Peer Checking and the Practical Cookery Skills of Grade 10 Learners

An Action Research

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

Conducted July to September 2025

Abstract

This action research tested whether peer checking, in which a partner checks the cook against the performance criteria during practice and says right after the turn what was missed, would raise the share of Grade 10 learners judged competent in practical cookery tasks. Two Grade 10 Cookery sections of one TLE teacher in a public secondary school took part in School Year 2025–2026, in a two-period crossover: a coin toss gave one section peer checking while practicing a vegetable omelet (Unit 1) and the usual practice while practicing stir-fried mixed vegetables (Unit 2), and the other section the reverse. After each unit, two assessors, not told which section had practiced which way, judged every learner on ten performance criteria, Competent or Not Yet Competent, and scored the plated dish on a rubric. Of the 76 learners analyzed, 53.9% were competent on all ten criteria after peer checking and 30.3% after the usual practice, a paired difference of 23.7 percentage points, 95% CI [10.0, 36.1]. The Mainland–Gart test, which takes the unit into account, was significant, p = .004, odds ratio = 4.00, 95% CI [1.55, 10.33]; the difference held in Unit 1 alone. The gains lay in mise en place, cleaning as one works, heat control, presentation, and the safe handling of tools; the rubric totals did not differ significantly, 0.60 points of 20, 95% CI [−0.30, 1.50]. Kappa ranged from .26 to .73 by criterion. Peer checking with the performance checklist is recommended for practical lessons in Cookery.

Keywords: competency-based assessment, peer assessment, Cookery, crossover design, Cohen's kappa

Acknowledgment

The researchers thank the school head, who approved the study, kept both sections' TLE periods in the kitchen laboratory for two units, and gave the assessors time from their own classes; the two TLE teachers who judged all 152 performances; the Cookery teacher of a neighboring school who judged, from the recordings, the criteria on which they disagreed; the TLE department head, who witnessed the coin toss and observed the practice periods; the school nurse, who checked the health records for food allergies; the Grade 11 Cookery learners and their parents, who allowed the assessors to train on their practice; the Grade 10 advisers, who gave out and collected the consent forms; and the 81 Grade 10 learners and their parents. The ingredients were the school's regular TLE consumables; the study's other costs were paid by the researchers.

Peer Checking and the Practical Cookery Skills of Grade 10 Learners

Context and Rationale

A Cookery learner is judged on what they do in a kitchen, not on what they can say about it. The Cookery NC II training regulations of the Technical Education and Skills Development Authority (TESDA) set the standard as performance criteria, such as "Ingredients are assembled according to correct quantity, type and quality required" and "Eggs are presented hygienically and attractively using suitable garnishes and side dishes sequentially within the required timeframe" (TESDA, 2014, p. 58), and they assess "through: 5.1 Direct observation of the candidate while preparing dishes" (p. 60). DepEd's Cookery curriculum guide asks learners to "apply core competencies in cookery as prescribed in the TESDA Training Regulation" (Department of Education [DepEd], 2016a, p. 22), and in its first lesson, egg dishes, to "present egg dishes hygienically and attractively using suitable garnishing and side dishes sequentially within the required time frame" (p. 25).

Our own practical assessments showed where our learners fell short. In the previous school year, learners who could cook a dish lost the competent rating on routines more often than on cooking itself: ingredients not ready when the pan was hot, a station left wet and cluttered, a plate finished past the time. In the usual practice period, the teacher moves between five stations and sees each learner for a minute or two of a turn.

Peer assessment puts a second pair of eyes beside every turn. It is "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar

status" (Topping, 1998, p. 250). In a meta-analysis of 54 control-group studies, peer assessment improved academic performance compared with no assessment (g = 0.31) and with assessment by the teacher (g = 0.28), and its effect for secondary students was g = 0.44 (Double et al., 2020, pp. 481, 493); the meta-analysis counted practical skills as performance, though none of its studies was in cooking. A closer procedural skill is basic life support: when university students practiced in pairs with task cards, and the partner checked the doer's actions against the cards before each switch of roles, significantly more of them performed every step in the correct order than students who practiced in pairs without the check (Iserbyt et al., 2009, p. 1396).

Whether peer checking against the performance criteria helps Grade 10 learners in a public school kitchen laboratory, with the same number of cooking turns as the usual practice, is therefore a question for our own classes. The study addresses the Teaching and Learning theme of the Basic Education Research Agenda (DepEd, 2016b).

Innovation, Intervention, and Strategy

The innovation was peer checking during the practice periods of a Cookery unit. Both Grade 10 sections studied two units of the Cookery curriculum guide in their usual order: egg dishes in Quarter 1 and vegetable dishes in Quarter 2 (DepEd, 2016a, pp. 24–28), the 2016 curriculum that Grade 10 still followed in School Year 2025–2026 (DepEd, 2024a, paras. 8 and 35). The school's class program gave TLE four periods of 60 minutes a week, an option DepEd allows for Grades 3 to 10 (DepEd, 2024b, para. 2(a)

(ii)). In each unit, both sections had the same lessons, the same demonstration by the teacher, the same recipe card, and four practice periods of 60 minutes in the kitchen laboratory, from July 28 to August 8, 2025 for Unit 1 (vegetable omelet) and from September 1 to 12, 2025 for Unit 2 (stir-fried mixed vegetables).

The Kitchen Laboratory and the Practice Period

The kitchen laboratory has five cooking stations, each with a two-burner stove, a sink, and a counter. In a practice period, eight or nine learners worked at each station in pairs that stayed the same for the unit, with a group of three where a section's number was odd. Each learner had one cooking turn per period. Two pairs at a time took the station's two burners; in each pair, one partner cooked the dish from mise en place to plating while the other did not cook, and then they changed places. The ten performance criteria of the unit's checklist (Appendix C) were explained in the first lesson of the unit and posted at every station in both practices, as competency-based training requires learners to know the standard they will be judged against.

