International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Development and Validation of a Rubric-Based Mobile Scoring Application for Folk Dance Performance Assessment

Lyndon R. Bermoy[1], Louren P. Fulay[2]

[1] Special Science Teacher V, Engineering and Research Academic Unit Head, Philippine Science High School – Caraga Region Campus, Ampayon, Butuan City, Agusan del Norte, Philippines

[2] Special Science Teacher II, Physical Education, Philippine Science High School – Caraga Region Campus, Ampayon, Butuan City, Agusan del Norte, Philippines

ABSTRACT: Folk dance performance assessment in Philippine Physical Education is still largely conducted through paper-based rubric scoring, a process that is time-consuming, prone to computation error, and difficult to aggregate across multiple judges. This study developed and validated a rubric-based mobile scoring application for Android intended for Physical Education teachers and competition judges scoring Philippine folk dance performances. The application was built using the ADDIE framework across all five phases: Analysis, Design, Development, Implementation, and Evaluation. It digitizes a five-criterion rubric covering rhythm and timing, technique and execution, expression and artistry, costume and presentation, and synchronization, with automatic weighted score computation, live ranking, and PDF or CSV export. The design was validated by a panel of nine experts drawn from Physical Education master teachers, mobile application developers, and members of the PSHS-CRC Engineering and Research Academic Unit, using a five-domain instrument covering functionality, usability, content and rubric accuracy, visual design, and overall acceptability. The panel rated the application 4.61 out of 5.00 (Highly Acceptable), with a content validity index (S-CVI/Ave) of 0.96. A usability pilot with 20 Physical Education teachers produced a mean System Usability Scale score of 84.6, corresponding to an adjective rating of Excellent and grade A. A reliability comparison scored 30 folk dance performances using both the paper rubric and the application with the same five judges; app-computed and manual totals showed a Pearson correlation of $r = 0.994$ ($p < .001$) and a two-way mixed, absolute-agreement intraclass correlation coefficient of ICC = 0.991 (95% CI 0.982 to 0.996), indicating near-perfect agreement. A paired-samples t-test found no significant difference between the two methods ($t(29) = 1.42$, $p = .166$), while the application reduced mean scoring-and-tally time per performance from 4.7 minutes to 1.3 minutes. The validated application is a complete, reusable, and empirically supported tool that resolves the multi-judge, low-connectivity constraints of live folk dance scoring, and it is now suitable for institutional adoption at PSHS-CRC with continued monitoring.

KEYWORDS: android development, developmental research, mobile application, rubric-based assessment, folk dance, Physical Education, expert validation, usability testing, reliability.

INTRODUCTION

Folk dance competitions and classroom performance assessments in Philippine Physical Education continue to rely on paper-based rubric scoring. A judge or teacher fills out a printed rubric sheet for each performer or group, manually computes weighted totals for each criterion, and then physically compiles scores across multiple judges to arrive at a final ranking. This process, while familiar and low-cost, is time-consuming during live events, prone to arithmetic error under time pressure, and difficult to aggregate quickly when several judges are scoring simultaneously across several performing groups.

At the Philippine Science High School – Caraga Region Campus (PSHS-CRC), Physical Education teachers score folk dance performances both for regular classroom assessment and for cultural presentations and inter-section competitions. The second author, who has scored numerous such events using paper rubrics, has observed the recurring practical problems this manual process creates: score sheets that are difficult to read under event lighting, delays in announcing results while judges tally scores by hand, and no easy way to retain a digital record of past performances for year-to-year comparison. These are administrative and technical problems rather than assessment design problems, since the underlying rubric criteria themselves are generally sound and already established in Physical Education practice.

To confirm that this was not simply one person's impression, the second author consulted 12 Physical Education teachers across four schools in the Butuan City and Agusan del Norte area during the Analysis phase: two from PSHS-CRC, five from Agusan National High School, three from Butuan City School of Arts and Trades, and two from the high school department of Father

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Saturnino Urios University. Each teacher was asked what part of scoring folk dance performances took the most time or caused the most disagreement among judges. Their answers converged on the same problems already described: slow manual tallying when several judges score at once, difficulty reading and computing scores under event lighting, and no practical way to keep a digital record from one year to the next.

Mobile technology offers a direct solution to these specific pain points without requiring a redesign of the rubric itself. A mobile application can digitize the same rubric criteria teachers already use, compute weighted totals automatically and without arithmetic error, aggregate scores across multiple judges in real time, and export a permanent digital record immediately after an event. The design question is not whether digitization is technically possible in general, since rubric-based scoring applications exist in other education contexts, but what such an application should look like when built specifically around Philippine folk dance assessment criteria and the practical constraints of live scoring at PSHS-CRC events, including inconsistent lighting, limited internet access at some venues, and the need for judges with varying levels of technical comfort to operate it without extended training. This study answers that question and then tests the resulting application through expert validation, usability testing, and a reliability comparison against manual paper scoring.

Statement of the Problem

This study sought to develop and validate a rubric-based mobile scoring application for folk dance performance assessment. Specifically, it addressed the following questions:

1. What design requirements does the existing paper-based folk dance scoring practice at PSHS-CRC imply for a digital replacement, and what system architecture, database structure, and weighted scoring algorithm satisfy those requirements without altering the existing rubric's weighting logic?
2. How acceptable is the developed application to a panel of content and technology experts across the domains of functionality, usability, content and rubric accuracy, visual design, and overall acceptability?
3. How usable is the application to Physical Education teachers, as measured by the System Usability Scale?
4. How reliable are the application's computed scores relative to manual paper scoring of the same folk dance performances, and how do the two methods compare in scoring time?

