; later work traced much of that pattern to regression toward the mean and to a general better-than-average bias, so that the gap between self-rating and measured performance is not by itself proof that the weakest do not know they are weak (Gignac & Zajenkowski, 2020; Krueger & Mueller, 2002).

Studies of teachers point the same way, with differences by what is asked. A survey composite of how often teachers used reform practices agreed well with observation, but it failed "to capture the quality with which teachers engage in reform practices" (Mayer, 1999, p. 29), and teachers' daily logs were more valid for frequent and coarse-grained activities than for detail (Camburn & Barnes, 2004). In a three-year program, "as the teachers' practice approached exemplary, the discrepancy between observed and self-perceptions decreased," and where it remained, teachers saw themselves

closer to exemplary than observers did (Van Zoest et al., 2002, p. 245). Wiggs et al. (2023) found only small correlations between teachers' and administrators' ratings on the same instrument, with no difference in means. Self-assessment helps teachers grow when it is "bundled with other professional growth strategies: peer coaching, observation by external change agents, and focused input on teaching strategies" (Ross & Bruce, 2007, p. 146).

Philippine evidence is consistent. In the national Public Education Expenditure Tracking and Quantitative Service Delivery Study, teachers rated their subject knowledge "very highly," in "stark contrast with the same teachers' results on the objective subject matter tests," and the association between self-ratings and test scores was "relatively weak" (World Bank, 2016, pp. 43–44). In a classroom observation study of Grades 1 to 3 in Mindanao, 65% of teachers believed they asked questions throughout the lesson to check understanding, "only 25% of them do so in practice," and 72% rated themselves effective at adjusting teaching to their learners' level against 10% who scored high on it (Molina et al., 2018, pp. 37–38). In a national digital course on early language, literacy, and numeracy, teachers' self-assessment rose far more than their tested knowledge (Cohen's d = 0.68 against 0.14; Oakley et al., 2018, pp. 76–77).

What One Classroom Observation Can Tell

A single observation by a single rater is a weak measure of a teacher. In the Measures of Effective Teaching (MET) project, "single observations produced reliabilities ranging from 0.14 to 0.37" (Kane & Staiger, 2012, p. 8).

With school personnel as observers, administrators rated their own teachers higher than administrators from other schools did, and a large share of score variance lay in the pairing of a rater with a teacher, which only more raters, not more lessons, can reduce (Ho & Kane, 2013, pp. 4, 14). High agreement between raters is not enough either: agreement is driven by how often a behavior occurs and by chance, and an item on which raters "match" mostly because the behavior was absent tells little about teachers (Hill et al., 2012).

In Philippine schools, observation has often served evaluation more than development. A baseline study in Regions V and VI found that a third of school heads saw no strong link between their classroom observations and the Learning Action Cell (LAC) program, suggesting that "many school heads view classroom observations more as an administrative function than an earnest effort to enhance teaching quality" (RTI International, 2020, p. 3), and teachers in a case study described observation as "only to look for mistakes" (Torres et al., 2024, p. 462). Observation notes, which record what happened, can be read for the practices behind a rating. Content analysis turns such notes into codes that can be counted, and the reliability of the coding should be reported per code with Krippendorff's alpha, with thresholds set before the coding (Krippendorff, 2004; O'Connor & Joffe, 2020).

Turning Measured Needs Into Technical Assistance

Needs assessment in teacher education commonly follows Borich (1979), who defined a training need as "a discrepancy between an educational goal and trainee performance in relation to this goal" (p. 3) and ranked needs by the discrepancy weighted by its average importance. Borich proposed

checking teachers' ratings against supervisors' and withholding competencies on which the two differ "pending clarification from other data sources, such as the classroom observation of teachers" (pp. 7–8). Because the weighted score treats ordinal ratings as numbers, Narine and Harder (2021) proposed a ranked discrepancy model based on the share of respondents whose ability falls below, above, or at the importance they assign; in their data the two rankings correlated at .98.

Targeted support has better evidence than one-size-fits-all training. A meta-analysis of 60 causal studies of teacher coaching found average effects of 0.49 standard deviations on instruction and 0.18 on achievement, smaller in large programs, and larger when coaching was paired with group training (Kraft et al., 2018). In a field experiment, pairing teachers who scored low on specific observation skills with colleagues who scored high on the same skills raised student achievement by 0.12 standard deviations (Papay et al., 2020). Desimone and Garet (2015) describe the aim as "collective experiences that are targeted to groups of teachers with similar needs" (p. 255), while noting that the effects of differentiated professional development are not yet known. Across 104 randomized trials, the average effect of professional development on learners was small, an effect size of .05; programs balanced across building knowledge, motivating teachers, developing teaching techniques, and embedding practice averaged .15, though the difference from other programs was not statistically significant (Sims et al., 2025, pp. 231, 237).

DepEd's own diagnosis is that support has been driven from the top. NEAP's transformation report found professional learning made up of

"uncoordinated and supply-driven activities that have uneven quality and impact" (DepEd, 2019, p. xxvii). The Basic Education Development Plan 2030 notes that the Schools Division Offices' "provision of technical assistance is largely program-driven" (DepEd, 2022, p. 97). A national study found that "only 17 percent of school division superintendents (SDSs) said that division offices based their decisions about teacher training on the actual needs of teachers" (World Bank, 2016, p. 46), and teachers in recent focus groups described training needs assessments as "merely procedural" (Rivera et al., 2025, p. 18).

Technical Assistance as a Service

Technical assistance is a service to teachers, and services can be rated by those who receive them. The government's harmonized Client Satisfaction Measurement covers internal services, including those requested by an agency's own personnel, and scores each service quality item by "the percentage of respondents that rated 'Agree' and 'Strongly Agree'" (Anti-Red Tape Authority [ARTA], 2022, items 4.1.1.2, 4.5.2). Three of its dimensions, responsiveness, reliability, and assurance, carry the names of Parasuraman et al.'s (1988) dimensions. No Philippine study found for this review measured teachers' satisfaction with the assistance their own school gives them.

Synthesis

Self-ratings tell a school what teachers believe and want; observations tell it what one lesson showed. Neither is enough alone, and the literature gives reasons to expect them to disagree most where practice is weakest. This study compares them indicator by indicator in one school, reads the

observers' notes for the practices behind the ratings, and asks teachers how they rated the assistance they received, so that the school's plan rests on all three.

Research Questions

The study answers four questions about the Teachers I to VII of one public junior high school in SY 2025–2026:

1. How do the teachers rate their own capability and development priorities on the 37 PPST indicators in the e-SAT, and what ratings did they receive on the six indicators of the first full-period classroom observation?
2. How closely do self-assessed and observed needs agree on those six indicators, and are years in teaching and career stage related to the needs that the self-assessment does not flag?
3. What teaching practices do the observers' notes record on four of the indicators, and do lessons with each practice receive different ratings from lessons without it?
4. Which development needs should the school's technical assistance address first and in what form, and how did teachers rate the technical assistance they received in SY 2025–2026?

The study tested four null hypotheses at the .05 level:

H01: For each of the six indicators, the proportion of teachers with a self-assessed need equals the proportion with an observed need.

H02: Years in teaching and career stage are not related to the odds that an observed need goes unflagged in the e-SAT.

H03: For each coded practice, the indicator's ratings do not differ between lessons whose notes recorded the practice and lessons whose notes did not.

H04: Teachers' satisfaction with the post-observation conference does not differ by the number of days between the observation and the conference.

Scope and Limitation

The study covers the Teachers I to VII of one large public junior high school who consented, their e-SAT of November 2025, their first full-period classroom observation of SY 2025–2026, its notes, the date of its post-observation conference, and their answers to a questionnaire in May 2026. Master teachers and head teachers were not participants: they were the raters, and the PPST indicators they are rated on belong to another career stage.

Four limits shape what the results can say. First, the observation is one lesson rated by one rater or by a team agreeing on one rating; it measures the practice seen that day, not the teacher's competence in general. Second, the e-SAT asks for capability and priority on four labels that the tool does not define, so a "Low" may mean a skill the teacher lacks or a situation, such as learners from Indigenous groups, that the teacher has never met. Third, the COT rates only six of the 37 indicators in the first observation of SY 2025–2026, so the comparison covers those six; the other 31 rest on the e-SAT alone. Fourth, the study is of one school, and its results describe that school's teachers and raters, not DepEd's.

Research Methodology

Research Design

The study is descriptive-comparative research with a convergent mixed-methods strand (Fetters et al., 2013). Two measures of the same indicators, the teacher's self-assessment and the observation, were compared within teachers; the observers' notes were coded and counted to explain the ratings; and a questionnaire measured satisfaction with the assistance received. All four sources were records the school already kept or produced in the normal course of the PMES, except the questionnaire, so the study added no observation and no test to any teacher's year.