Peer Checking

In peer checking, the partner who was not cooking held the checklist and marked each criterion Competent or Not Yet Competent as the cook worked, then, right after the plate was set down, told the cook, criterion by criterion, which ones were marked Not Yet Competent and what they had seen. The feedback took about two minutes. The partners then changed places with a fresh checklist. In the unit's first lesson, the class practiced marking: the teacher cooked the dish once with three planted errors, every

learner marked the checklist, and the class compared marks against the criteria for 15 minutes. The checklists were collected after each period, kept by the teacher, and never graded.

The Usual Practice

In the usual practice, the partner who was not cooking washed the pair's tools and readied the station for the next turn, and could watch the cook. In the first lesson, the 15 minutes the peer-checking section spent on marking went to questions on the demonstration and a second look at the plating. In both practices the teacher moved between the stations and gave feedback as usual; the only difference between the two was the partner's checklist and the feedback after the turn.

The Crossover

A coin toss on July 11, 2025, witnessed by the TLE department head, gave Section A peer checking in Unit 1 and the usual practice in Unit 2 (sequence PU), and Section B the usual practice in Unit 1 and peer checking in Unit 2 (sequence UP). Each section thus practiced once in each way, and each learner was assessed once after each.

Strategy and Evaluation Plan

The study compared each learner's practical assessment after peer checking with their assessment after the usual practice, taking the unit into account. The TLE department head observed periods 2 and 4 of each section in each unit and, at one moment of each, counted the ten pairs at the burners and how many were checking with the checklist. Three criteria for success were set in the proposal: (a) all practice periods held as planned and the

partner's checklist completed on at least 85% of the cooking turns with peer checking; (b) the two assessors agreeing, before any resolution, on at least 85% of the judgments on every criterion; and (c) a higher share of learners competent on the task after peer checking, significant at the .05 level with the unit taken into account, and in the same direction in Unit 1 alone and with each assessor's judgments alone.

Action Research Questions

Main question: Does peer checking against the performance criteria raise the share of Grade 10 learners judged competent in practical cookery tasks?

1. What share of learners were judged competent on each performance criterion, and on the whole task, after peer checking and after the usual practice?
2. How closely did the two assessors agree on each criterion and on the overall decision before their disagreements were resolved, and on the rubric totals?
3. Is the share of learners competent on the task significantly higher after peer checking than after the usual practice, with the unit taken into account?
4. Did the rubric totals of the plated dishes differ between the two practices?

Two null hypotheses were tested at the .05 level of significance:

$H_0 1$: The share of learners competent on the task does not differ between peer checking and the usual practice.

$H_0$2: The rubric totals of the plated dishes do not differ between peer checking and the usual practice.

Action Research Methods

Research Design

The study used a two-period, two-sequence crossover with two intact sections (Hills & Armitage, 1979). A class is taught as a section, and learners at five shared stations see what every pair does, so the two practices could not be mixed within a section. In a crossover, each section, and each learner, practiced once each way, so differences between the sections cancel, and the coin toss decided only the order. The cost is that the two practices came with two different dishes in two different months; the analysis takes the unit into account, and the order was balanced across the two sections.

Participants and Other Sources of Data and Information

The participants were the learners of the researcher-teacher's two Grade 10 Cookery sections in School Year 2025–2026. All 81 enrolled learners were invited; no sampling was done. Table 1 shows how many were analyzed.

Table 1

Learners Enrolled, Excluded, and Analyzed, by Section

Learners	Section A (PU)	Section B (UP)	Total
Enrolled	41	40	81
No parental consent	1	1	2
Egg allergy (alternative task in Unit 1)	0	1	1
Transferred to another school after Unit 1	1	0	1
Absent from the Unit 2 assessment and its make-up	0	1	1
Analyzed	39	37	76

Note. Learners who were not analyzed took part in the lessons and practicals as usual; only their records were left out. A learner was analyzed only with an assessment in both units.

The learners analyzed were 21 girls and 18 boys in Section A and 20 girls and 17 boys in Section B.

Two TLE teachers of the school who taught neither section were the assessors: a senior high school Cookery teacher who holds Cookery NC II and is an accredited competency assessor, and a Grade 9 TLE teacher who holds Cookery NC II. Where the two disagreed, a Cookery teacher of a neighboring school who holds Cookery NC II, and who knew neither the sections nor the order in which they practiced, made the final judgment. The TLE department head witnessed the coin toss and observed the practice periods. Ten Grade 11 Cookery learners, with their parents' consent, cooked the vegetable omelet for the assessors' calibration rounds.

Sample Size

Because whole sections took part, the number of learners was fixed. In a paired comparison of shares, only the learners competent after one practice and not after the other carry information, so the power depends on how many there are. Assuming that .35 of the learners would be competent after one practice only, a share in line with the practical assessments of the previous school year, Connor's (1987) formula for McNemar's test gave 77 learners, the number expected after consent and absences, a power of .80 to detect a difference of 18.6 percentage points in the share competent at α = .05, two-sided. The Mainland–Gart test, which splits these learners by sequence, has somewhat less power. A difference of 15 points would have needed 120. In the study, .39 of the learners analyzed were competent after one practice only.

Data Gathering Methods

The Practical Assessment

After each unit, every learner cooked the unit's dish alone, from the recipe card, in 25 minutes from the first step of mise en place to the plated dish. Unit 1's assessment ran from August 11 to 14, 2025, and Unit 2's from September 15 to 18, 2025, in each section's TLE periods, with a make-up day for absences. Learners came to the kitchen in batches of five, drawn by lot within the section, one learner to a station, with stations drawn by lot within the batch. The ingredients were set out in bulk on a common table, so that each learner measured their own; the tools at every station were the same; the time started for the whole batch at once. Learners wore aprons with numbers in place of name tags.