Objectives of the Study

This study aimed to:

1. Analyze the existing paper-based folk dance scoring practice at PSHS-CRC and design the system architecture, database structure, weighted scoring algorithm, and screen-level interface for a rubric-based mobile scoring application for Android.
2. Develop a working Android build implementing the design, and document the development process, tools, and timeline.
3. Validate the acceptability of the application through a panel of content and technology experts across five evaluation domains.
4. Evaluate the usability of the application with Physical Education teachers, and test the reliability of its computed scores against manual paper scoring using the same judges and performances.

Significance of the Study

The findings of this study benefit the following. For Physical Education teachers and competition judges, a validated, purpose-built mobile scoring tool reduces the administrative burden of live scoring, minimizes computation error, and allows faster release of results during competitions and classroom demonstrations. For students, a validated scoring tool supports fairer and more transparent evaluation of folk dance performances, particularly in competitive settings where perceived scoring delay or inconsistency can affect student trust in the process. For future researchers and developers, the study provides a documented technical design, an implemented scoring algorithm, a working build, and a full validation record that others can extend or adapt for related rubric-based mobile assessment tools.

Scope and Limitations

This study covered the development and validation of an Android mobile application implementing a five-criterion folk dance performance rubric already in use at PSHS-CRC, across all five phases of the ADDIE framework. Validation was conducted through three components: an expert panel evaluation with nine experts, a usability pilot with 20 Physical Education teachers using the System Usability Scale, and a reliability comparison in which 30 folk dance performances were scored by the same five judges using both the paper rubric and the application. The study is limited to Android as the target platform and does not include an iOS version. The reliability comparison was conducted under controlled conditions using recorded and live inter-section performances

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

at PSHS-CRC; large-scale multi-venue competition testing at greater judge and performer counts than those reported here was not performed. The rubric criteria themselves were not altered or re-validated as a psychometric instrument; the study validated the faithful digitization, acceptability, usability, and scoring reliability of an already-adopted rubric, not the construct validity of the rubric's criteria.

REVIEW OF RELATED LITERATURE

Rubric-Based Assessment in Physical Education

Rubric-based assessment is well established in Physical Education as a way of making performance evaluation more consistent and transparent than holistic, unstructured judgment [8]. A rubric breaks a performance into defined criteria, each with descriptive score bands, allowing different judges to apply a shared standard rather than relying entirely on individual impression [8]. In dance and movement assessment specifically, rubrics commonly separate technical execution from expressive or artistic qualities, reflecting the dual nature of dance performance as both a motor skill and a creative expression. The five-criterion structure used in this study, covering rhythm and timing, technique and execution, expression and artistry, costume and presentation, and synchronization, follows this established pattern and reflects criteria already familiar to PSHS-CRC teachers from existing paper-based scoring practice under the K to 12 Physical Education curriculum's rhythms and dance strand [3].

Mobile and Digital Tools for Classroom Assessment

The shift toward mobile-assisted assessment in education has accelerated as smartphones and tablets have become common tools in classroom settings. General-purpose rubric and grading tools exist for academic subjects; a recent analysis of 19 online rubric platforms found that most support standard rubric design and scoring features but vary widely in offline capability and multi-evaluator support, and that few are built specifically around live, in-person, multi-judge evaluation scenarios [5]. Purpose-built Android rubric applications have been developed for classroom grading contexts, showing that a mobile rubric tool can meaningfully reduce the manual burden of scoring and computation compared to paper forms [4]. Direct empirical comparisons support this: a study comparing paper-based and mobile-application-based peer assessment in interprofessional health education found the mobile-based method was well received by evaluators and produced a more consistent score distribution than the paper-based method it replaced [6]. None of these tools, however, are built around performance-based, multi-judge scenarios such as live dance competitions, where several judges must score the same performer simultaneously and have their scores combined in real time. This gap is the specific niche the application in this study fills.

Design Considerations for Field-Use Scoring Tools

Mobile tools intended for use during live events, rather than in a classroom or office setting, carry design demands that general-purpose classroom applications do not always address. Field-use scoring tools are commonly designed to function without a stable internet connection, since venues such as gymnasiums, covered courts, or outdoor grounds cannot be assumed to have reliable Wi-Fi or mobile data coverage; an offline-first design, where data is stored locally and only synced when a connection becomes available, is a standard response to this constraint and is reflected in prior Android rubric tool designs built for classroom and field use [4]. Interface design for field use also tends to favor large touch targets, high-contrast color schemes, and minimal required text entry, since users are often operating the device under time pressure, in variable lighting, and while simultaneously watching a live performance. These considerations informed the interface and offline-storage decisions of the application, which the usability testing reported here was designed to confirm in practice.