Sampling

The population was all 114 Teachers I to VII of the school in SY 2025–2026: 86 at the Beginning towards Proficient career stage (Teachers I to III, rated on COT Levels 2 to 6) and 28 at the Proficient stage (Teachers IV to VII, Levels 3 to 7; DepEd, 2025a, Table 10). There was no sampling: every one was invited. Of these, 104 consented (91.2%); the 10 who did not were never identified to the researchers. Consent did not differ by career stage (Fisher's exact test, $p = .257$) or by years in teaching, $x^2(3) = 1.60$, $p = .660$ (Table 1). Of the consenting teachers, 96 (92.3%) answered the questionnaire.

Table 1

Eligible and Consenting Teachers by Career Stage and Years in Teaching

Group	Eligible	Consented	Rate
Beginning towards Proficient (Teachers I–III)	86	80	93.0%
Proficient (Teachers IV–VII)	28	24	85.7%
1–4 years in teaching	18	17	94.4%

Group	Eligible	Consented	Rate
5–9 years	45	40	88.9%
10–19 years	37	35	94.6%
20 years or more	14	12	85.7%
All	114	104	91.2%

Note. Eligible counts are from the school's list of personnel; consenting counts are from the custodian's coded list. Years in teaching are kept in bands so that no teacher can be recognized.

The teachers were rated by fourteen raters, as DepEd Memorandum No. 089, s. 2025 provides for junior high school teachers (DepEd, 2025a, Table 18): seven head teachers heading seven departments, a master teacher heading the eighth, and six more master teachers sharing the ratees of the six largest departments. Each rater had 4 to 12 consenting ratees.

With about 100 consenting teachers expected, a share near 50% would be estimated within ±9.8 percentage points. In 20,000 simulated samples of 100 teachers in which 15% flag themselves and 35% are flagged by observation (5% by both), McNemar's exact test at Holm's strictest threshold for six tests (.0083) rejected in 68.3% of samples, and in 16.5% when the observed share was 25%; the half-width of kappa's 95% interval would be about .17. The study could therefore detect only large gaps with confidence, and its kappas would be estimated roughly.

Data Collection

Four records and one questionnaire were used (Table 2).

Table 2

Sources of Data

Source	What it holds	When	Who produced it
e-SAT 2025–2028, full self-assessment	Level of Capability and Priority for Development, each 1 Low, 2 Moderate, 3 High, 4 Very High, on the 37 PPST indicators at the Proficient career stage	November 17 to 28, 2025	Each teacher (DepEd, 2025a, Annex A)
COT rating sheet or inter-observer agreement form, first full-period observation	The rating on the six classroom observable indicators of the first observation: 1.1.2, 1.4.2, 1.5.2, 2.3.2, 3.1.2, 4.1.2	October 13, 2025 to March 26, 2026	The rater, or the observers by consensus (items 55–56)
COT Observation Notes Form	The observer's written record of the lesson	Same	The rater
Performance Monitoring and Coaching Form (PMCF)	The date of the post-observation conference	Same	The rater (item 39)
Technical assistance questionnaire (Appendix B)	Forms of assistance received in SY 2025–2026, satisfaction with each, preferences for SY 2026–2027, an open question	May 4 to 15, 2026	Each teacher who chose to answer

Note. The COT rating was converted to the PMES five-point rating with Table 13 of DM 089, s. 2025: for Teachers I to III, Levels 6, 5, 4, 3, and 2 are ratings 5 to 1; for Teachers IV to VII, Levels 7 to 3.

The e-SAT. DM 089, s. 2025 requires a full self-assessment of all 37 indicators once every three years. The school ran the e-SAT 2025–2028 from November 17 to 28, 2025. The tool's own instructions state that "other school personnel, including the School Head, are not allowed to see the results of this tool" (DepEd, 2025a, Annex A, Home sheet). The study respected this: no researcher saw a named e-SAT (see Ethical Issues).

The observations. DM 089, s. 2025 set two full-period observations a year, the first in the first or second quarter, rating six indicators (Table 7).

The memorandum was issued on October 1, 2025, and the school's first observations ran from October 2025 to February 2026. A guidance memorandum of March 3, 2026 then required only one observation for SY 2025–2026: teachers already observed once were deemed compliant, and teachers not yet observed had one observation covering the nine classroom observable indicators of the year, up to three of which could remain unobserved (DepEd, 2026b, pp. 1–3). By the end of the school year, 33 consenting teachers had two observations, 62 had one in the first observation's window, and 9 had their only observation after March 3. For every teacher the study used the first observation, as recorded: a teacher's final IPCRF rating, which under the guidance could take the higher of two observations, was not used. An indicator marked Not Observed, which the rating forms score at the lowest level of the teacher's range, was treated as missing; this left three teachers without a rating on 2.3.2 and three without one on 3.1.2. The second observation of the 33 teachers observed twice, which repeated indicators 1.1.2, 1.4.2, and 1.5.2, served only to check how stable a rating is from one lesson to the next.

The order of the two measures. 80 teachers took the e-SAT before their first observation and 24 after it, when their post-observation conference may already have shaped their self-ratings. The comparison was repeated for the first group alone.

The notes. The custodian removed names, sections, and dates from each first-observation notes form. Two master teachers of a neighboring school, who did not know the ratings, coded each note for ten practices tied

to four indicators (Appendix C). The codebook was drafted from the COT rubric's descriptions of the indicators, tried on ten notes from the previous school year's observations, which the custodian also redacted, and revised before the formal coding. The two coders coded every note independently; disagreements were settled by rereading the note together, and the agreed code was used. Reliability was set in the proposal as Krippendorff's alpha of at least .800 for a code to support a conclusion and at least .667 for a tentative one (Krippendorff, 2004).

The questionnaire. The technical assistance questionnaire (Appendix B) asked which of six forms of assistance the teacher had received in SY 2025–2026 and, for each one received, six statements rated from 1 (Strongly disagree) to 5 (Strongly agree). The statements adapt the Service Quality Dimensions of ARTA's Client Satisfaction Measurement to assistance within a school (ARTA, 2022, 2023): the help came when I needed it (responsiveness); it addressed the needs we had identified (reliability); the person who helped knew the subject and the class (assurance); I was treated with respect (assurance, which ARTA's form asks as courtesy); I was able to use what I learned (outcome); and I am satisfied overall. Teachers then ranked three forms they wanted for SY 2026–2027 and answered one open question: what help would most improve your teaching this year? Two teachers outside the study read the draft for clarity. The same two coders coded the open answers into six categories.

Ethical Issues

The principal, who approves every teacher's PMES rating, led the study, so the protocol was built so that no teacher's standing could depend on taking part. The National Ethical Guidelines for Research Involving Human Participants name employees among those "liable to the authority of others" (Philippine Health Research Ethics Board [PHREB], 2022, pp. 91–92) and, for research in organizations, advise recruiting "in the absence of supervisors" and that "the organization should not be informed about who among their employees or staff agreed or refused to participate" (pp. 253–254). Accordingly:

1. A PHREB-accredited research ethics committee reviewed the protocol in full before any record was read.
2. The research coordinator, a teacher who rates no one, explained the study and collected consent forms in sessions with no rater, principal, or assistant principal present. The forms went in sealed envelopes to the custodian; the researchers never learned who consented.
3. The school's e-SAT focal person, who already held the e-SAT files for the school's consolidation and was not a researcher, acted as data custodian. The custodian linked each consenting teacher's e-SAT, COT ratings, notes, and conference date under a random code, replaced the raters' names with random codes, kept the key, and destroyed it after the report was accepted.
4. Consent covered the research use of the records named in the form and nothing else (National Privacy Commission [NPC], 2023, Secs. 8 and

16). The questionnaire had its own consent page; answering it was voluntary.

5. The study made no decision about any teacher. Its results are reported for groups, and the technical assistance plan names indicators and numbers of teachers, never a person. Which teacher receives which assistance is settled in each rater's IS Plan with the teacher's own development plan, as DM 089 already provides (items 26–30).
6. The data set submitted with this report holds codes, career stage, and years in bands; it omits the rater codes, which anyone who knows the school's rating assignments could use to recognize a teacher.

   No learner took part, and no learner's data were read.

Plan for Data Analysis

RQ1. The e-SAT answers were summarized as shares at each level and as each indicator's mean. The observed ratings were summarized by indicator, and the six indicators were compared with the Friedman test, with Kendall's *W* as its effect size.

RQ2. A self-assessed need was a capability of Low or Moderate; an observed need was a rating of Satisfactory or lower (PMES rating 3 or less: COT Level 4 or lower for Teachers I to III and Level 5 or lower for Teachers IV to VII, below the level the transmutation table calls Very Satisfactory). For each indicator, the teachers fell into four groups: a need seen by both, by observation only, by self-assessment only, or by neither. McNemar's exact test compared the two proportions (McNemar, 1947), with Holm's (1979) adjustment across the six indicators. Agreement was measured by Cohen's

kappa with a 95% interval from 4,000 bootstrap samples of teachers, and, since kappa depends on how common a need is, by the prevalence-adjusted bias-adjusted kappa (PABAK; Byrt et al., 1993). Spearman's rho related the four-level capability to the five-level rating. Two sensitivity analyses repeated the comparison with an observed need set at a rating of 2 or lower, and with only the teachers whose e-SAT preceded their observation. For H02, among the teacher-indicator pairs with an observed need, the odds that the e-SAT did not flag it were modeled by generalized estimating equations with a logit link and exchangeable correlation within teacher (Liang & Zeger, 1986), with career stage, years band, and indicator as predictors.