The Performance Checklist

Each learner was judged on ten performance criteria, each Competent or Not Yet Competent (Appendix C). Six are the same for both dishes: kitchen attire, mise en place, safe handling of tools, presentation, cleaning as one works, and finishing within time; presentation and time follow the curriculum guide's competency on presenting dishes "within the required time frame" (DepEd, 2016a, p. 25). Four are written for each dish: hand washing, ingredient preparation, heat control, and doneness. The criteria come from the performance criteria and evidence guides of the Cookery NC II units the lessons teach (TESDA, 2014): "Prepare egg dishes" and "Prepare vegetables dishes" for the dish, "Clean and maintain kitchen premises" for the station, and the units on hygiene and on occupational health and safety for attire,

hand washing, and the safe handling of tools, which the evidence guide for egg dishes names: "Followed safety practices in handling tools and equipment" (p. 60). Doneness of the omelet was judged as no liquid egg left, following the World Health Organization's advice to "Cook eggs and seafood until piping hot throughout" (World Health Organization [WHO], 2006, p. 17).

A learner was judged competent on the task only when the final judgment on all ten criteria was Competent. The rule follows competency-based assessment, in which a candidate must meet every performance criterion of a unit, not an average of them.

The Product Rubric

After plating, both assessors scored the dish on four attributes, appearance, texture, flavor, and portion and plating, each from 1 to 5 with a descriptor for every level (Appendix C), for a total of 4 to 20. Each assessor tasted with a clean spoon. The curriculum guide's egg lesson asks learners to "rate the finished products using rubrics" (DepEd, 2016a, p. 25), and the class used the same rubric to rate their own dishes during the units. The final total was the mean of the two assessors' totals.

Two Assessors and a Third

The two assessors watched all five stations of each batch from opposite sides of the kitchen, marked their checklists independently, and did not confer. Where their judgments on a criterion differed, the second researcher cut the station's recording of the moment the criterion concerned into a short clip, gave it a random code, and shuffled it with the clips from both sections and both units. A Cookery teacher of a neighboring school who holds Cookery

NC II judged each clip and gave the final judgment, without the two assessors' marks and without knowing the section, the order in which it practiced, or which practice came before the unit. One phone camera on a stand at each station recorded the assessment, framed on the counter, the stove, and the sink, below the learners' shoulders.

Assessor Training

The assessors were trained on July 22, 2025 on the checklist and the rubric, criterion by criterion. They then watched ten Grade 11 Cookery learners cook the vegetable omelet in a TVL class and judged the ten criteria alone: they agreed on 81.0% of the 100 judgments and on the overall decision for 8 of the 10 performances. After discussing every disagreement against the wording of the criteria, they judged ten more performances by the same class on July 25, 2025, agreeing on 91.0% and on the overall decision for 9 of 10. The rule set in advance, at least 90% agreement and the same overall decision for at least 9 of the 10 performances, was met in the second round, and the study's assessments began.

Keeping the Practice From the Assessors

The assessors were told that the study compared two ways of practicing but not which section practiced which way in which unit. The checklists and the practice records stayed in the teacher's cabinet, and the learners were asked not to talk about how they had practiced. Within a unit, every learner in a batch had practiced the same way, so an assessor who guessed one learner's practice could guess the batch's. After each unit's assessment, each

assessor wrote down which section they believed had practiced with peer checking.

Implementation Records

The teacher kept a log of each learner's attendance and cooking turns in every practice period and counted the completed checklists. The TLE department head observed periods 2 and 4 of each section in each unit and, at one moment of each, counted the ten pairs at the burners and how many were checking with the checklist. A safety log recorded every injury in the kitchen.

Storage

Paper checklists, scoring sheets, and consent forms were kept in a locked cabinet in the TLE office. The data set carries only learner codes; the key linking codes to names stayed with the second researcher. The recordings were kept on a password-protected school laptop; only the coded clips of the criteria in dispute were shown to the assessor from the neighboring school, at the school, and the recordings will be deleted when the school's research committee accepts this report.

Safety and Food Handling

The kitchen rules were taught in the first lesson of each unit and posted at the stations. Before Unit 1, the school nurse checked the learners' health records for food allergies. One learner was allergic to eggs; on the parent's and the school nurse's advice, the learner did an alternative task in Unit 1, the knife work and plating of a vegetable garnish, and was not analyzed. Eggs came from one supplier, were kept refrigerated in their trays, and were used

first in, first out. The first-aid kit and a burn station at the main sink were ready in every period. Every dish was eaten by the class at the end of the period; none was sold.

Ethical Considerations

The school head approved the study in writing on July 4, 2025. Parents received a letter in English and Filipino that explained the purpose, the two ways of practicing, the assessments, the recordings and that a Cookery teacher from another school would view short clips of them, the risks and benefits, and the right to withdraw, and signed a consent form; learners gave their assent in the TLE period on July 14, 2025, after the same explanation. Every learner had the same lessons, practice, and assessments whether or not they took part; only the records of consenting learners were kept for the study. A learner judged Not Yet Competent practiced again in the week after the assessment, with peer checking in both sections, and was reassessed, and the grade for the unit's performance task used the reassessment, so that no learner's grade depended on the section's place in the crossover. The study's data are the first assessments only.

A recording that shows a learner's hands and apron number is personal information when it can be traced to the learner (Republic Act No. 10173, 2012, Sec. 3(g)). The recordings were therefore framed below the shoulders, kept on a protected laptop, limited to the assessment days, and shown, as coded clips of the criteria in dispute, only to the assessor from the neighboring school. The ingredients for both units' practice and assessments were the TLE consumables in the school's Annual Implementation Plan, paid from school

MOOE (DepEd, 2019, Sec. V.D.3); no learner paid for or brought ingredients (DepEd, 2024c, items 2 and 5). The study followed the Research Management Guidelines (DepEd, 2017). The researchers' use of an artificial intelligence (AI) tool is declared in Appendix F (DepEd, 2026).

Data Analysis

For the first question, the share of learners judged competent was computed for each criterion and for the whole task after each practice, with Wilson 95% confidence intervals, and the paired difference with Newcombe's (1998b) interval for paired data, his method 10.