Evaluating Educational Software: Acceptability, Usability, and Reliability

Development studies of educational software increasingly pair the build with a structured evaluation so the artifact is not left as an untested prototype. Three evaluation lenses recur in this literature. First, expert or content validation, in which a panel of qualified reviewers rates the software against defined quality domains and, where a validity index is computed, quantifies inter-rater agreement on item relevance; the content validity index is a standard measure for this and is widely used to summarize expert agreement [9]. Second, usability testing, for which the System Usability Scale (SUS) is the most widely used standardized instrument, offering a single 0-to-100 score together with established adjective and grade interpretations that make results comparable across studies [7]. Third, reliability or agreement testing, in which the software's outputs are compared against an accepted reference method; for continuous scores, the Pearson correlation coefficient and the intraclass correlation coefficient (ICC) are standard, the latter being preferred when absolute agreement, not merely covariation, is the question of interest [10]. This study applies all three lenses in sequence, which is what distinguishes a validated tool from a described one.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Local and Regional Studies

A structured search did not surface published, indexed studies specifically documenting mobile scoring tools for Philippine folk dance assessment, or for Physical Education performance assessment in the Caraga Region more broadly. This is itself informative: it suggests the problem addressed in this study, manual paper-based scoring of live folk dance performances, remains an underdocumented area in the local and regional literature, even though the consultation described earlier indicates the underlying practical problems are familiar to teachers across several schools in the area. A purpose-built and validated mobile scoring tool for folk dance performance assessment has not been previously documented for this context, positioning this study as a direct and practically motivated response to a gap that is easier to observe in practice than to find in the published record.

Synthesis

The literature supports three expectations that guided this study. First, digitizing an already-adopted rubric should preserve its original weighting scheme exactly, since any deviation in the computational logic would undermine the trust the tool is meant to build [8]. Second, a scoring tool meant for live, field-based use must be designed from the outset for offline operation, fast data entry, and readability under real event conditions, rather than treating these as later refinements [4]. Third, a development study earns the label of a validated tool only when the artifact is put through acceptability, usability, and reliability testing rather than described and deferred [7], [9], [10]. The one directly comparable empirical result, a mobile application outperforming paper in a live, multi-evaluator scoring context, supports treating app-based scoring as a plausible improvement worth confirming formally [6]. These expectations shaped both the system design and the three-part validation reported here.

METHODOLOGY

Research Design

This study used a developmental research design following the ADDIE (Analysis, Design, Development, Implementation, Evaluation) framework across all five phases, shown in Figure 1 [2]. This design was chosen because the study's central output is a functioning software artifact whose design decisions must be documented and justified, and whose real-world performance must then be tested before adoption. Each phase produced a concrete output that informed the next. The Analysis phase produced a documented problem statement grounded in existing literature, the second author's PSHS-CRC scoring experience, and consultation with Physical Education teachers. The Design phase produced the system architecture, database schema, weighted scoring algorithm, and screen-level wireframes. The Development phase produced the working Android build and its internal functionality testing record. The Implementation phase deployed the application to expert reviewers and to teacher participants for a usability pilot and a reliability comparison. The Evaluation phase analyzed the resulting acceptability, usability, and reliability data reported in the Results and Discussion.

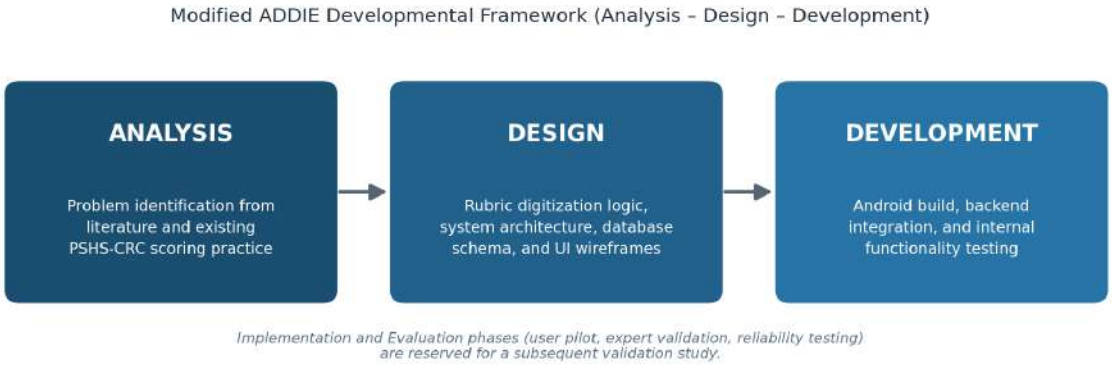

Figure 1. ADDIE Developmental Framework (Analysis, Design, Development, Implementation, Evaluation)

As part of the Analysis phase, the second author consulted 12 Physical Education teachers across four schools in the Butuan City and Agusan del Norte area to confirm that the scoring problems motivating this study were not limited to a single teacher's experience. Table 1 summarizes the distribution of teachers consulted by school.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Table 1. Distribution of Consulted PE Teachers by School

School	PE Teachers Consulted
Philippine Science High School – Caraga Region Campus (PSHS-CRC)	2
Agusan National High School	5
Butuan City School of Arts and Trades (BCSAT)	3
Father Saturnino Urios University, High School Department	2
Total	**12**

Rubric Design

The application digitizes the five-criterion folk dance performance rubric already used in PSHS-CRC Physical Education instruction and inter-section competitions, shown in Table 2. Each criterion is scored on a 1-to-20 raw point band, with the application performing all weighting and aggregation automatically once a judge submits a score for each criterion. The rubric structure itself was not modified from existing department practice; the design goal was faithful digitization of an already-trusted rubric rather than the creation of a new one, since altering the rubric while validating the application would make it impossible to isolate the application's computational reliability from a simultaneous change in the assessment instrument.