Two checks asked how far one observation can be trusted. A linear mixed model of the rating, with the raters and the teachers within them as random effects, estimated the share of variance lying between raters' groups of ratees. For the teachers observed twice, the two ratings of each repeated indicator were compared by exact agreement and quadratic weighted kappa (Fleiss & Cohen, 1973).

RQ3. For each code, the coders' reliability was Krippendorff's alpha, reported per code and as the smallest value rather than pooled (Krippendorff, 2004). Each code's indicator ratings were compared between lessons with and without the practice by the Mann-Whitney test, with the rank-biserial correlation as effect size and Holm's adjustment across the ten codes. Only codes meeting the reliability threshold support a conclusion.

RQ4. Each of the 37 indicators was ranked by a mean weighted discrepancy score in Borich's (1979) form, using the e-SAT's own ratings: the

gap to Very High (4 minus capability), weighted by the indicator's mean priority for development. Intervals and rank ranges came from 4,000 bootstrap samples of teachers. As a check, the indicators were also ranked by the ranked discrepancy model (Narine & Harder, 2021): the share of teachers who rated capability above priority minus the share who rated it below. For the six observed indicators, the observed share with a need decided the form of assistance under a rule set in the proposal: 40% or more of observed teachers, school-wide sessions; 15% to 39%, small-group coaching; under 15%, individual coaching in the IS Plan. Satisfaction was reported as ARTA scores it: the share who agree or strongly agree, with Wilson 95% intervals. H04 was tested by Fisher's exact test (conference within three days or later) and by logistic regression of overall satisfaction on the number of days. Preferences were summed as Borda points (3 for a first choice, 2 for a second, 1 for a third). The open answers were counted by category.

The analysis was done in Python (pandas, SciPy, statsmodels) and checked in Microsoft Excel, where every statistic that Excel can compute is a formula in the executive data set.

Discussion of Results and Recommendations

How Teachers Rated Themselves

Teachers rated their capability High in 58.8% of their 3,848 e-SAT answers and Very High in 23.4%; 15.8% were Moderate and 2.1% Low. One teacher in eight (12.5%) rated every indicator High or Very High; the rest named a median of 6 indicators at Moderate or Low. Capability and priority

moved against each other, ρ = −.46: teachers wanted to grow where they felt weakest.

The lowest self-ratings were on indicators about particular groups of learners and tools: 3.5.2, learners from Indigenous groups (mean 2.16); 3.3.2, learners with disabilities, giftedness, and talents (2.28); 3.4.2, learners in difficult circumstances (2.52); and 1.3.2, the use of ICT (2.59). The highest were 7.2.2, upholding the dignity of teaching (3.61), and 6.4.2, complying with school policies (3.47). On the six indicators the first observation rates, between 5.8% and 16.3% of teachers rated their capability Low or Moderate.

What the Observations Found

The first observations rated most teachers Very Satisfactory or Outstanding on content knowledge (1.1.2) and sequenced lessons (4.1.2), and many fewer on differentiated instruction (3.1.2) and higher-order thinking (1.5.2) (Table 3). The six indicators differed, $x^2(5) = 172.46$, $p < .001$, $W = .35$, $n = 98$ teachers rated on all six. Of the 104 teachers, 76 (73.1%) had at least one observed need, and 26 had three or more.

Table 3

PMES Ratings on the Six Indicators of the First Full-Period Observation

Indicator	*n*	1 Poor	2 Unsatisfactory	3 Satisfactory	4 Very Satisfactory	5 Outstanding	*M*	Need (3 or lower)
1.1.2 Content knowledge	104	1	2	5	40	56	4.42	7.7%
1.4.2 Literacy and numeracy strategies	104	6	6	17	49	26	3.80	27.9%
1.5.2 Higher-order thinking	104	8	14	26	45	11	3.36	46.2%
2.3.2 Classroom structure and hands-on work	101	4	7	16	52	22	3.80	26.7%
3.1.2 Differentiated instruction	101	13	13	29	38	8	3.15	54.5%

Indicator	*n*	1 Poor	2 Unsatisfactory	3 Satisfactory	4 Very Satisfactory	5 Outstanding	*M*	Need (3 or lower)
4.1.2 Sequenced teaching and learning	104	2	3	5	48	46	4.28	9.6%

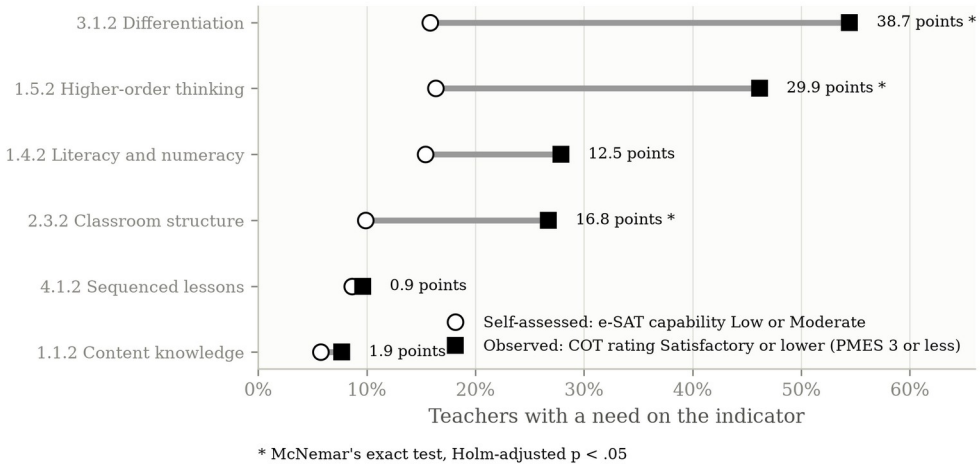
Note. Counts of teachers. Ratings are the PMES five-point ratings converted from COT levels (DepEd, 2025a, Table 13). Indicators marked Not Observed are excluded.

Where Self-Assessment and Observation Agree

On four of the six indicators, the observation flagged far more teachers than the e-SAT did (Figure 1). On differentiated instruction, 16 teachers (15.8%) rated their capability Low or Moderate, but 55 (54.5%) were observed at Satisfactory or lower; on higher-order thinking, 16.3% against 46.2%. The difference was significant after Holm's adjustment for 3.1.2 and 1.5.2 (p < .001 each) and for 2.3.2 (p = .009), and not for 1.4.2 (p = .088), 1.1.2, or 4.1.2 (p > .999) (Table 4). The first null hypothesis is rejected for three indicators.

Figure 1

Self-Assessed and Observed Need on the Six Indicators



Note. Self-assessed need: e-SAT capability Low or Moderate. Observed need: PMES rating of 3 or lower. *McNemar's exact test, Holm-adjusted* p* < .05.

Where the two measures did flag a teacher, they seldom flagged the same one. Kappa ranged from .02 on 4.1.2 to .39 on 1.1.2, and the 95% interval of every kappa included zero. The four groups make the pattern plain (Figure 2): on differentiated instruction, 10 teachers had a need both measures saw, 45 had a need only the observation saw, and 6 had a need only they saw. PABAK was higher than kappa where needs were rare (1.1.2, .85; 4.1.2, .67), showing that most of the agreement on those indicators was agreement that no need existed.