For the second, agreement was computed on the two assessors' independent judgments, before the third assessor resolved any difference, because resolved judgments agree partly by construction. For each criterion and for the overall decision, we report the percentage of agreement, Cohen's (1960) kappa with a 95% confidence interval from 2,000 bootstrap resamples of the learners, and the prevalence index, the bias index, and the prevalence-adjusted bias-adjusted kappa (PABAK) of Byrt et al. (1993), since "it can be misleading to report kappa values alone" (p. 429). Kappa was labeled by Landis and Koch's (1977) benchmarks. For the rubric totals, we report the intraclass correlation coefficient from a two-way random-effects model for absolute agreement, ICC(A,1) for a single assessor and ICC(A,2) for the mean of the two (McGraw & Wong, 1996), read against Koo and Li's (2016) guidelines.

For the third, the primary analysis was the Mainland–Gart test (Gart, 1969; Hills & Armitage, 1979). Learners competent in both units or in neither

carry no information about the practice; for the others, a 2 × 2 table counts, in each sequence, those competent in Unit 1 only and in Unit 2 only, and Fisher's exact test compares the two sequences. The odds ratio for peer checking, with its 95% confidence interval, came from conditional logistic regression with learners as strata and the unit as a covariate. Three checks followed: the same test on each assessor's judgments alone; McNemar's exact test, which ignores the unit; and Unit 1 alone, a comparison of the two sections before either could carry anything over from an earlier unit, by the difference in shares with Newcombe's (1998a) hybrid score interval and Fisher's exact test. A crossover can be biased if a practice carries over into the next unit, and a test for it "might be a rather insensitive test" (Hills & Armitage, 1979, p. 12), so Unit 1 alone was reported beside the main analysis, as planned, and not chosen by a preliminary test.

For the fourth, each learner's Unit 1 total minus Unit 2 total was compared between the two sequences with a two-sample *t* test; half the difference between the sequences' means estimates the effect of peer checking, and half their sum the difference between the units (Hills & Armitage, 1979). The normality of the differences was checked with the Shapiro–Wilk test (Shapiro & Wilk, 1965) on each learner's difference from their sequence's mean.

The analyses were run in Python 3.12 with pandas, SciPy, statsmodels (Seabold & Perktold, 2010), and pingouin (Vallat, 2018); the Excel data set computes the final judgments, the agreement statistics, the shares and their intervals, the Mainland–Gart table with Fisher's exact test, the odds ratio, the

ICC, and the crossover *t* test as formulas, and the notebook reproduces every result (Appendix A).

Discussion of Results and Reflection

Implementation

All practice periods were held. The learners analyzed had a mean of 3.91 cooking turns of 4 with peer checking and 3.83 with the usual practice; 69 and 63 of the 76 had all four. The partner's checklist was completed on 275 of the 297 cooking turns with peer checking (92.6%). At the moments observed, 37 of the 40 pairs at the burners were checking with the checklist in peer checking and 0 of 40 in the usual practice. The first criterion for success was met. Minor injuries were logged three times, two burns and one cut, all treated with first aid; none needed referral.

Question 1: The Share Competent

After peer checking, 53.9% of the learners were competent on all ten criteria; after the usual practice, 30.3% (Table 2 and Figure 1). The paired difference was 23.7 percentage points, 95% CI [10.0, 36.1]. The largest differences were in mise en place, 18.4 points, and cleaning as one works, 17.1 points, followed by heat control, presentation, and the safe handling of tools. On ingredient preparation and doneness the two practices did not differ. The ten criteria were described with their intervals, not tested one by one, so no correction for multiple tests was made; an interval that excludes zero on one criterion describes where the difference lay.

Table 2

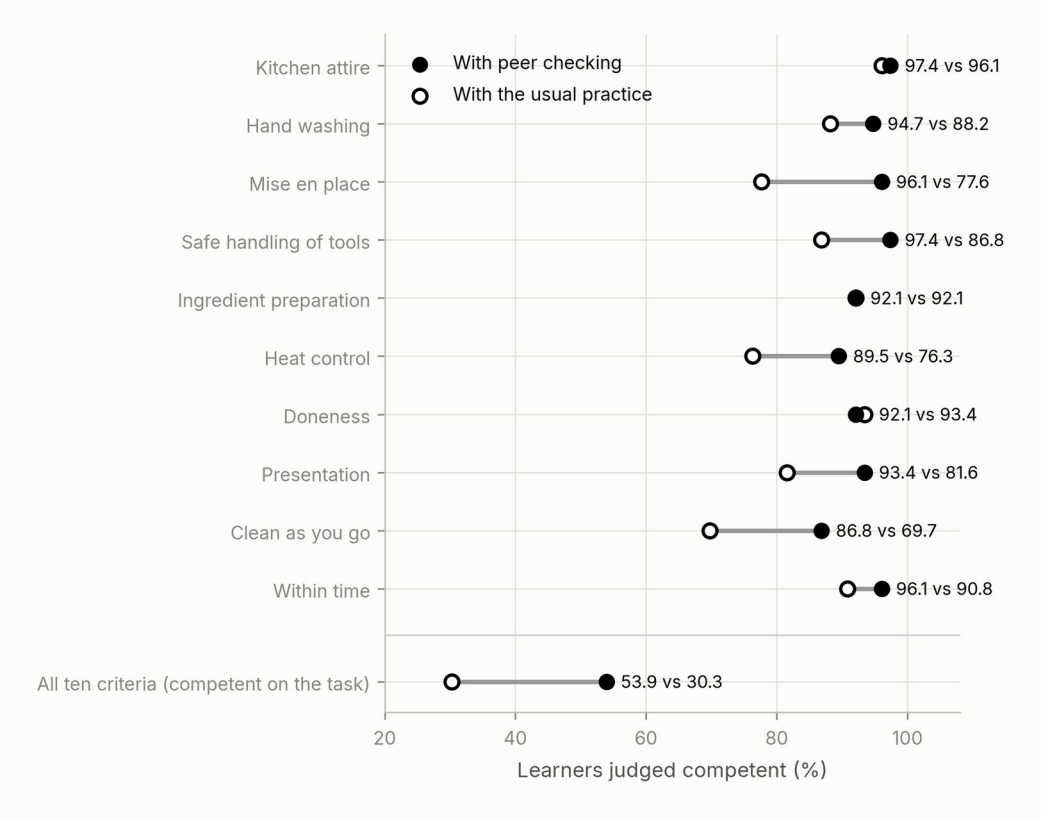
Learners Judged Competent After Each Practice, by Criterion (%)