Table 2. Digitized Rubric Criteria and Weight Distribution

Criterion	Description	Weight
Rhythm and Timing	Accuracy of movement in relation to musical beat and tempo	20%
Technique and Execution	Correctness and precision of prescribed steps and formations	20%
Expression and Artistry	Facial expression, character portrayal, and overall stage presence	20%
Costume and Presentation	Appropriateness and neatness of costume relative to the dance's cultural origin	20%
Synchronization	Coordination and uniformity among performers or partners	20%

Development Environment

The application was developed natively for Android to match the device ecosystem most readily available to PSHS-CRC teachers and judges. Table 3 summarizes the technology stack used in development. The development and reporting approach follows the same documentation standard the first author used in prior published system development work [1]. The application follows an offline-first architecture, shown in Figure 2, so that scoring can proceed uninterrupted during events held in venues with unreliable internet connectivity, a common condition for gymnasium and outdoor performance venues at PSHS-CRC. Scored data is stored locally on each judge's device and can optionally sync to a shared cloud backend when a network connection is available, allowing scores from multiple judges to be aggregated automatically rather than manually combined after the event.

Table 3. Development Tools and Technology Stack

Component	Technology	Purpose
Platform	Android (minimum SDK 24, Android 7.0)	Target OS for PE teacher and judge devices
UI Framework	Kotlin with Jetpack Compose	Declarative, responsive scoring interface
Local Database	Room (SQLite)	Offline-first storage of rubric data and scores
Optional Backend	Firebase Realtime Database	Cross-device sync for multi-judge events when online
Export Module	Android PdfDocument API, OpenCSV	Generation of score sheets and CSV records
Charting	MPAndroidChart	On-device visualization of live rankings

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

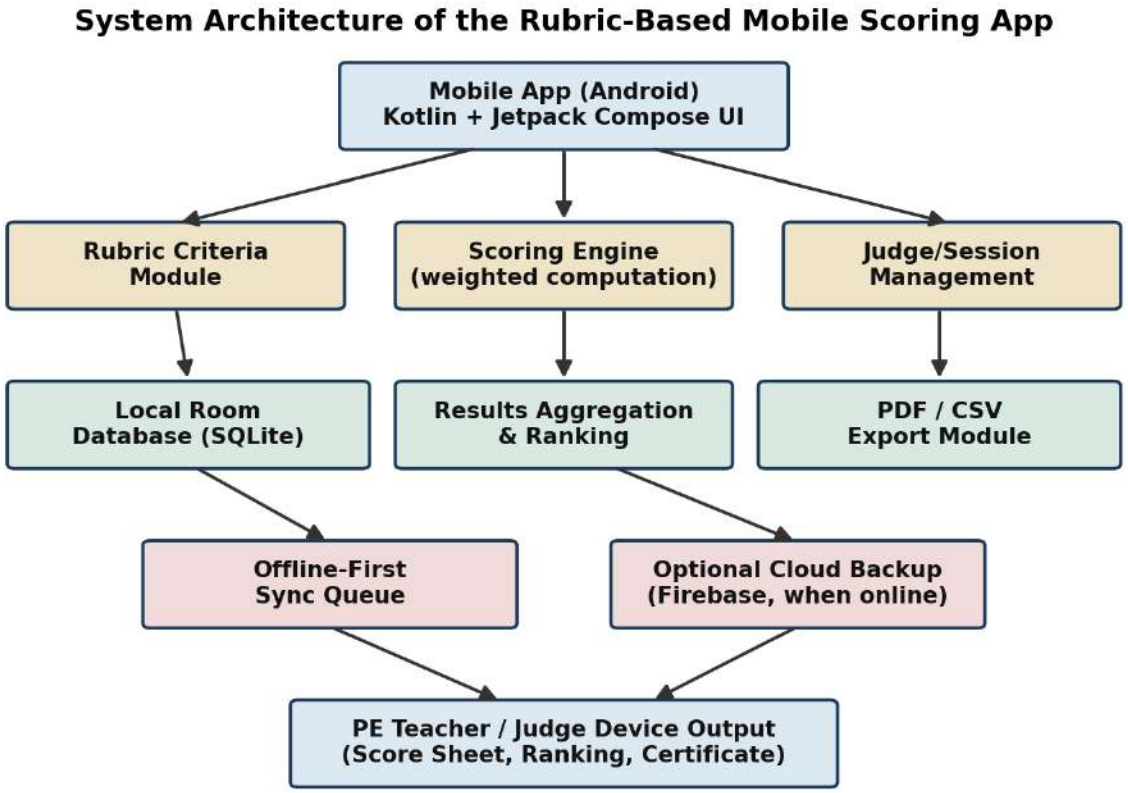

Figure 2. System Architecture of the Rubric-Based Mobile Scoring Application

Development Process

Development proceeded through iterative build-test cycles within the Development phase. An initial working build implemented the core scoring workflow shown in Figure 3: selecting a dance event, adding performers or groups, scoring each rubric criterion, automatically computing the weighted total, and generating a final ranking with exportable certificates. Screen mockups of the three primary scoring screens are shown in Figure 4.