Table 4

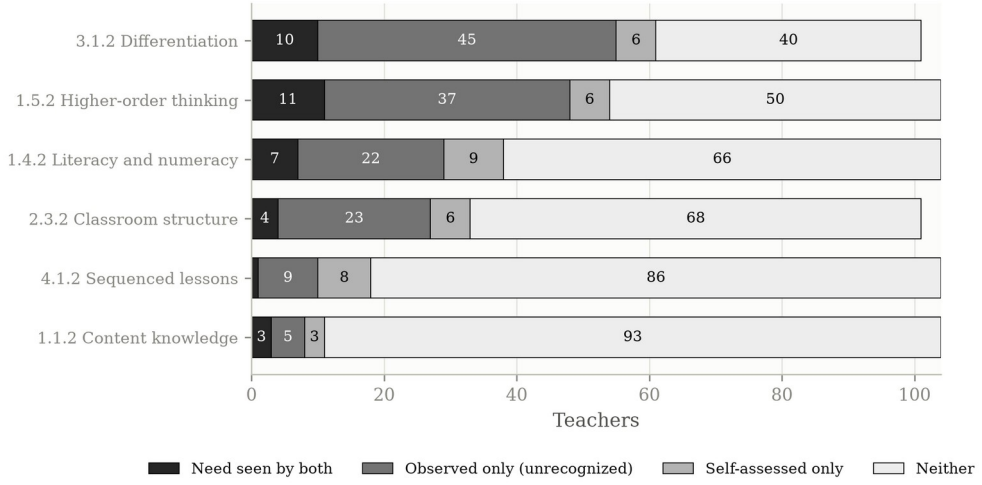
Agreement of Self-Assessed and Observed Need, by Indicator

Indicator	*n*	Both	Observed only	Self only	Neither	Holm *p*	κ	95% CI	PABAK	ρ
1.1.2	104	3	5	3	93	> .999	.39	[−.03, .71]	.85	.39
1.4.2	104	7	22	9	66	.088	.14	[−.06, .34]	.40	.11
1.5.2	104	11	37	6	50	< .001	.13	[−.02, .28]	.17	.22
2.3.2	101	4	23	6	68	.009	.08	[−.09, .27]	.43	.23
3.1.2	101	10	45	6	40	< .001	.05	[−.09, .18]	−.01	.29
4.1.2	104	1	9	8	86	> .999	.02	[−.12, .25]	.67	.34

Note. Counts of teachers. Holm *p*: McNemar's exact test, Holm-adjusted for six indicators. Kappa's interval is from 4,000 bootstrap samples of teachers. PABAK = prevalence-adjusted bias-adjusted kappa. ρ: Spearman's correlation of the four-level capability with the five-level rating.

Figure 2

Four Groups of Teachers on Each Indicator



Note. Counts of teachers observed on the indicator. Observed only: an observed need the teacher's e-SAT did not flag.

The pattern held in both sensitivity analyses. With an observed need set at a rating of 2 or lower, the observed shares fell to between 2.9% and 25.7% and no McNemar test remained significant, while kappa stayed between .09 and .28: the e-SAT missed needs at the stricter cut as well, but at that cut the observation flagged too few teachers for a test to separate the proportions. Among the 80 teachers whose e-SAT came before their observation, differentiated instruction and higher-order thinking still showed significant gaps ($p < .001$ and $p < .001$), so the gap is not an effect of feedback received before the self-assessment.

Which Observed Needs the e-SAT Did Not Flag

Of the 177 observed needs across the six indicators, 141 (79.7%) belonged to teachers who had rated the indicator High or Very High; 69 teachers had at least one such need. The share fell with experience: 92.7% among teachers with one to four years, 84.1% with five to nine, 77.3% with

ten to 19, and 47.8% with 20 or more. In the GEE model, each band of experience multiplied the odds that a need went unflagged by 0.41, 95% CI [0.21, 0.84], p = .014, holding career stage and indicator constant; career stage itself was not related, OR = 2.26, 95% CI [0.46, 11.00], p = .314. With years alone in the model the odds ratio was 0.49, p = .015. The second null hypothesis is rejected for years in teaching and retained for career stage. Only 23 observed needs belonged to teachers with 20 or more years, so the last band's share is uncertain.

How Far One Observation Can Be Trusted

The raters' groups of ratees differed in their mean ratings, from 3.30 to 4.33, and in their share of observed needs, from 7.7% to 45.2%. Once the teachers were modeled, the raters accounted for 3.5% of the variance in ratings, the teachers for 46.8%, and the indicator within the lesson for the rest. Since each teacher had one rater, this share cannot separate a rater's leniency from real differences between the teachers each rater happened to have; it shows only that the raters' groups did not differ much beyond their teachers.

For the 33 teachers observed twice, the two ratings on the three repeated indicators were identical in 65.7% of 99 pairs and within one point in 96.0%, weighted κ = .74, 95% CI [.60, .83]. Still, 11 of the 99 pairs (11.1%) crossed the line between a need and no need from one lesson to the next. Both observations were usually by the same rater, so this measures how much a teacher's lessons differ, not how much raters disagree.

What the Observers' Notes Recorded

The notes ran from 35 to 640 words, median 188; the median by rater ranged from 51 to 430 words, and 5 notes had fewer than 60. Eight of the ten codes met the threshold of .800 set in the proposal (Table 5). Two did not reach even the tentative threshold of .667: "recall questions only" (α = .651) and "one task for the whole class" (α = .578). Both record the absence of something, which a coder must infer from what a note leaves out, and their disagreements (18 and 22 notes) arose mostly from short notes. Their associations with the ratings are reported in Table 5 but support no conclusion.

Table 5

Practices Recorded in the Notes and the Indicator's Ratings With and Without Them

Code	Practice (indicator)	Recorded	Krippendorff's α	*M* with	*M* without	*r*rb	*p* (Holm)
Q1	Asked why or how questions (1.5.2)	44.2%	.922	3.93	2.90	.55	< .001
Q2	Recall questions only (1.5.2)	46.2%	.651	2.81	3.82	−.52	< .001
Q3	Learners explained or justified answers (1.5.2)	31.7%	.910	3.73	3.18	.30	.030
D1	One task for the whole class (3.1.2)	51.9%	.578	2.67	3.70	−.51	< .001
D2	Tiered tasks or choices (3.1.2)	26.9%	.877	3.75	2.92	.42	.004
D3	Helped learners one by one during seatwork (3.1.2)	52.9%	.943	3.45	2.78	.32	.020
G1	Group work with assigned roles (2.3.2)	40.4%	.820	4.21	3.51	.39	.002
G2	Learners handled materials (2.3.2)	31.7%	.826	4.18	3.62	.30	.029
L1	Explicit reading or vocabulary strategy (1.4.2)	46.2%	.884	4.04	3.59	.24	.045
L2	Numeracy or computation strategy (1.4.2)	37.5%	.855	4.08	3.63	.23	.045

Note. Recorded: share of the 104 first-observation notes with the practice, as agreed by the two coders. *M*: mean PMES rating on the code's indicator in lessons with and without the

practice. *r*rb: rank-biserial correlation from the Mann-Whitney test. Codes Q2 and D1 did not reach α = .667 and support no conclusion.

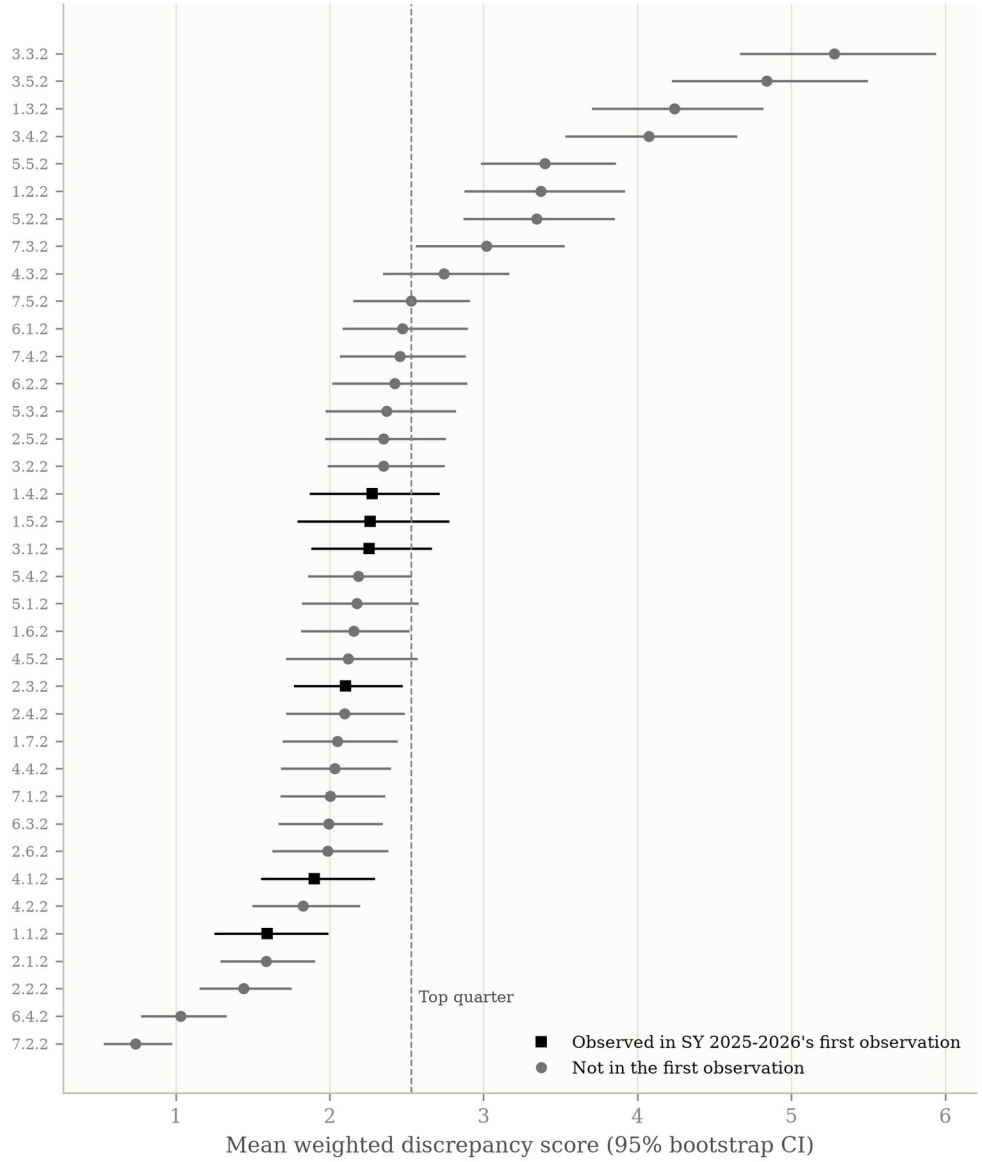
Among the reliable codes, the practices that separated higher from lower ratings were the ones the rubric describes. On higher-order thinking, lessons whose notes recorded why or how questions were rated 3.93 on average against 2.90 without them (*r*rb = .55), and lessons in which learners explained or justified answers, 3.73 against 3.18. On differentiated instruction, tiered tasks or choices were recorded in only 26.9% of lessons, and those lessons were rated 3.75 against 2.92 (*r*rb = .42); help given to learners one by one was more common (52.9%) and went with a smaller difference. On classroom structure, group work with assigned roles went with 4.21 against 3.51 (*r*rb = .39). The third null hypothesis is rejected for all eight reliable codes; the literacy and numeracy codes showed the smallest differences (*r*rb = .24 and .23, $p = .045$ after adjustment).

