

The Effectiveness of the Master Teachers' Classroom Observation Practices in Improving the Teaching and Learning Process in the Public Elementary Schools in Ragay I District, Division of Camarines Sur

Wilanie D. Badong¹

¹School of Graduate Studies, University of Northeastern Philippines

ARTICLE INFORMATION	ABSTRACT
Article history: Published: October 2026	The study described master teachers' classroom observation practices and their perceived effectiveness for teaching and learning in public elementary schools of Ragay I District, Division of Camarines Sur, during school year 2025–2026. A descriptive–evaluative–correlational survey involved 53 respondents: 19 master teachers and 34 teachers in 17 schools grouped by size. A modified questionnaire used 61 practice and 43 perceived-effectiveness items. Weighted means summarized six practice domains and four perceived-effectiveness domains; Kendall's W tested concordance of item ranks across big, medium, and small schools. Classroom observation practice averaged 4.44 (Much Employed), highest for classroom culture (4.49). Perceived effectiveness averaged 4.56 (Very Much Effective), highest for classroom management and learning environment (4.60). No domain's reported chi-square exceeded its .05 critical value: the six practice W values ranged from 0.21 to 0.57, and the four effectiveness W values from 0.13 to 0.46. The findings reflect respondent appraisals and school-size rank orders, not observed changes in teacher practice or learner achievement. The study points toward attention to curriculum implementation and assessment feedback, while direct outcome evidence and several instrument details remain to be established.
Keywords: Master Teachers Classroom Observation Instructional Supervision Teaching And Learning Ragay Kendall Concordance	

1. Introduction

Classroom observation can make instruction visible for reflection, feedback, and professional learning. Its value depends on what the observer attends to and what happens afterward. An observation that documents routines but does not help a teacher act on evidence of pupils' understanding may remain an evaluation event. O'Leary (2020) discusses observation as a practice with formative and evaluative purposes, while Hattie and Timperley (2007) distinguish the information a useful feedback process provides about goals, current performance, and next steps. These perspectives are relevant to the study's focus on master teachers' classroom observation in Ragay I District.

The study examined both the extent of observation practices and respondents' judgments of their effectiveness for the teaching and learning process. Practice domains covered classroom culture, instructional delivery, classroom organization, curriculum, assessment, and implementation. Perceived-effectiveness domains covered lesson planning and preparation, instructional strategies and delivery, classroom management and learning environment, and assessment and feedback. These are connected but distinct: a practice rating describes what respondents think master teachers do, while an effectiveness rating expresses an appraisal of how useful it is. Neither, on its own, measures a change in teaching or pupil achievement.

The Philippine Professional Standards for Teachers offer a policy frame for teaching practice and professional development (Department of Education [DepEd], 2017). School-based Learning Action Cells provide a formal mechanism for teachers to examine shared instructional concerns (DepEd, 2016). Neither policy guarantees that a particular observation conference is constructive. The study supplies a district-level profile of what master teachers and teachers reported, grouped by school size, and considers where subsequent observation and feedback processes may need closer assessment.

The research questions were: (1) how extensively were six sets of classroom observation practices employed, (2) did rankings across big, medium, and small school groups agree for each set, (3) how effective were four sets of observation-related teaching and learning processes perceived to be, (4) did the groups' rankings agree for these sets, and (5) what policy recommendations followed? This manuscript presents the reported means and tests, with analytical discussion of each table and a clear limit on causal interpretation.

2. Methods

2.1 Design, setting, and respondents

The study used a descriptive–evaluative–correlational design in 17 public elementary schools of Ragay I District, Division of Camarines Sur, for school year 2025–2026. It purposively selected the district, included all available master teachers according to the methods, and selected two teachers per school, one from Key Stage 1 and one from Key Stage 2, through a non-proportional stratified procedure. The respondent table lists 19 master teachers and 34 teachers (17 in each key-stage group), totaling 53. Schools were grouped as big (four schools), medium (four), and small (nine). A classification rule for school size was not supplied.

Table 1. Reported respondent composition by school-size group

School-size group	Schools	Master teachers	KS1 teachers	KS2 teachers	Total	Printed %
Big	4	13	4	4	21	40
Medium	4	4	4	4	12	22
Small	9	2	9	9	20	38
Grand total	17	19	17	17	53	100

Note. KS1 = Key Stage 1; KS2 = Key Stage 2. Values follow the study's respondent table; percentages are as printed and rounded. The big-school group includes thirteen of the nineteen master teachers, whereas the small-school group includes only two master teachers across nine schools. This uneven composition matters for comparison: responses from a teacher and a master teacher may reflect different experiences of the same observation process, and a school-size mean can mix those roles in different proportions. The study reports group means but does not separate them by respondent role or school. Consequently, group contrasts should not be interpreted as pure effects of school size.