Criterion	Peer checking	Usual practice	Difference (points) [95% CI]
Kitchen attire	97.4	96.1	1.3 [−5.7, 8.7]
Hand washing	94.7	88.2	6.6 [−3.1, 16.5]
Mise en place	96.1	77.6	18.4 [7.3, 29.6]
Safe handling of tools	97.4	86.8	10.5 [1.5, 20.2]
Ingredient preparation	92.1	92.1	0.0 [−9.6, 9.6]
Heat control	89.5	76.3	13.2 [0.6, 25.4]
Doneness	92.1	93.4	−1.3 [−10.3, 7.6]
Presentation	93.4	81.6	11.8 [0.5, 23.1]
Clean as you go	86.8	69.7	17.1 [4.2, 29.4]
Within time	96.1	90.8	5.3 [−3.4, 14.4]
All ten criteria (competent on the task)	**53.9**	**30.3**	**23.7 [10.0, 36.1]**

Note. 76 learners, each assessed once after each practice. Final judgments: the two assessors' common judgment, or the third assessor's where they differed. Differences are peer checking minus usual practice, in percentage points, with Newcombe's 95% confidence interval for paired data.

Figure 1

Learners Judged Competent After Peer Checking and After the Usual Practice, by Criterion



Note. Filled points: after peer checking. Open points: after the usual practice. The last row is the share competent on all ten criteria.

Small differences on single criteria added up on the whole task, which needs all ten: after the usual practice, 33 of the 76 learners missed two or more criteria, against 9 after peer checking.

Question 2: The Assessors' Agreement

The two assessors differed on 130 of the 1,520 criterion judgments (8.6%), in 85 of the 152 performances; the third assessor sided with Assessor 1 on 65 and with Assessor 2 on 65. Agreement by criterion ranged from 87.5% on presentation to 97.4%, and every criterion met the second criterion for

success of 85% (Table 3). Kappa ranged from .26 on hand washing to .73 on within time, from "Fair" to "Substantial" in Landis and Koch's (1977) terms.

Table 3

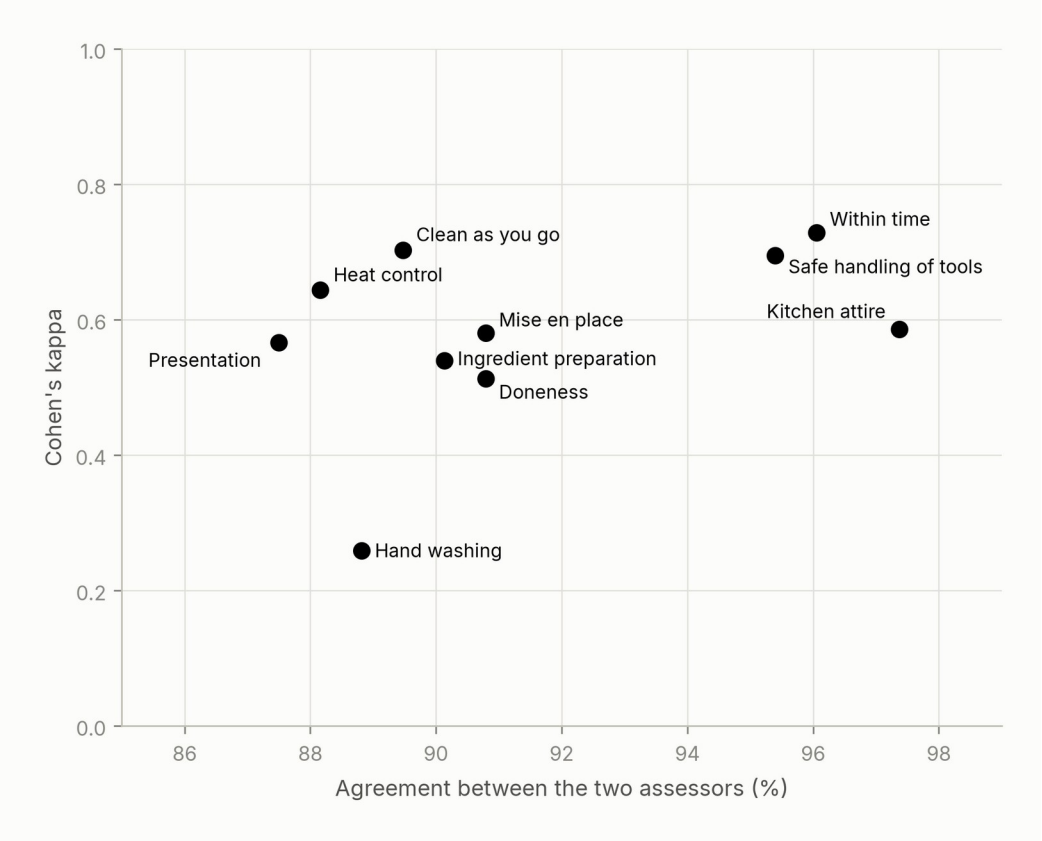
Agreement Between the Two Assessors Before Resolution, by Criterion

Criterion	Agreement (%)	Cohen's κ [95% CI]	Prevalence index	Bias index	PABAK
Kitchen attire	97.4	.59 [−.01, .91]	.93	.00	.95
Hand washing	88.8	.26 [−.01, .50]	.84	.01	.78
Mise en place	90.8	.58 [.34, .78]	.75	.04	.82
Safe handling of tools	95.4	.69 [.44, .88]	.84	.01	.91
Ingredient preparation	90.1	.54 [.31, .74]	.76	−.03	.80
Heat control	88.2	.64 [.49, .80]	.58	.01	.76
Doneness	90.8	.51 [.27, .71]	.79	.04	.82
Presentation	87.5	.57 [.38, .73]	.65	.03	.75
Clean as you go	89.5	.70 [.53, .84]	.54	−.01	.79
Within time	96.1	.73 [.47, .91]	.84	−.03	.92
Overall decision (all ten)	77.6	.50 [.35, .65]	−.33	.03	.55

Note. 152 performances, each judged by both assessors. κ = Cohen's kappa; its 95% confidence interval is from 2,000 bootstrap resamples of the learners. Prevalence index = (both Competent − both Not Yet Competent) ÷ *n*; bias index = (Assessor 1 only Competent − Assessor 2 only Competent) ÷ *n*; PABAK = 2 × agreement − 1 (Byrt et al., 1993).