Which Needs to Address First

The e-SAT's own priorities did not point to the needs the observations found (Figure 3). Ranked by the mean weighted discrepancy score, the top of the e-SAT's list was 3.3.2 (learners with disabilities, giftedness, and talents), 3.5.2 (learners from Indigenous groups), 1.3.2 (ICT), and 3.4.2 (learners in difficult circumstances), with ranks that the bootstrap held within one or two places. Differentiated instruction, the most common observed need, ranked 19th of 37, and higher-order thinking 18th, with rank ranges of 11 to 28 and 10 to 30. The ranked discrepancy model gave nearly the same order, ρ = .94, with nine of the top ten indicators in common.

Figure 3

e-SAT Development Priorities for the 37 PPST Indicators



Note. Mean weighted discrepancy score = (4 − capability) × the indicator's mean priority for development, averaged over teachers; 95% intervals from 4,000 bootstrap samples of teachers. Squares: the six indicators of the first observation. The dashed line marks the top quarter.

The rule set in the proposal gave each observed indicator its form of assistance (Table 6). Two indicators, differentiated instruction and higher-order thinking, crossed the 40% line for school-wide sessions; literacy and numeracy strategies and classroom structure fell in the small-group band;

content knowledge and sequenced lessons called for individual coaching. Of the e-SAT's top-quarter indicators, the first four are classroom observable indicators that SY 2025–2026 did not observe. Two of them, 3.5.2 and 3.3.2, depend on whether a teacher has such learners, which the e-SAT does not ask.

Table 6

Forms of Technical Assistance Set by the Proposal's Rule

Indicator	Teachers with an observed need	Share	Of them, not flagged in the e-SAT	Form of assistance
3.1.2 Differentiated instruction	55 of 101	54.5%	45	School-wide sessions
1.5.2 Higher-order thinking	48 of 104	46.2%	37	School-wide sessions
1.4.2 Literacy and numeracy strategies	29 of 104	27.9%	22	Small-group coaching
2.3.2 Classroom structure and hands-on work	27 of 101	26.7%	23	Small-group coaching
4.1.2 Sequenced teaching and learning	10 of 104	9.6%	9	Individual coaching (IS Plan)
1.1.2 Content knowledge	8 of 104	7.7%	5	Individual coaching (IS Plan)

Note. Rule set in the proposal: 40% or more of observed teachers with a need, school-wide sessions; 15% to 39%, small-group coaching; under 15%, individual coaching in the rater's IS Plan.

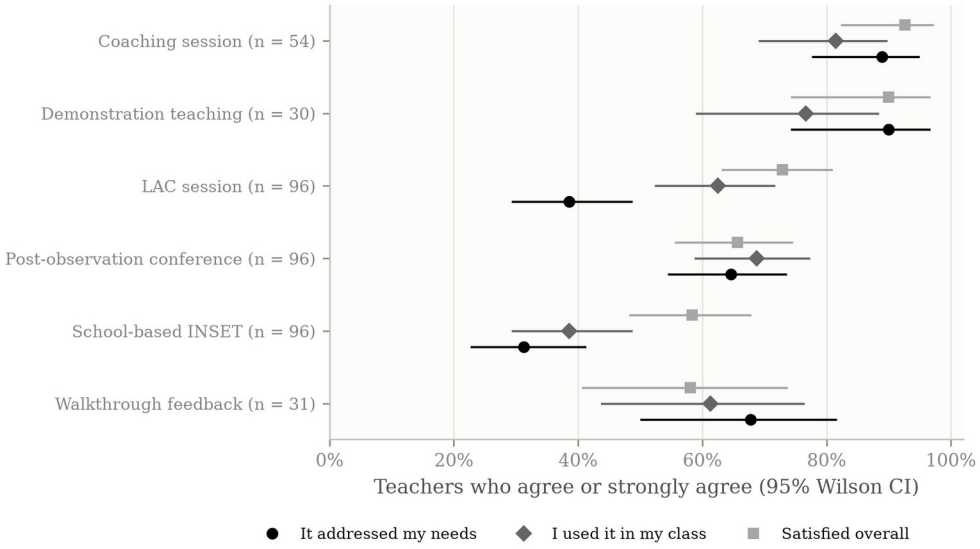
How Teachers Rated the Assistance They Received

All 96 respondents had a post-observation conference, attended LAC sessions, and attended school-based in-service training (INSET); 54 had a coaching session with their rater, 30 watched a demonstration lesson, and 31 received feedback after a walkthrough. Overall, 70.5% of the ratings of assistance received were Agree or Strongly Agree that the teacher was satisfied. The forms differed widely (Figure 4). Coaching sessions (92.6%

satisfied) and demonstration lessons (90.0%) rated highest; school-based INSET (58.3%) and walkthrough feedback (58.1%) lowest. The weakest single item was whether the help addressed the teacher's needs: 38.5% agreed for LAC sessions and 31.3% for INSET, against 88.9% for coaching.

Figure 4

Satisfaction With Each Form of Technical Assistance Received



Note. Share of teachers who received the form and agreed or strongly agreed, with 95% Wilson intervals.

The post-observation conference was held a median of 9 days after the observation (interquartile range 3 to 22). DM 089, s. 2025 asks for it "immediately after the observation or within three (3) days thereafter" (DepEd, 2025a, item 57) without saying whether the days are calendar or school days (school days counted from the SY 2025–2026 calendar; DepEd, 2025b): 27 of 104 conferences (26.0%) met three calendar days, and 40 (38.5%) met three school days. Teachers whose conference came within three days were satisfied with it in 92.6% of cases, against 55.1% when it came later (Fisher's exact

test, $p < .001$); each five days of delay multiplied the odds of satisfaction by 0.69, 95% CI [0.57, 0.84]. The fourth null hypothesis is rejected.

For SY 2026–2027, teachers most often ranked demonstration lessons first (24 first choices, 137 Borda points), then peer observation (22, 130) and coaching (19, 116); school-based INSET received 35 points. Asked what help would most improve their teaching, 39 teachers (40.6%) asked to see the practice done, 30 (31.3%) asked for feedback specific to their class, 16 for time, and 14 for help with particular learners. The coders' alpha ranged from .557 (materials) to .812; the materials category did not reach .667 and is not interpreted.

Discussion

The e-SAT and the observation measured different things, and the school needed both. Teachers' self-ratings were high and their priorities sensible: they wanted help where they felt weakest, with learners they seldom meet and with tools they lack. The observations found something else, a large and shared gap in differentiation and higher-order questioning, and the teachers' own ratings missed four in five of the needs the observers saw. This is the pattern the national study and the Mindanao observations found (Molina et al., 2018; World Bank, 2016), and it is consistent with Van Zoest et al. (2002), who found self-perceptions furthest from observation where practice was weakest. The fall in unflagged needs with experience fits the same account, though in one school with few senior teachers it is a finding to recheck, not a rule. A plan built from the e-SAT's consolidation alone would

have sent the school's training to ICT and inclusion, and left its most common observed needs unaddressed.