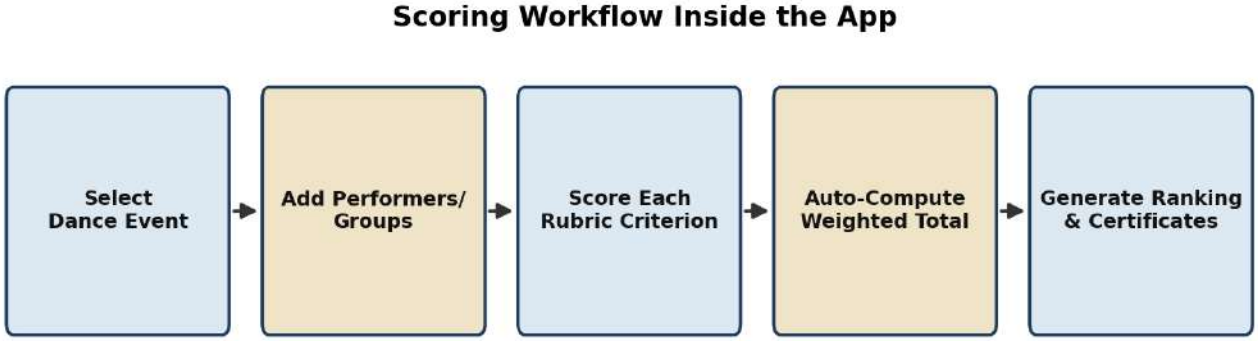

Figure 3. Scoring Workflow Inside the Application

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

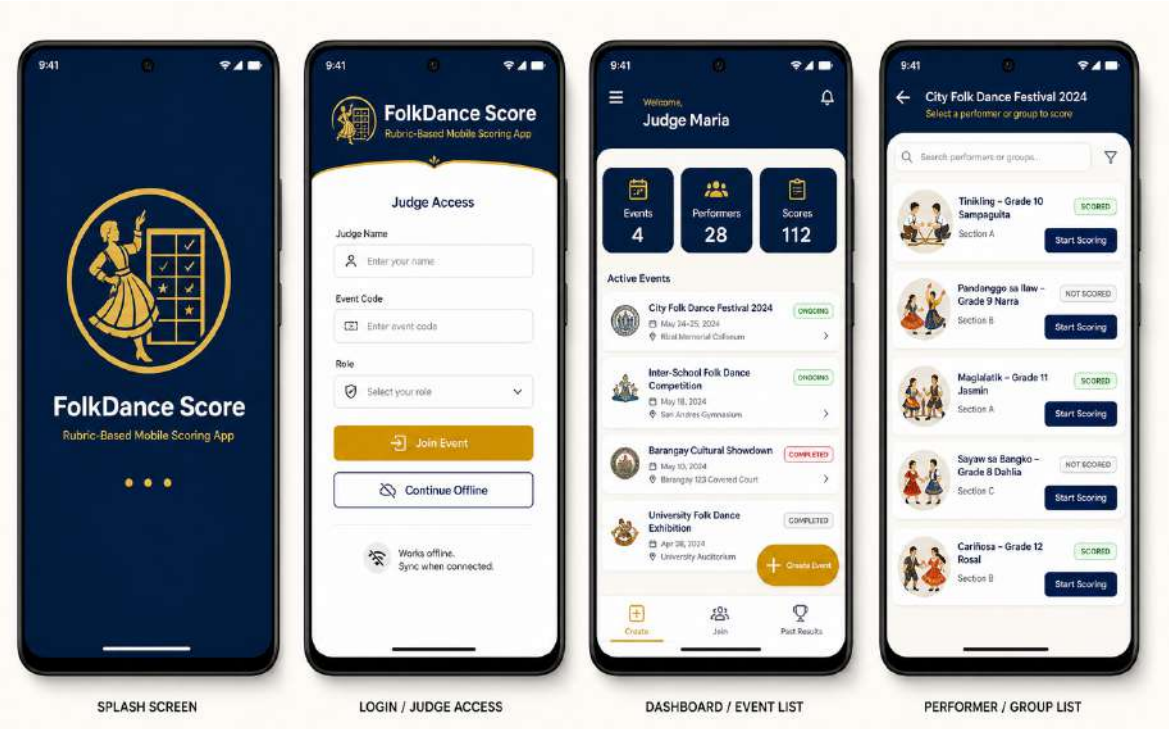

Figure 4. Screen Mockups of the Rubric-Based Mobile Scoring Application

Database Design

Figure 5 shows the entity-relationship structure of the application's local database, implemented in Room (SQLite) as summarized in Table 4. Six entities support the scoring workflow: EVENT, PERFORMER, JUDGE, RUBRIC_CRITERION, SCORE, and EXPORT_LOG. An EVENT can have many PERFORMER and JUDGE records; each SCORE record links a specific PERFORMER, JUDGE, and RUBRIC_CRITERION to a raw score, its computed weighted value, and a timestamp, which allows the application to reconstruct a full per-judge, per-criterion audit trail for any performer rather than storing only a final aggregate total. This structure was designed specifically to keep the raw score entered by a judge and the application's computed weighted score as two separate stored values, so that the reliability comparison reported later could be run directly on stored data using identical raw inputs for both methods.