Figure 2

Agreement and Kappa for the Ten Criteria



Note. Each point is a criterion, before resolution. Nearly every learner was competent on four criteria, kitchen attire, hand washing, safe handling of tools and within time (prevalence index .80 or more). Hand washing, one of them, has the lowest kappa; the other three have higher agreement and kappas of .59 to .73.

Kappa and agreement told different stories where nearly every learner was competent. On hand washing, the assessors agreed on 88.8% of the performances, yet kappa was .26; on heat control they agreed on 88.2%, and kappa was .64. The difference is the prevalence: both assessors judged 131 of the 152 performances competent on hand washing, so two assessors guessing would have agreed often, and kappa, which counts only agreement beyond chance, had little room left (prevalence index .84, against .58 for heat control). The PABAK values, .78 and .76, show the two criteria judged about equally

well. Neither assessor judged Competent much more often than the other: the bias index was between −.03 and .04 on every criterion.

On the overall decision, competent on all ten or not, the assessors agreed on 77.6% of the performances, κ = .50, 95% CI [.35, .65]. The overall decision agrees less than any single criterion because one disagreement on any of ten criteria splits it, which is why every disagreement went to the third assessor before the decision was made.

The rubric totals agreed closely: ICC(A,1) = .87, 95% CI [.72, .93], for a single assessor, and ICC(A,2) = .93, 95% CI [.83, .96], for the mean of the two: moderate to excellent for a single assessor and good to excellent for the mean, read on the confidence intervals against Koo and Li's (2016) guidelines. Assessor 1's totals averaged 0.89 points higher than Assessor 2's, $t(151) = 7.39$, $p < .001$. Because both assessors scored every performance of both practices, this leniency entered both alike.

The practice was kept from the assessors as far as the design allowed. After Unit 1, Assessor 1 could not tell which section had practiced with peer checking, and Assessor 2 named the wrong one; after Unit 2, neither could tell.

Question 3: The Difference Between the Practices

Of the 76 learners, 30 were competent after one practice only: 24 after peer checking only and 6 after the usual practice only. Table 4 splits them by sequence. The Mainland–Gart test, which takes the unit into account, was significant, $p = .004$, and the odds of being competent on the task were 4.00 times as high after peer checking, 95% CI [1.55, 10.33]. $H_0 1$ is rejected. With

the practice held constant, the odds of being competent were lower in Unit 2 than in Unit 1, odds ratio 0.50, $p = .152$, a difference that was not significant.

Table 4

Learners Competent on the Task in Each Unit, by Sequence

Sequence	Competent in both units	Competent in Unit 1 only	Competent in Unit 2 only	Competent in neither
PU (peer checking in Unit 1), $n = 39$	8	**16**	**2**	13
UP (peer checking in Unit 2), $n = 37$	9	**4**	**8**	16

Note. The bold counts form the Mainland–Gart table. In sequence PU, competent in Unit 1 only means competent only after peer checking; in sequence UP, competent in Unit 2 only does.

The checks agreed with the main result (Table 5). With Assessor 1's judgments alone the difference was significant, and with Assessor 2's alone it was in the same direction and not significant, $p = .087$. In Unit 1, before either section could carry a practice over, 24 of 39 learners were competent after peer checking and 13 of 37 after the usual practice, a difference of 26.4 points, 95% CI [4.0, 45.4], Fisher's $p = .024$. In Unit 2 the difference was 20.3 points, 95% CI [−1.2, 39.5]. The third criterion for success was met: the difference was significant with the unit taken into account and pointed the same way in Unit 1 alone and with each assessor's judgments.

Table 5

The Difference Between the Practices Under Four Checks

Analysis	Result
Main analysis: final judgments, Mainland–Gart test	OR = 4.00, 95% CI [1.55, 10.33], $p = .004$
Assessor 1's judgments alone	OR = 3.54, $p = .004$
Assessor 2's judgments alone	OR = 2.55, $p = .087$
McNemar's test, ignoring the unit	24 against 6 learners, $p = .001$

Analysis	Result
Unit 1 alone (two sections, before any carryover)	24 of 39 against 13 of 37; difference 26.4 points, 95% CI [4.0, 45.4], Fisher's $p = .024$

Note. OR = odds ratio for being competent on the task after peer checking, from conditional logistic regression with learners as strata and the unit as a covariate. The *p* values of the Mainland–Gart test are from Fisher's exact test.

Question 4: The Plated Dishes

The rubric totals were close after both practices (Table 6). Taking the unit into account, the difference was 0.60 points of 20 in favor of peer checking, 95% CI [−0.30, 1.50], $t(74) = 1.32$, $p = .189$. $H_0 2$ is not rejected. Unit 1's totals minus Unit 2's, with the practice held constant, came to −0.44 points, 95% CI [−1.34, 0.45]. The learners' differences between the units, centered on their sequence's mean, looked normal, Shapiro–Wilk $W = .98$, $p = .421$, but varied more in sequence UP, $SD = 4.62$, than in sequence PU, $SD = 3.14$; Welch's test, which does not assume equal variances, gave $p = .194$, and each assessor's totals alone gave 0.45 and 0.75 points.