The observations were not a perfect yardstick either. One lesson decides a rating, and about one in nine need flags changed between two lessons of the same teacher; the raters' groups differed in their means, though little beyond their teachers. The study therefore used the observations to set where the school should act, not to judge any teacher, and gave the individual decision back to the rater and the teacher in the IS Plan, where the teacher's own lessons are the evidence. The coded notes added what neither rating could: which practices the higher-rated lessons contained. Questions that ask why or how, learners explaining their answers, tiered tasks, and group work with roles are concrete, observable, and teachable, and they give the sessions and the coaching their content. That the two codes recording an absence were unreliable is itself useful: an observer who wants a note to support feedback should write what happened, not what did not.

The satisfaction results point to the form the assistance should take. Teachers valued coaching and seeing a practice done, and rated whole-staff sessions lowest on fit, the same complaint EDCOM II heard nationally (EDCOM II, 2026, p. 288). Timeliness mattered: a conference within three days satisfied nearly everyone, and most conferences came later. Pairing group sessions with coaching and with demonstrations by colleagues who already use the practice follows the evidence on coaching and peer pairing (Kraft et al., 2018; Papay et al., 2020).

Recommendations

1. **Adopt the technical assistance plan in Appendix D for SY 2026–2027.** It sets school-wide LAC sessions on differentiated instruction and higher-order thinking, small-group coaching on literacy and numeracy strategies and on classroom structure, and individual coaching on content knowledge and sequenced lessons, each with its success indicator and means of verification, and puts the school-wide sessions in the INSET days and Learning Action Cell sessions of the three-term calendar (DepEd, 2026c). The raters copy its rows into their monthly IS Plans.
2. **Hold every post-observation conference within three calendar days**, scheduled in the IS Plan when the observation is scheduled, and record its date in the PMCF.
3. **Make walkthrough feedback specific.** Raters use the reliable codes of the codebook as look-fors and write in the PMCF what was seen. Since 3.1.2 and 2.3.2 are not rated again in the PMES cycle (DepEd, 2025a, Table 7), walkthroughs are how the school will know whether the assistance worked.
4. **Read e-SAT results with the teachers' classes in view.** Before planning assistance on 3.3.2, 3.4.2, or 3.5.2, the school should check which teachers actually teach such learners.
5. **Keep observation notes as evidence.** Notes should record what the teacher and learners did; ratings should be read with their notes.

6. **For the Schools Division Office:** consolidate e-SAT results together with observation results, not alone, when planning division training, and repeat this comparison in other schools before drawing division-wide conclusions.
7. **For further research:** a study that pairs a second, independent observer on a sample of lessons would separate rater leniency from teacher differences; a follow-up in SY 2027–2028 can test whether the targeted assistance changed the observed practices.

Dissemination and Advocacy Plans

The results go first to the people who produced the data (Table 7). DepEd Order No. 16, s. 2017 asks researchers to "disseminate and discuss their research results and recommendations in the area/office/governance level where the study was conducted, preferably attended by the respondents" (DepEd, 2017, Sec. V.B.viii).

Table 7

Dissemination and Advocacy Plan

Audience	What they receive	How	When
The school's teachers	The findings for groups, the codebook's practices, the plan	A LAC session in the Instructional Block	June 26, 2026
The fourteen raters	The plan's rows for their IS Plans; the codebook as walkthrough look-fors; the three-day conference target	A planning meeting	June 22, 2026
The School Governing Council	A one-page summary	The council's regular meeting	Term 1 of SY 2026–2027
The Schools Division Office	The report, the executive data set, a two-page brief	Submission to the Schools Division Research Committee; the division's research forum if invited	July 2026
Other school heads of the district	How to compare e-SAT and observation records with consent	The district school heads' meeting	Term 1 of SY 2026–2027

Note. Presentation in research conferences and forums, the division's included, and publication need the written permission of the National and Regional Research Committees (DepEd, 2017, Sec. VI.B).

References

Anti-Red Tape Authority. (2022). *Guidelines on the implementation of the harmonized client satisfaction measurement* (Memorandum Circular No. 2022-05). https://arta.gov.ph/wp-content/uploads/2022/09/MC-2022-05-GUIDELINES-ON-THE-IMPLEMENTATION-OF-THE-HARMONIZED-CLIENT-SATISFACTION-MEASUREMENT.pdf

Anti-Red Tape Authority. (2023). *Amendment to the guidelines on the implementation of the harmonized client satisfaction measurement* (Memorandum Circular No. 2023-05). https://arta.gov.ph/wp-content/uploads/2023/06/MC-2023-05_Amendment-to-CSM-1-1.pdf

Borich, G. D. (1979). *A needs assessment model for conducting follow-up studies of teacher education and training*. Research and Development Center for Teacher Education, The University of Texas at Austin. https://files.eric.ed.gov/fulltext/ED189079.pdf

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*(5), 423–429. https://doi.org/10.1016/0895-4356(93)90018-V

Camburn, E., & Barnes, C. A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal, 105*(1), 49–73. https://doi.org/10.1086/428802

Department of Education. (2016). *Adoption of the Basic Education Research Agenda* (DepEd Order No. 39, s. 2016).

https://www.deped.gov.ph/wp-content/uploads/2016/06/DO_s2016_039.pdf

Department of Education. (2017). *Research management guidelines* (DepEd Order No. 16, s. 2017). https://www.deped.gov.ph/wp-content/uploads/2017/03/DO_s2017_016.pdf

Department of Education. (2019). *Implementation of the NEAP transformation* (DepEd Order No. 011, s. 2019). https://www.deped.gov.ph/wp-content/uploads/2019/05/DO_s2019_011.pdf

Department of Education. (2020). *National adoption and implementation of the Philippine Professional Standards for School Heads* (DepEd Order No. 024, s. 2020). https://www.deped.gov.ph/wp-content/uploads/2020/09/DO_s2020_024-.pdf

Department of Education. (2022). *Adoption of the Basic Education Development Plan 2030* (DepEd Order No. 024, s. 2022). https://www.deped.gov.ph/wp-content/uploads/2022/05/DO_s2022_024.pdf

Department of Education. (2024). *Guidelines on the school-level development and implementation of school-based professional development (PD) programs* (Enclosure No. 1 to Memorandum DM-OUHROD-2024-1576).

Department of Education. (2025a). *Guidelines on the multi-year Performance Management and Evaluation System for teachers from school years*

2025–2026 to 2027–2028 (DepEd Memorandum No. 089, s. 2025). https://www.deped.gov.ph/wp-content/uploads/DM_s2025_089r.pdf

Department of Education. (2025b). *Multi-year implementing guidelines on the school calendar and activities* (DepEd Order No. 012, s. 2025). https://www.deped.gov.ph/wp-content/uploads/DO_s2025_012.pdf

Department of Education. (2026a). *Foundational guidelines on artificial intelligence in basic education* (DepEd Order No. 003, s. 2026). https://www.deped.gov.ph/wp-content/uploads/DO_s2026_003.pdf

Department of Education. (2026b). *Guidance in the implementation of the Performance Management and Evaluation System (PMES) for teachers for SY 2025–2026* (Memorandum DM-OSEC-OUHRODI-2026-0001).

Department of Education. (2026c). *Guidelines on the implementation of the three-term school calendar in basic education* (DepEd Order No. 009, s. 2026). https://www.deped.gov.ph/wp-content/uploads/DO_s2026_009r.pdf

Desimone, L. M., & Garet, M. S. (2015). Best practices in teachers' professional development in the United States. *Psychology, Society, & Education, 7*(3), 252–263. https://doi.org/10.25115/psye.v7i3.515

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs: Principles and practices. *Health Services*

Research, 48(6 Pt 2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3), 613–619. https://doi.org/10.1177/001316447303300309

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence, 80*, Article 101449. https://doi.org/10.1016/j.intell.2020.101449

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (MET Project Research Paper). Bill & Melinda Gates Foundation. https://files.eric.ed.gov/fulltext/ED540957.pdf

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (MET Project Research Paper). Bill & Melinda Gates Foundation. https://files.eric.ed.gov/fulltext/ED540960.pdf

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*(2), 180–188. https://doi.org/10.1037/0022-3514.82.2.180

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29–45. https://doi.org/10.3102/01623737021001029

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

Molina, E., Fatima, S. F., Trako, I., & Wilichowski, T. (2018). *Teacher practices in Mindanao: Results of the Teach classroom observation study*. World Bank. https://documents1.worldbank.org/curated/en/604141549910104822/pdf/134535-WP-PUBLIC-11-2-2019-11-50-40-TEACHPhilippinesReportSubmissiontoTeam.pdf

Narine, L. K., & Harder, A. (2021). Comparing the Borich model with the Ranked Discrepancy Model for competency assessment: A novel approach. *Advancements in Agricultural Development, 2*(3), 96–111. https://doi.org/10.37433/aad.v2i3.169

National Privacy Commission. (2023). *Guidelines on consent* (NPC Circular No. 2023-04). https://privacy.gov.ph/wp-content/uploads/2023/11/NPC-Circular-No.-2023-04_Guidelines-on-Consent_07Nov2023.pdf

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 1–13. https://doi.org/10.1177/1609406919899220

Oakley, G., King, R., & Scarparolo, G. (2018). *An evaluation of ELLN Digital: Technology-supported teacher professional development on early language, literacy, and numeracy for K-3 teachers*. Foundation for

Information Technology Education and Development. https://dl4d.org/wp-content/uploads/2019/03/ELLN-Digital-Evaluation.pdf

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy, 12*(1), 359–388. https://doi.org/10.1257/pol.20170709

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12–40.