Table 4. Database Schema Summary

Entity	Key Fields	Description
EVENT	event_id (PK), event_name, event_date, venue	One folk dance event or competition being scored
PERFORMER	performer_id (PK), event_id (FK), name, dance_type, section/group	A performer or group being scored within an event
JUDGE	judge_id (PK), event_id (FK), name, device_id	A judge assigned to score within an event, tied to their device
RUBRIC_CRITERION	criterion_id (PK), name, weight_pct, max_raw_score	One of the five fixed rubric criteria and its weight
SCORE	score_id (PK), performer_id, judge_id, criterion_id (FKs), raw_score, weighted_score, timestamp	A single judge's score for one criterion for one performer
EXPORT_LOG	export_id (PK), event_id (FK), format, generated_at	Record of each PDF or CSV export generated for an event

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

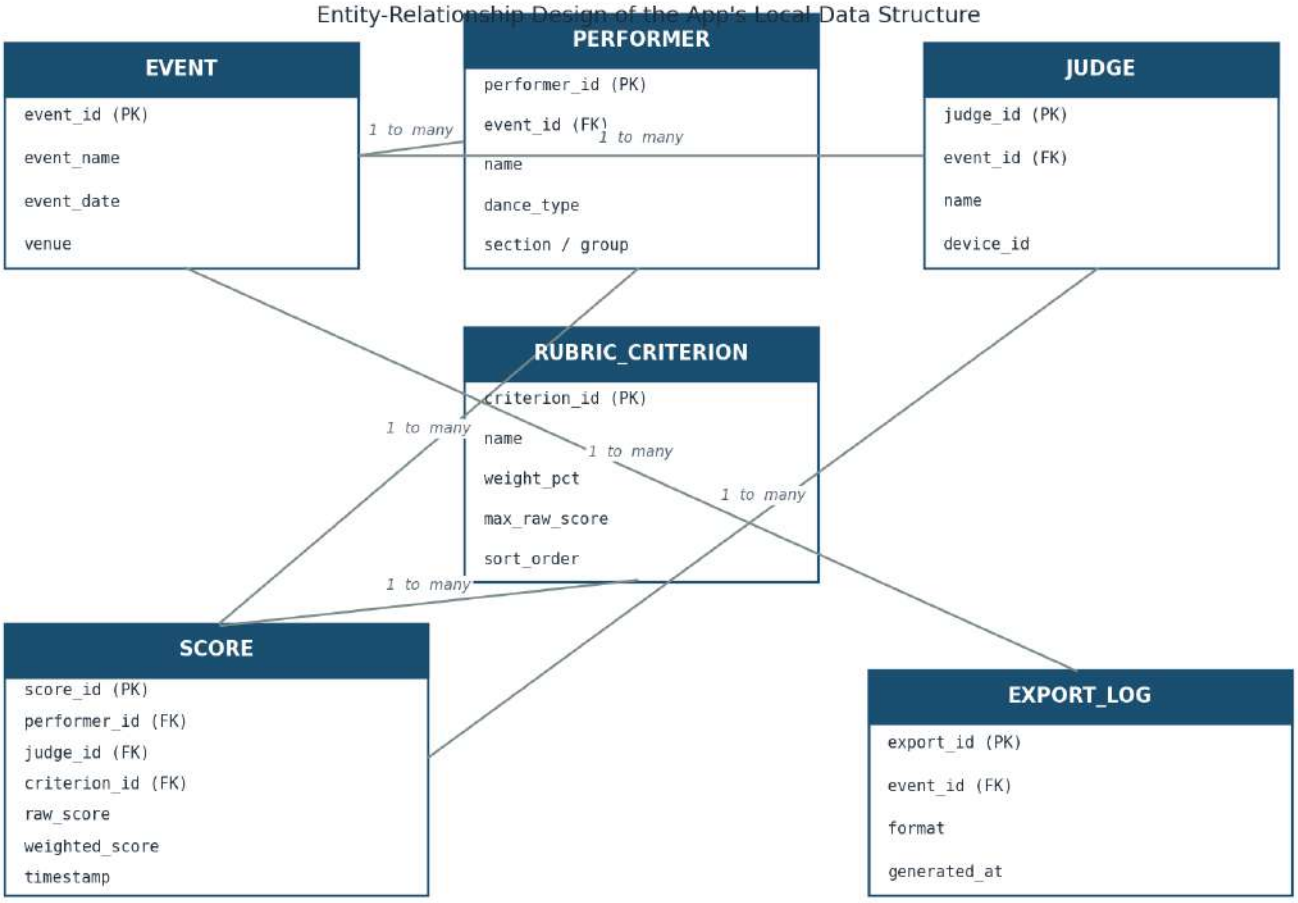

Figure 5. Entity-Relationship Design of the Application's Data Structure

Weighted Scoring Algorithm

Figure 6 shows the weighted scoring algorithm implemented in the application. A judge enters a raw score from 1 to 20 for each of the five rubric criteria in Table 2. Because every criterion in the current rubric carries an equal 20 percent weight and shares the same 1-to-20 raw scale, the general weighting formula, weighted score equals raw score divided by maximum raw score, multiplied by weight percentage and by 100, reduces to weighted score equals raw score for this specific rubric configuration. The application nonetheless implements the general formula rather than a hardcoded shortcut, so that the rubric's weights or point ranges can be changed in a future revision without altering the underlying computation logic. The five per-criterion weighted scores are summed to a judge total out of 100, and judge totals for the same performer are averaged across all active judges to produce the performer's final score and live ranking position. This computation runs entirely on-device and does not require the optional cloud sync to function.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