Table 6

Rubric Scores of the Plated Dishes After Each Practice

Score	Peer checking *M* (*SD*)	Usual practice *M* (*SD*)
Total (4–20), Unit 1	13.72 (3.08)	12.80 (3.70)
Total (4–20), Unit 2	13.84 (3.61)	13.56 (2.55)
Total (4–20), both units	13.78 (3.33)	13.19 (3.16)
Appearance (1–5)	3.47	3.43
Texture (1–5)	3.42	3.14
Flavor (1–5)	3.33	3.18
Portion and plating (1–5)	3.55	3.43

Note. Each score is the mean of the two assessors'. 76 learners, each assessed once after each practice.

Reflection

What We Found

Peer checking raised the share of learners competent on the whole task, from 30.3% to 53.9%, with the same number of cooking turns. The gain lay in the routines a partner with a checklist can see and the cook can miss: having everything measured before the heat, cleaning while working, keeping the heat right, plating to the card. On ingredient preparation and doneness, the two practices ended level. The plated dishes scored about the same; a nonsignificant difference of 0.60 points does not show that the dishes were equal, and its interval runs from a small loss to a gain of 1.50 points.

How It Compares With the Research

Double et al. (2020) found peer assessment better than no assessment by $g = 0.31$ across subjects, and an effect of 0.44 for secondary students; our outcome is a share, not a mean, and the difference of 23.7 points in the share competent on all ten criteria is larger than the difference on any single criterion, 18.4 points at most. Iserbyt et al. (2009) found that peer evaluation with task cards raised the number of learners who performed all the steps of a procedure in the correct order, without a significant difference in the total score; our result has the same shape: the share meeting every criterion changed more than the mean number of criteria met, 9.4 against 8.5.

What Else Could Explain the Difference

The two practices came with two dishes in two months, and the order was balanced only across two sections; a difference between the sections in how they took to one dish could look like an effect of the practice. Unit 1

alone, which compares the two sections on the same dish, gave a similar difference, but it rests on two sections. The assessors watched each section in its own periods, so the practice could not be fully hidden; neither could tell which section had practiced with peer checking, but a guess check of four guesses shows little. The disputed judgments were settled by a teacher from another school who judged coded clips without knowing the section or its order. The same teacher taught both practices and knew which was which. Peer checking may also work partly through the partner's attention rather than the checklist; the study cannot separate the two.

A practice can carry over. Section A, which checked in Unit 1, may have kept checking itself in Unit 2; if so, the usual practice in Unit 2 looked better than it was, and the difference is smaller than peer checking's full effect, not larger.

Limits

The assessments came within a week of the last practice period, so the study shows what the learners could do at the end of a unit, not whether it lasted. Agreement on hand washing was high but its kappa was low, because nearly every learner met it, and kitchen attire's kappa has a wide interval, [−.01, .91], for the same reason; a study that needs to tell assessors apart on such criteria needs more performances where learners fall short, such as the first practice period. Assessor 2's judgments alone did not give a significant difference, $p = .087$; the final judgments rest on the outside assessor's reading of the coded clips for the 130 judgments on which the two assessors differed.

What We Would Change

We would add a reassessment of a sample of learners a month later, train the class in marking with two planted-error demonstrations instead of one, and, with more sections, give each practice to several sections in each unit, so that a section's response to one dish weighs less.

Conclusion

Peer checking against the performance criteria, with the same practice time and the same number of cooking turns, raised the share of Grade 10 learners judged competent on a whole practical task, chiefly on the routines of mise en place, cleaning, heat, and plating. The TLE department will use the checklists and peer checking in its Cookery and Food Preparation practicals from School Year 2026–2027.

Action Plan

Table 7

Action Plan

Objective	Activities	Persons involved	Timeline	Resources (PhP)	Success indicator
Use peer checking in every practical unit	Peer checking with each unit's checklist in the practice periods of the Grade 9 and 10 TLE Food Preparation term (salads, soups, bread)	TLE teachers	June 2026 to March 2027	1,200 (printing of checklists)	Checklists completed on at least 85% of cooking turns
Share the checklists and the agreement check with the TLE department	LAC session: each TLE teacher judges the same recorded practice performances with a checklist and compares, criterion by criterion	TLE department, school head	June 2026	300 (printing)	Teachers agree on at least 85% of judgments in the second round
Find out whether the gain lasts	Add to the next study a second assessment one month after each unit, with the same assessors	Researchers	SY 2026–2027	Included below	A delayed assessment in the approved proposal
Test peer checking across more classes	Proposal for a study of four Grade 9 TLE classes in the Food Preparation term, two drawn by lot for peer checking in each unit	Researchers, school research committee	Next call for proposals	Up to 30,000 (BERF)	Proposal submitted and approved

References

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Connor, R. J. (1987). Sample size for testing differences in proportions for the paired-sample design. *Biometrics, 43*(1), 207–211. https://doi.org/10.2307/2531961

Department of Education. (2016a). *K to 12 basic education curriculum, junior high school technical livelihood education and senior high school technical-vocational-livelihood track: Home economics – cookery (NC II)* [Curriculum guide]. https://www.deped.gov.ph/wp-content/uploads/2019/01/Cookery-NC-II-CG.pdf

Department of Education. (2016b). *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). *Revised implementing guidelines on the direct release, use, monitoring and reporting of maintenance and other operating expenses allocation of schools, including other funds managed by schools* (DepEd Order No. 008, s. 2019). https://www.deped.gov.ph/2019/05/02/may-2-2019-do-008-s-2019-revised-implementing-guidelines-on-the-direct-release-use-monitoring-and-reporting-of-maintenance-and-other-operating-expenses-allocation-of-schools-including-other-fu/

Department of Education. (2024a). *Policy guidelines on the implementation of the MATATAG curriculum* (DepEd Order No. 010, s. 2024). https://www.deped.gov.ph/wp-content/uploads/DO_s2024_010.pdf

Department of Education. (2024b). *Amendment to DepEd Order No. 010, s. 2024* (DepEd Order No. 012, s. 2024). https://www.deped.gov.ph/wp-content/uploads/DO_s2024_012.pdf