Philippine Health Research Ethics Board. (2022). *National ethical guidelines for research involving human participants 2022*. Philippine Council for Health Research and Development.

Rivera, J. P. R., Sinsay-Villanueva, L. M. V., Tanyag, I. H., Berroya, J. D., Garcia, G. D. V., & Lim, V. L. (2025). *Revitalizing the Philippine education system: Facilitating access and participation to in-service training (INSET) and teacher professional development (TPD)* (PIDS Discussion Paper No. 2025-14). Philippine Institute for Development Studies. https://pidswebs.pids.gov.ph/CDN/document/pidsdps2514.pdf

Ross, J. A., & Bruce, C. D. (2007). Teacher self-assessment: A mechanism for facilitating professional growth. *Teaching and Teacher Education, 23*(2), 146–159. https://doi.org/10.1016/j.tate.2006.04.035

RTI International. (2020). *ABC+: Advancing Basic Education in the Philippines—Baseline report*. United States Agency for International Development.

Second Congressional Commission on Education. (2026). *Turning point: A decade of necessary reform (2026–2035). EDCOM II final report*. https://edcom2.gov.ph/media/2026/01/EDCOM-II-Y03-for-web-012626_final.pdf

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., Van Herwegen, J., & Anders, J. (2025). Effective teacher professional development: New theory and a meta-analytic test. *Review of Educational Research, 95*(2), 213–254. https://doi.org/10.3102/00346543231217480

Torres, M. V., Zerrudo, A. P., Labad, V. S., & Gabales, B. G., Jr. (2024). Teachers' views on the conduct of class observation: The Philippine DepEd setting. *International Journal of Instruction, 17*(3), 453–474. https://doi.org/10.29333/iji.2024.17325a

Van Zoest, L. R., Breyfogle, M. L., & Ziebarth, S. W. (2002). Self-perceived and observed practices of secondary school mathematics teachers. *Teacher Development, 6*(2), 245–268. https://doi.org/10.1080/13664530200200167

Wiggs, N. B., Reddy, L. A., Kettler, R. J., Hua, A. N., Dudek, C. M., Lekwa, A. J., & Bronstein, B. (2023). Convergence between teacher self-report and school administrator observation ratings using the Classroom

Strategies Assessment System. *Assessment for Effective Intervention, 48*(2), 113–123. https://doi.org/10.1177/15345084221112858

World Bank. (2016). *Assessing basic education service delivery in the Philippines: The Philippines Public Education Expenditure Tracking and Quantitative Service Delivery Study* (Report No. AUS6799). https://documents1.worldbank.org/curated/en/507531468325807323/pdf/AUS6799-REVISED-PH-PETS-QSDS-Final-Report.pdf

Financial Report

The study was funded by a school-level grant of the Basic Education Research Fund. BERF is output-based: the tranches were released on the deliverables, which serve as the liquidation documents (DepEd, 2017, Sec. V.C.i.5). No equipment, software, or honorarium was charged, since DO 16 makes none eligible; the coders gave their time through the district's partnership between schools.

Table 8

Financial Report

Item	Approved	Spent
Supplies and materials (consent forms, envelopes, coding sheets)	2,400.00	2,310.00
Reproduction, printing, and binding (questionnaire copies for teachers without devices, report copies)	3,600.00	3,480.00
Communication (prepaid load for the custodian, the coders, and the online questionnaire)	1,500.00	1,500.00
Domestic travel (the coders' fares between schools, six coding days)	2,880.00	2,880.00
Food during the coding sessions and the consent sessions	4,620.00	4,560.00
Dissemination (the LAC session and the raters' meeting)	3,000.00	2,950.00
Total	18,000.00	17,680.00

Note. In Philippine pesos. First tranche (80%), PhP 14,400.00, on the work plan and MOA; last tranche (20%), PhP 3,600.00, on the final report (DepEd, 2017, Sec. V.C.i.5).

Appendices

Appendix A. Consent Form for the Research Use of PMES Records

The form, in English and Filipino, named the records to be read (the e-SAT 2025–2028, the first observation's rating sheet and notes, the date of the post-observation conference), who would see them coded and who would not, how long the key would be kept, and that refusing would change nothing in the teacher's PMES. The research coordinator read it aloud in each session and collected it in a sealed envelope.

Appendix B. The Technical Assistance Questionnaire

Before you begin

This questionnaire is part of a study of the help our school gives teachers to improve their teaching. It asks about the technical assistance you received in School Year 2025–2026, how you rate it, and what you want for SY 2026–2027. It takes about ten minutes.

Answering is voluntary. You may skip any question or stop at any time, and choosing not to answer changes nothing in your performance rating, your assignments, or your standing in the school. The form carries only the code the data custodian gave you. No researcher, rater, or school head will see your answers with your name, and results are reported only for groups of teachers.

By continuing to Part A, you agree to take part.

Part A. The help you received in SY 2025–2026

For each form of help below, first say whether you received it this school year. If you did, rate each statement: 1 Strongly disagree, 2 Disagree, 3 Neither agree nor disagree, 4 Agree, 5 Strongly agree, or N/A if it does not apply.

A1. A post-observation conference after your full-period classroom observation

- ○ Received: Yes / No
- ○ The help came when I needed it. (1 2 3 4 5 N/A)
- ○ The help addressed the needs we had identified. (1 2 3 4 5 N/A)
- ○ The person who helped me knew the subject and the class. (1 2 3 4 5 N/A)
- ○ I was treated with respect. (1 2 3 4 5 N/A)
- ○ I was able to use what I learned in my class. (1 2 3 4 5 N/A)
- ○ I am satisfied with the help I received. (1 2 3 4 5 N/A)

A2. A coaching session with your rater (other than the post-observation conference)

- ○ Received: Yes / No
- ○ The same six statements as A1.

A3. Feedback after a short walkthrough of your class

- ○ Received: Yes / No
- ○ The same six statements as A1.

A4. Learning Action Cell (LAC) sessions

- ○ Received: Yes / No
- ○ The same six statements as A1.

A5. A demonstration lesson by a colleague that you watched

- ○ Received: Yes / No
- ○ The same six statements as A1.

A6. School-based in-service training (INSET)

- ○ Received: Yes / No
- ○ The same six statements as A1.

Part B. What you want in SY 2026–2027

B1. Which three forms of help would most improve your teaching next school year? Write 1 beside your first choice, 2 beside your second, and 3 beside your third.

- ○ ___ Post-observation conferences held soon after the observation
- ○ ___ Coaching sessions with my rater
- ○ ___ Walkthroughs with feedback
- ○ ___ LAC sessions
- ○ ___ Demonstration lessons by colleagues
- ○ ___ School-based INSET
- ○ ___ Observing a colleague's class and being observed by a colleague (peer observation)

Part C. In your own words

C1. What help would most improve your teaching this year?

_____

_____

Thank you. Seal the printed form in the envelope provided, or submit the online form.

Appendix C. Codebook for the Observation Notes

Rules for every code

The unit is one first-observation notes form, with names, sections, and dates removed. Code 1 when the note records the practice; code 0 when it does not. Code only what the note records: a coder never infers a practice the note does not describe, and never uses what they know of the teacher, the subject, or the rating, which the coders do not see. A practice recorded once is enough. A note may receive any number of codes.

Two codes record an absence, Q2 and D1. Code them 1 only when the note records the lesson's questions or tasks fully enough to show the absence: for Q2, at least two of the teacher's questions are written and every one asks for recall; for D1, the note describes the seatwork or activity and every learner had the same task. A short note that does not describe the questions or the tasks is coded 0 on both.