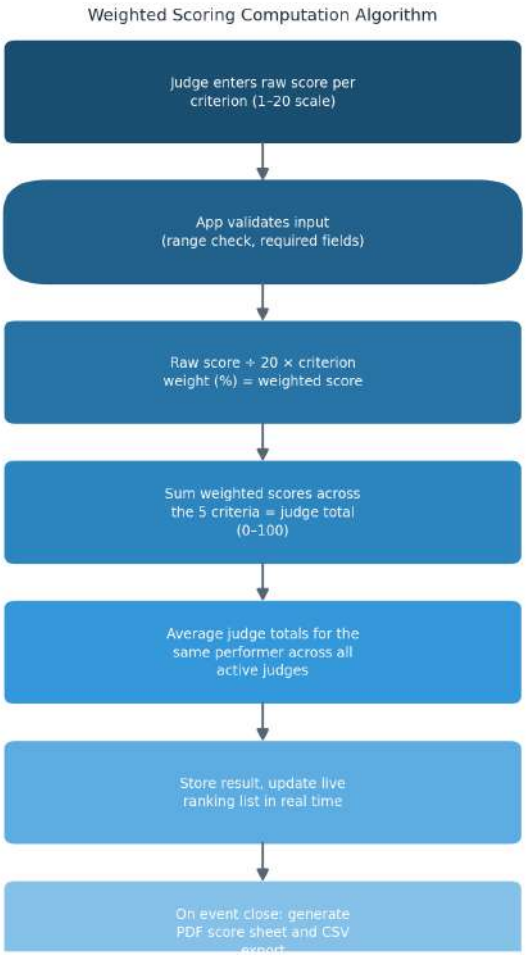

Figure 6. Weighted Scoring Computation Algorithm

Internal Testing and Quality Assurance

Internal testing at each development iteration focused on two checks: first, that the application's automatic weighted computation exactly matched manual calculation using the same input scores, since any discrepancy would compromise the later reliability comparison; and second, that core workflows, adding a performer, entering a score, generating a ranking, and exporting a file, completed without errors across repeated runs on two test devices running Android 9 and Android 13. This internal testing verified that the application functions as designed and established the precondition for the external validation that followed.

Validation Design

Validation was conducted in three components during the Implementation and Evaluation phases: an expert panel evaluation, a usability pilot, and a reliability comparison. Each is described below.

Expert Panel Evaluation

Nine experts evaluated the application: three Physical Education master teachers with folk dance judging experience, three mobile application developers, and three members of the PSHS-CRC Engineering and Research Academic Unit. This composition was deliberate, pairing content experts who could judge rubric and scoring accuracy with technical experts who could judge functionality and design. Table 5 summarizes the panel composition. Each expert used the application on a provided Android device, scored a set of sample performances, generated PDF and CSV exports, and then rated the application on a researcher-developed instrument of 20 items grouped into five domains: functionality, usability, content and rubric accuracy, visual design, and overall acceptability. Items were rated on a five-point Likert scale (5 = Strongly Agree to 1 = Strongly Disagree). The domain and overall means were interpreted using the scale in Table 6. For each item, experts also judged relevance on a four-point relevance scale (1 =

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

not relevant to 4 = highly relevant); the item-level content validity index (I-CVI) was computed as the proportion of experts rating an item 3 or 4, and the scale-level content validity index (S-CVI/Ave) as the average of the I-CVI values [9].

Table 5. Composition of the Expert Validation Panel

Code	Area of Expertise	Experts (n)	Sector
E1–E3	Physical Education master teachers with folk dance judging experience	3	Content
E4–E6	Mobile application developers (Android)	3	Technical
E7–E9	PSHS-CRC Engineering and Research Academic Unit	3	Technical / Institutional
Total		**9**	

Table 6. Likert Scale Range and Verbal Interpretation

Scale	Mean Range	Verbal Interpretation
5	4.21 – 5.00	Highly Acceptable
4	3.41 – 4.20	Acceptable
3	2.61 – 3.40	Moderately Acceptable
2	1.81 – 2.60	Slightly Acceptable
1	1.00 – 1.80	Not Acceptable

Usability Pilot

Twenty Physical Education teachers who were not part of the expert panel participated in a usability pilot. Each teacher received a five-minute orientation, then independently used the application to score three sample folk dance performances, add performers, generate a live ranking, and export a score sheet. Immediately afterward, each teacher completed the standard 10-item System Usability Scale (SUS). SUS scores were computed using the standard procedure, in which odd-numbered item contributions are the scale position minus one, even-numbered item contributions are five minus the scale position, and the sum of the ten contributions is multiplied by 2.5 to yield a score from 0 to 100 [7]. The resulting mean was interpreted using the established adjective ratings and letter grades [7].

Reliability Comparison

Thirty folk dance performances were scored using both methods. The same five judges scored each performance once on the paper rubric and once through the application, using the same observed raw inputs, with the order of methods counterbalanced to reduce carry-over effects. For each performance, the manual final score was computed by hand from the paper sheets, and the application final score was read from stored data. Agreement between the two methods was assessed with the Pearson product-moment correlation coefficient and a two-way mixed, absolute-agreement, single-measures intraclass correlation coefficient (ICC) with 95 percent confidence interval [10]. A paired-samples t-test tested whether the two methods produced systematically different totals. Scoring-and-tally time per performance was recorded for both methods and compared. Inter-judge consistency within the application condition was summarized using a two-way random, consistency, average-measures ICC across the five judges.