Department of Education. (2024c). *Reiteration of the "no collection policy" in schools* (DepEd Memorandum No. 041, s. 2024). https://www.deped.gov.ph/2024/08/02/august-2-2024-dm-041-s-2024-reiteration-of-the-no-collection-policy-in-schools/

Department of Education. (2026). *Foundational guidelines on artificial intelligence in basic education* (DepEd Order No. 003, s. 2026). https://www.deped.gov.ph/2026/02/20/february-20-2026-do-003-s-2026-foundational-guidelines-on-artificial-intelligence-ai-in-basic-education/

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*(2), 481–509. https://doi.org/10.1007/s10648-019-09510-3

Gart, J. J. (1969). An exact test for comparing matched proportions in crossover designs. *Biometrika, 56*(1), 75–80. https://doi.org/10.1093/biomet/56.1.75

Hills, M., & Armitage, P. (1979). The two-period cross-over clinical trial. *British Journal of Clinical Pharmacology, 8*(1), 7–20. https://doi.org/10.1111/j.1365-2125.1979.tb05903.x

Iserbyt, P., Elen, J., & Behets, D. (2009). Peer evaluation in reciprocal learning with task cards for acquiring basic life support (BLS). *Resuscitation, 80*(12), 1394–1398. https://doi.org/10.1016/j.resuscitation.2009.07.006

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Newcombe, R. G. (1998a). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine, 17*(8), 873–890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8%3C873::AID-SIM779%3E3.0.CO;2-I

Newcombe, R. G. (1998b). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine, 17*(22), 2635–2650. https://doi.org/10.1002/(SICI)1097-0258(19981130)17:22%3C2635::AID-SIM954%3E3.0.CO;2-C

Republic Act No. 10173. (2012). *Data Privacy Act of 2012*. https://lawphil.net/statutes/repacts/ra2012/ra_10173_2012.html

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 92–96). https://doi.org/10.25080/Majora-92bf1922-011

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Technical Education and Skills Development Authority. (2014). *Training regulations: Cookery NC II* (Amended). https://www.tesda.gov.ph/Downloadables/TR%20Cookery%20NC%20II.pdf

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software, 3*(31), 1026. https://doi.org/10.21105/joss.01026

World Health Organization. (2006). *Five keys to safer food manual.* https://www.who.int/publications/i/item/9789241594639

Financial Report

Table 8

Statement of Expenses

Date	Particulars	Receipt no.	Amount (PhP)	Fund source
July 7, 2025	Consent letters in English and Filipino and assent scripts, 90 copies	OR 2146	180	Researchers
July 21, 2025	Performance checklists for peer checking, 320 copies, and scoring sheets for the assessors, 340 copies	OR 2178	660	Researchers
July 21, 2025	Five phone stands and a 128 GB memory card for the station recordings	OR 0937	1,150	Researchers
July 25, 2025	Snacks for the assessors' two calibration sessions, 4 persons	OR 5512	480	Researchers
July 26, 2025	Eggs, 27 trays of 30, kept refrigerated; cooking oil, salt and pepper (Unit 1)	OR 7741	7,000	School MOOE
July 26, 2025	LPG refills, two 11-kg tanks (both units)	OR 3305	2,240	School MOOE
July 29 to August 7, 2025	Onions and tomatoes for the Unit 1 practice periods, bought the morning of each period (4 receipts)	OR 7752 to 7755	560	School MOOE
August 11 to 15, 2025	Onions and tomatoes for the Unit 1 assessment days and make-up day, bought each morning (5 receipts)	OR 7761 to 7765	400	School MOOE
August 29, 2025	Garlic, onions, cooking oil, salt and pepper (Unit 2)	OR 7790	1,240	School MOOE
September 2 to 11, 2025	Carrots, cabbage, chayote and string beans for the Unit 2 practice periods, bought the morning of each period (4 receipts)	OR 7791 to 7794	4,760	School MOOE
September 15 to 19, 2025	The same vegetables for the Unit 2 assessment days and make-up day, bought each morning (5 receipts)	OR 7801 to 7805	3,840	School MOOE
March 12, 2026	Printing and binding of the report, 3 copies	OR 2311	900	Researchers
	Researchers' own funds		**3,370**	
	School MOOE (TLE consumables)		**20,040**	

Date	Particulars	Receipt no.	Amount (PhP)	Fund source
Total			**23,410**	

Note. The ingredients and the LPG were the regular TLE consumables for both units, planned in the school's Annual Implementation Plan and bought by the school; both sections used them alike, and no learner paid for or brought ingredients. The researchers paid the study's own costs. Receipt numbers are invented for this sample.

Appendices

Appendix A. Data Set and Notebook

The Excel data set (the learners, both assessors' judgments of all 152 performances, the third assessor's judgments, the rubric scores, the practice, fidelity, calibration, and safety records, and every statistic Excel can compute as a formula) and the Jupyter notebook that reproduces every result, the power calculation from its stated assumptions included.

Appendix B. Consent Letter and Assent Script

Kept with the school's copy, with the signed forms, and left out of the archived report.

Appendix C. Performance Checklists and Product Rubric

The ten performance criteria for each dish, each judged Competent or Not Yet Competent, and the four-attribute rubric with its descriptors, as one scoring sheet.

Appendix D. Recipe Cards

The vegetable omelet and the stir-fried mixed vegetables, with ingredients, quantities, steps, and the plating, as given to the learners.

Appendix E. Lesson Outline and Practice Records

Each unit's lessons, the practice period's plan for each practice, the practice log, and the fidelity checklist.

Appendix F. Declaration of AI Use

We, the researchers, declare that we used [name of the AI tool] only to check the grammar of the English text of this report, for the purpose of finalizing the research study. The prompt used was: "Check the grammar of

this paragraph. Do not change its meaning or add content." We have reviewed, edited, and understood all the content, and we take full accountability for it. No checklists, names, scores, recordings, or other personal information were entered into the AI tool (adapted from Annex A of DepEd Order No. 003, s. 2026).