The codes

Code	Indicator	Code 1 when the note records	Example in a note	Not this
Q1	1.5.2 Higher-order thinking	The teacher asked at least one question that asks why or how, or asks learners to compare, predict, or decide	"Asked: Why did the farmer choose the earlier planting date?"	A question with one factual answer, even if it begins with "what"
Q2	1.5.2	Every question recorded asks for recall, and at least two are recorded	"Questions: What is the capital? Who wrote the poem? When did it happen?"	A note that records no questions
Q3	1.5.2	Learners explained or justified an answer, aloud or in writing	"L3 explained why 3/4 is greater than 2/3 using the drawing"	A learner reading an answer aloud
D1	3.1.2 Differentiated instruction	The seatwork or activity is described and every learner had the same task	"All learners answered items 1 to 10 on the board"	A note that does not describe the seatwork
D2	3.1.2	Learners worked on tasks set at different levels, or chose among tasks	"Three sets of activity cards: A for those who still need the steps, C for those who finished"	Groups doing the same task
D3	3.1.2	The teacher helped learners one by one or in a small group during seatwork	"Teacher sat with two learners who had not started and modeled the first item"	Checking answers at the end of the lesson
G1	2.3.2 Classroom structure	Group work in which learners had assigned roles	"Roles: reader, recorder, reporter, timekeeper"	Groups without roles
G2	2.3.2	Learners handled materials, models, or tools themselves	"Each pair measured the leaves with a ruler"	The teacher demonstrating with materials
L1	1.4.2 Literacy and numeracy	An explicit reading, vocabulary, or writing strategy was taught or used	"Pre-taught three terms with pictures; learners used a word map"	Learners reading a text aloud without a strategy
L2	1.4.2	An explicit numeracy or computation strategy was taught or used	"Showed estimation before computing the total"	Computation without a strategy named or modeled

Agreement

Both coders code every note independently. Krippendorff's alpha is computed for each code. A code with alpha of at least .800 supports a conclusion; one between .667 and .800 supports a tentative one; one below .667 is reported and not interpreted. Disagreements are settled by the two coders rereading the note together, and the agreed code is used.

Appendix D. Technical Assistance Plan for SY 2026–2027

School: [Name of school], [Schools Division] (blind copy)

Teachers covered: The 114 Teachers I to VII; the counts below are of the 104 whose records the study read

Based on: The school's study of its SY 2025–2026 e-SAT, first full-period observations, observation notes, PMCF dates, and technical assistance questionnaire

Columns: The first six of the Instructional Supervisory Plan (DM 089, s. 2025, Annex B), with the teachers each row covers and the persons responsible; each rater copies the rows for their ratees into their monthly IS Plan, which adds Accomplishment/Status and Remarks

Prepared by: The School Principal, with the fourteen raters, June 22, 2026

The plan names indicators and numbers of teachers, never a teacher. Which teacher joins which group is settled by each rater with the teacher, in the teacher's Individual Development Plan and the rater's IS Plan (DM 089, s. 2025, items 26 to 30). School-wide needs are those of 40% or more of observed teachers; small-group needs, 15% to 39%; individual needs, under 15%.

Teaching and learning: differentiated instruction (PPST 3.1.2)

Objective: Teachers set tasks at different levels or offer choices, and help learners one by one during seatwork

Teachers: School-wide: 55 of 101 observed teachers (54.5%) had a need; 45 of them had rated themselves High or Very High

Enabling projects and activities: Four LAC sessions in the Instructional Block on designing one tiered task from the coming week's lesson; two demonstration lessons by teachers whose lessons already use tiered tasks, each followed by a LAC debrief; peer-observation pairs across departments

Success indicators: Walkthrough entries recording tiered tasks or choices: from 26.9% of first-observation notes to 50% of walkthroughs by the end of Term 3

Means of verification: PMCF walkthrough entries, coded with the study's codebook; LAC outputs (the tiered tasks)

Time frame: Terms 1 to 3; LAC sessions in Term 1

Persons responsible: Department heads and master teachers (raters); the assistant school principal (LAC); the principal (approval)

Teaching and learning: higher-order thinking (PPST 1.5.2)

Objective: Teachers ask why and how questions and have learners explain or justify their answers

Teachers: School-wide: 48 of 104 (46.2%); 37 of them unflagged in the e-SAT

Enabling projects and activities: Three LAC sessions on planning questions for a lesson; one day of the September 10–11, 2026 INSET on questioning; demonstration lessons; peer observation

Success indicators: Teachers rated 3 or lower on 1.5.2 in the first observation of SY 2026–2027: from 46.2% to 30% or less; notes recording why or how questions: from 44.2% to 65%

Means of verification: COT rating sheets (1.5.2 is rated in both observations of SY 2026–2027); observation notes, coded

Time frame: Terms 1 and 2; INSET in the End-of-Term Block

Persons responsible: Raters; the assistant school principal; INSET proposal to the SDO 45 working days before

Teaching and learning: literacy and numeracy strategies (PPST 1.4.2)

Objective: Teachers of every subject teach an explicit reading or numeracy strategy their lesson needs

Teachers: Small groups: 29 of 104 (27.9%)

Enabling projects and activities: Coaching cycles in groups of four to six across departments (observe, debrief, try, observe again), one group per master teacher

Success indicators: Teachers rated 3 or lower on 1.4.2 in the first observation of SY 2026–2027: from 27.9% to 15% or less

Means of verification: COT rating sheets (1.4.2 is rated in both observations of SY 2026–2027); PMCF coaching entries

Time frame: Terms 1 and 2

Persons responsible: Master teachers

Teaching and learning: classroom structure and hands-on work (PPST 2.3.2)

Objective: Teachers run group work with assigned roles and let learners handle the materials

Teachers: Small groups: 27 of 101 (26.7%)

Enabling projects and activities: Coaching cycles in groups of four to six; one demonstration of group roles per department

Success indicators: Walkthrough entries recording group work with roles: from 40.4% of first-observation notes to 60%

Means of verification: PMCF walkthrough entries, coded (2.3.2 is not rated again in the PMES cycle)

Time frame: Terms 1 and 2

Persons responsible: Department heads

Teaching and learning: sequenced lessons and content knowledge (PPST 4.1.2, 1.1.2)

Objective: Each teacher with a need on these indicators improves on them

Teachers: Individual: 10 on 4.1.2 and 8 on 1.1.2

Enabling projects and activities: Individual coaching in each rater's IS Plan; for 1.1.2, by a subject specialist; for 4.1.2, through the lesson plan review

Success indicators: Each teacher's IDP names the indicator, with at least one coaching entry per term

Means of verification: IDPs; PMCF; lesson plan reviews

Time frame: Terms 1 to 3

Persons responsible: Each teacher's rater

Teaching and learning: inclusion and ICT (PPST 3.3.2, 3.4.2, 3.5.2, 1.3.2)

Objective: The e-SAT's top priorities are confirmed before the school acts on them

Teachers: To be confirmed: the e-SAT's top four needs; 3.5.2 and 3.3.2 depend on whether a teacher has such learners

Enabling projects and activities: List the teachers who teach learners with disabilities, learners in difficult circumstances, and learners from Indigenous groups from the class records; then plan sessions for them

Success indicators: The list made by the end of Term 1; sessions for listed teachers by Term 2

Means of verification: The list; session records

Time frame: Terms 1 and 2

Persons responsible: The principal; the guidance designate; raters

Instructional supervision: post-observation conferences

Objective: Every conference is held within three days of the observation (DM 089, s. 2025, item 57)

Teachers: All teachers

Enabling projects and activities: Each rater schedules the conference with the observation in the IS Plan

Success indicators: Conferences within three calendar days: from 26.0% to 90%

Means of verification: PMCF dates

Time frame: Each observation

Persons responsible: Raters

Instructional supervision: walkthrough feedback

Objective: Every teacher receives specific written feedback after a walkthrough each term

Teachers: All teachers

Enabling projects and activities: Raters use the codebook's practices as look-fors and write what they saw in the PMCF

Success indicators: Teachers satisfied with walkthrough feedback: from 58.1% to 80%

Means of verification: The questionnaire, May 2027; PMCF entries

Time frame: Each term

Persons responsible: Raters

Monitoring and evaluation of this plan

Objective: The school knows whether the assistance worked

Teachers: All teachers

Enabling projects and activities: Repeat the questionnaire in May 2027; compare SY 2026–2027 ratings on 1.4.2 and 1.5.2 with the baselines; code a sample of walkthrough entries

Success indicators: "The help addressed the needs we had identified," LAC sessions: from 38.5% to 70%

Means of verification: The questionnaire; COT rating sheets; PMCF entries

Time frame: May 2027

Persons responsible: The research coordinator; the principal

The INSET days of SY 2026–2027 fall in the End-of-Term Blocks: September 10–11 and December 17–18, 2026, and April 2 and 5, 2027 (DepEd Order No. 009, s. 2026). LAC sessions are held within the Instructional Block.

Appendix E. The Custodian's Protocol

The custodian's written steps: receive the sealed consent forms; list the consenting teachers; draw a random code for each; copy the e-SAT answers, the first observation's ratings, the second observation's ratings on the three repeated indicators, and the conference date; replace each rater's name with a random code; redact names, sections, and dates from the notes; keep the key on a separate encrypted drive; delete it after the report's acceptance.

Appendix F. Declarations

The Declaration of Anti-Plagiarism and the Declaration of Absence of Conflict of Interest (DepEd, 2017, Annex 3), with a statement of the conflict the principal's role creates and how the protocol manages it; and the declaration of the use of artificial intelligence (DepEd, 2026a, Annex A): [name of the AI tool] was used to check the grammar of the report and the citation format of the references; no personal or sensitive information was entered into it.