2.2 Instrument, procedure, and statistics

The researcher-developed modified questionnaire had two parts. The table of specifications states 61 practice items: classroom culture (9), instructional delivery (10), organization (10), curriculum (11), assessment (11), and implementation (10). It states 43 perceived-effectiveness items: lesson planning/preparation (10), instructional strategies/delivery (10), classroom management/environment (11), and assessment/feedback (12), for 104 items total. Five-point response bands used 4.50–5.00 for Very Much Employed or Very Much Effective and 3.50–4.49 for Much Employed or Much Effective. Several percentages in the specification do not correspond neatly to the item counts; the scoring key should be confirmed.

The instrument underwent adviser and expert review and a dry run outside the study schools according to the methods. After district and school permission, the questionnaires were personally distributed, retrieved, and tabulated. The methods also mention seeking permission for informal interviews and observation, but no interview or independently observed data are presented in the results. The reported reliability coefficient is 0.5846 and a *t* value of 5.6769; the text attributes internal consistency to KR-21, a procedure suited to dichotomous scoring, while the questionnaire uses five categories. The pilot size, scoring, and domain-level reliability are not given, so reliability cannot be independently appraised from the available numbers.

Weighted means, ranks, frequencies, percentages, and a reported standard deviation were among the listed treatments. Kendall's coefficient of concordance *W* and a chi-square statistic assessed how similarly big, medium, and small school groups ranked the items within each domain at $\alpha = .05$. The results tables give three group means, *W*, chi-square, degrees of freedom, and critical values; they do not report exact *p* values. A nonsignificant concordance test fails to establish consistent item ranking across the three groups. It does not prove that groups hold opposite views, nor does a favorable effectiveness rating establish that observation caused improvement.

3. Results

3.1 Extent of classroom observation practices

Table 2. Summary of master teachers' classroom observation practice

Domain	Big Wx	Medium Wx	Small Wx	Overall Wx	Interpretation	Rank
Classroom culture	4.74	4.06	4.68	4.49	ME	1
Instructional delivery	4.68	4.03	4.73	4.48	ME	2
Classroom organization	4.59	4.07	4.63	4.43	ME	4
Curriculum	4.53	4.06	4.60	4.40	ME	5.5
Assessment	4.60	4.07	4.69	4.45	ME	3
Implementation Practices	4.57	4.03	4.60	4.40	ME	5.5
Average	4.62	4.05	4.66	4.44	ME	

Note. Wx = weighted mean; ME = Much Employed. Overall means and rank positions reproduce Table 9; big and small group averages are labeled VME, and the medium group ME in that table.

All six overall domains were Much Employed. Classroom culture led at 4.49, immediately below the 4.50 threshold for Very Much Employed. Curriculum and implementation tied at 4.40. The medium-school averages (4.03–4.07) were lower than the big and small school averages in each domain. However, school-size groups differ in master-teacher composition (Table 1), and no uncertainty intervals or role-specific response distributions are available. The pattern may identify a context worth investigating, but it cannot establish an effect of school size or explain the difference. The lower curriculum and implementation ratings suggest examining whether observation feedback connects to lesson content and later action.

Table 3. Kendall concordance for practice-item rankings across school-size groups

Domain	Items	W	χ^2	df	.05 critical χ^2	Decision
Classroom culture	9	0.40	9.60	8	15.51	NS
Instructional delivery	10	0.39	10.53	9	16.92	NS
Classroom organization	10	0.34	9.18	9	16.92	NS
Curriculum	11	0.57	17.10	10	18.31	NS
Assessment	11	0.48	14.40	10	18.31	NS
Implementation Practices	10	0.21	5.67	9	16.92	NS

Note. *W*, χ^2 , *df*, critical values, and decisions reproduce Table 10. NS = not significant at .05. The tests use three group rankings, not 53 independent raters.

None of the six chi-square statistics exceeded its .05 critical value. Curriculum had the largest W (0.57) yet remained below its critical chi-square threshold; implementation had the smallest W (0.21). This is a failure to detect significant concordance of the item order across school-size groups, not a direct test that the extent of practices differs among respondents. The domain means in Table 2 are consistently favorable while the item-level order is not statistically aligned. A domain average can therefore conceal which particular aspects of curriculum guidance or implementation are salient to different groups.

3.2 Perceived effectiveness for teaching and learning

Table 4. Summary of perceived effectiveness of observation practices

Domain	Big Wx	Medium Wx	Small Wx	Overall Wx	Interpretation	Rank
Lesson Planning and Preparation	4.85	4.13	4.75	4.58	VME	2
Instructional Strategies and Delivery	4.86	4.14	4.66	4.55	VME	3
Classroom Management and Learning Environment	4.80	4.18	4.81	4.60	VME	1
Assessment and Feedback Practices	4.74	4.19	4.66	4.53	VME	4
Average	4.81	4.16	4.72	4.56	VME	

Note. Wx = weighted mean; VME = Very Much Effective. Values, labels, and ranks follow Table 15. The medium group's entries and average are printed ME (Much Effective).

Classroom management and learning environment was rated highest at 4.60, while assessment and feedback was lowest at 4.53, though both were in the same Very Much Effective band. The medium-school group again gave lower means (4.13–4.19) than the big and small groups. That pattern invites a closer look at context and respondent role; the survey does not demonstrate that observation changed planning or classroom behavior. A favorable assessment-feedback rating is also not a record of the specificity of advice or whether the teacher used it. Direct review of conferences and subsequent lessons would be needed to evaluate such mechanisms.

Table 5. Kendall concordance for perceived-effectiveness item rankings

Domain	Items	W	χ^2	df	.05 critical χ^2	Decision
Lesson Planning and Preparation	10	0.21	5.67	9	16.92	NS
Instructional Strategies and Delivery	10	0.13	3.51	9	16.92	NS
Classroom Management and Learning Environment	11	0.34	10.20	10	18.31	NS
Assessment and Feedback Practices	12	0.46	15.18	11	19.68	NS

Note. Statistics and decisions reproduce Table 16. NS = not significant at .05. The tabulated thresholds are critical chi-square values rather than exact p values.

The four perceived-effectiveness domains showed no statistically significant concordance in item rankings across the three school-size groups. Instructional strategies and delivery had $W = 0.13$, the lowest reported agreement, while assessment and feedback had $W = 0.46$. These values concern the relative priority of questionnaire items and do not test whether teachers' instructional strategies improved. The table does not warrant the stronger claim that groups have "completely different views"; nonsignificance can arise from limited ranking groups and compressed ratings. A role-stratified analysis and the raw item distributions would help determine where perspectives diverge.

3.3 Selected item patterns

Table 6. Selected classroom-culture observation items

Item	Big Wx	Medium Wx	Small Wx	Overall Wx	Interpretation
Visual displays promote a positive and motivating learning space	4.86	4.08	4.90	4.61	VME
All learners treated with fairness, equality, and respect	4.86	4.08	4.85	4.60	VME
Respectful communication between learners	4.86	3.92	4.85	4.54	VME
Classroom feels safe, inclusive, and welcoming	4.48	4.17	4.50	4.38	ME
Sensitivity to learners' diverse cultural backgrounds	4.67	3.92	4.50	4.36	ME

Note. Selected rows preserve the means and interpretation labels of Table 3; wording is condensed. VME = Very Much Employed; ME = Much Employed.

Visual displays and equitable treatment were the two highest selected culture practices, while sensitivity to diverse cultural backgrounds and a safe, inclusive environment had lower overall ratings. The contrast suggests that observers may find physical displays and overt interactions easier to rate than nuanced inclusion. It does not establish that the classrooms were unsafe or that master teachers ignored culture. The medium-school mean was particularly lower for respectful communication and cultural sensitivity (3.92 each), but role composition and contextual differences remain unexamined. Observation rubrics and examples of feedback could show what evidence respondents used to form these ratings.

Table 7. Selected assessment and feedback effectiveness items

Item	Big Wx	Medium Wx	Small Wx	Overall Wx	Interpretation
Develop rubrics defining performance expectations	4.76	4.42	4.70	4.63	VME
Encourage learner self-assessment and reflection	4.81	4.33	4.65	4.60	VME
Use formative assessments to guide instructional decisions	4.86	4.08	4.75	4.56	VME
Provide immediate constructive feedback on learners' work	4.67	4.25	4.55	4.49	ME
Return graded assignments and tests promptly	4.67	4.08	4.65	4.47	ME
Involve learners in setting their own learning targets	4.67	4.00	4.55	4.41	ME

Note. Selected means and printed interpretations follow Table 14. The table reports respondents' perceived effectiveness, not direct evidence that the practice changed after observation.

Defining criteria with rubrics received the highest selected effectiveness rating (4.63), while learner-set targets received 4.41. Immediate feedback and timely return of work were also below the Very Much Effective threshold. This item pattern matters because formal criteria and a completed score are only part of an assessment cycle. Learners need information they can use to revise an answer or set a next goal. Hattie and Timperley (2007) provide a framework for examining feedback content, but this survey did not record feedback messages or pupil revisions. The figures indicate where to inspect practice rather than an established effect of observation on assessment quality. Tables 6 and 7 also show how a favorable overall domain mean can mask specific priorities. Classroom culture and assessment feedback are broad labels; their items span visible displays, relational climate, rubrics, prompt feedback, and learner agency. A follow-up could select one observable item in each domain and compare independent observation notes with the teacher's account of the conference. That would make the differences between big, medium, and small schools more interpretable. The study's existing W tests, by contrast, assess rank concordance of all items within a domain and do not identify a causal explanation for any selected item.

4. Discussion

The practice ratings suggest a visible emphasis on classroom culture and delivery, with somewhat lower emphasis on curriculum and implementation. Observation can be a means of collecting instructional evidence, but its value depends on an ensuing conversation and an opportunity to apply feedback. O'Leary (2020) describes the different functions of observation in educator development; the study did not document whether conferences were developmental, evaluative, or both. Thus the local rank order suggests what to inspect in the process, not a validated judgment about mastery of observational practice.

The four favorable effectiveness ratings require a separate interpretation. They reflect master teachers' and teachers' perceptions of lesson planning, strategies, management, and assessment. The high classroom-management mean may reflect practices that are immediately observable during a lesson, whereas assessment adjustments can unfold over several lessons and may be harder to attribute to one observation. Hattie and Timperley (2007) emphasize that feedback varies in type and usefulness. The study's broad assessment-feedback rating cannot show whether comments identified a specific gap, proposed a usable next step, or prompted an improved response to pupils' work. School-size differences recur in both descriptive tables. The medium group's means are markedly lower, but only four schools and twelve respondents are in that group. Master teachers are concentrated in big schools, and some small schools have none. Because the survey did not present ratings by respondent role, a big-versus-medium contrast blends school context, staffing, and observer/teacher viewpoint. The nonsignificant Kendall tests concern concordance of item rankings, not tests of differences in group-level means. The evidence does not justify attributing the descriptive gap to resources, class sizes, leadership, or training without separate records.

For the policy question, the study recommended pre-observation meetings, attention to pacing, dedicated planning time, peer observation, and coaching. Those are proposals derived from ratings, not interventions that were implemented and evaluated. The PPST can help specify a focus for the meeting and feedback (DepEd, 2017). The LAC policy offers a school-based space for colleagues to examine a common instructional concern (DepEd, 2016). A practical cycle would define a target skill, gather observation evidence, provide feedback, document an agreed trial, and revisit pupil work. The present survey supplies reasons to examine this process but cannot certify its effect.

Several limitations affect the strength of the claims. All outcomes were reported on the same questionnaire, creating scope for shared response patterns. No independent observation ratings, lesson artifacts before and after feedback, or pupil outcomes were analyzed. The school-size classification rule and role-specific response distributions were not supplied. The KR-21 reliability claim is not transparently suited to five-category items, and the pilot details are incomplete. The correlational label describes the design more broadly than the reported concordance tests; no association between frequency of observation and actual learning was estimated. The results are local to Ragay I and should not be generalized automatically to other districts.

A follow-up could record an observation's focus, evidence, feedback, agreed action, and next observation for the same teacher. Independent scoring of lesson plans, classroom practice, and samples of pupil work would permit evaluation of change. Teacher interviews could explain whether observation is experienced as useful and safe for reflection. If group comparisons remain a question, school-size rules, the role mix, and school-level clustering should be specified in advance. These steps would distinguish the perceived effectiveness documented here from a demonstrated improvement in teaching and learning.

5. Conclusion

The study reports classroom observation practices as Much Employed overall (4.44) and their perceived effectiveness as Very Much Effective (4.56). None of the six practice domains or four effectiveness domains showed significant concordance of item rankings across the three school-size groups at .05. The results identify favorable respondent appraisals and point to curriculum implementation and assessment feedback for closer review. They do not establish that master-teacher observation improved teaching or pupil learning; direct, longitudinal evidence would be needed for that claim.

References

- [1] Department of Education. (2016). DO 35, s. 2016: The Learning Action Cell as a K to 12 Basic Education Program school-based continuing professional development strategy for the improvement of teaching and learning. <https://www.deped.gov.ph/2016/06/07/do-35-s-2016-the-learning-action-cell-as-a-k-to-12-basic-education-program-school-based-continuing-professional-development-strategy-for-the-improvement-of-teaching-and-learning/>
- [2] Department of Education. (2017). DO 42, s. 2017: National adoption and implementation of the Philippine Professional Standards for Teachers. <https://www.deped.gov.ph/2017/08/11/do-42-s-2017-national-adoption-and-implementation-of-the-philippine-professional-standards-for-teachers-2/>
- [4] Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. O'Leary, M. (2020). *Classroom observation: A guide to the effective observation of teaching and learning* (2nd ed.). Routledge.