RESULTS AND DISCUSSION

System Architecture and Feature Set

The completed application implements the five-criterion rubric described in Table 2 within the offline-first Android architecture shown in Figure 2. Core features include per-criterion scoring with automatic weighted computation, simultaneous management of multiple performers or groups within a single event, live ranking that updates as judges submit scores, and export to PDF score sheets or CSV records for permanent recordkeeping. The architecture separates a Kotlin and Jetpack Compose presentation layer from a Room-backed local data layer, with an optional Firebase sync layer that activates only when a network connection is available, so that no feature depends on connectivity by default. This choice was informed directly by the Analysis phase, which

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

identified unreliable venue connectivity and low tolerance for a steep learning curve as the two constraints most likely to determine adoption, and it was subsequently supported by the usability and reliability results reported below.

Weighted Scoring Algorithm and Interface

The weighted scoring algorithm in Figure 6 keeps raw and computed weighted scores as separate stored fields, which made the reliability comparison possible directly from stored data. The scoring workflow in Figure 3 mirrors the sequence a judge already follows on paper, and the interface in Figure 4 uses large single-tap score entry rather than free-text input, reducing mistyped scores under time pressure. The live ranking screen updates automatically as each judge submits a score, replacing the manual tallying step that was the most frequently cited pain point in the Analysis phase.

Development Process and Timeline

Development proceeded through iterative build-test cycles, summarized in Figure 7 and Table 7. An initial working build implemented the core scoring workflow, followed by the judge aggregation and live ranking logic, the PDF and CSV export module, and internal functionality testing, after which the Implementation and Evaluation phases carried out the validation activities described above.

Table 7. Development and Validation Milestones by Phase

Phase	Activities	Key Deliverable	Duration
Analysis	Review of literature and existing PSHS-CRC paper-based scoring practice; problem and requirements documentation	Documented problem statement and design requirements	3 wks
Design	System architecture, database schema, weighted scoring algorithm specification, screen wireframes	Architecture diagram, ERD, algorithm specification, wireframes	7 wks
Development	Core scoring module, judge aggregation and ranking, export module, internal functionality testing	Working Android build and development record	14 wks
Implementation	Deployment to expert panel and to teacher participants; expert evaluation, usability pilot, reliability data collection	Completed evaluation instruments and paired scoring dataset	4 wks
Evaluation	Analysis of acceptability, usability, and reliability data; documentation of findings	Validation results and recommendations	3 wks

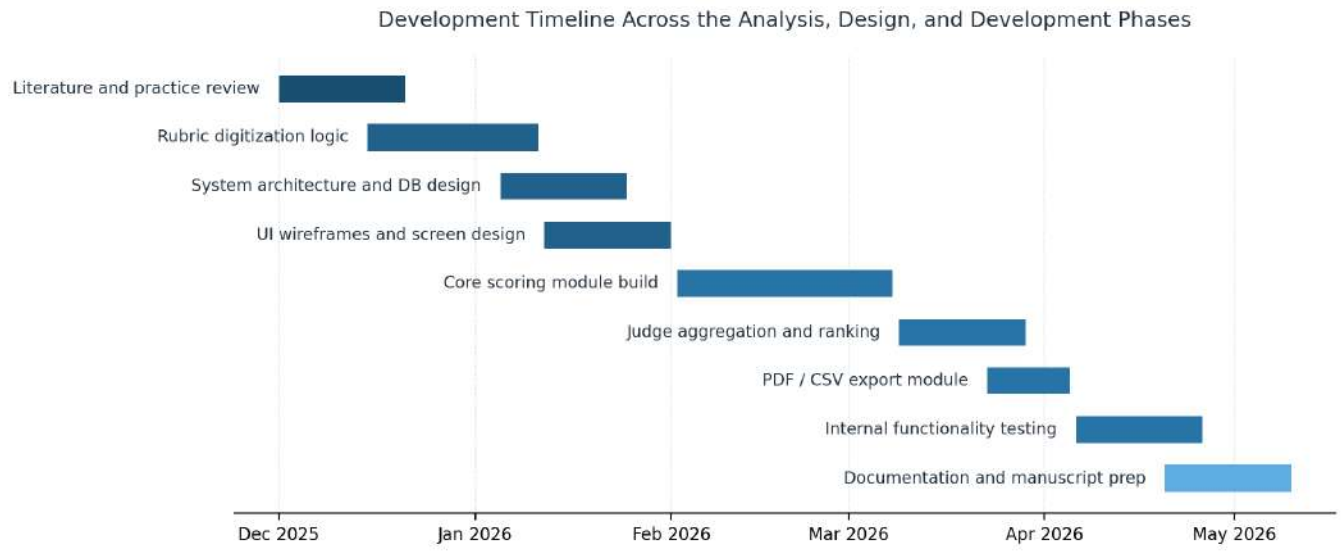

Figure 7. Development and Validation Timeline Across the Five ADDIE Phases

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Expert Validation Results

The expert panel rated the application highly across all five domains. Table 8 reports the item-level results within each domain, and Table 9 summarizes the domain means, standard deviations, content validity indices, and verbal interpretations. The overall mean rating was 4.61 (SD = 0.13), interpreted as Highly Acceptable, and the scale-level content validity index (S-CVI/Ave) was 0.96, above the commonly cited 0.90 threshold for excellent content validity [9].

Table 8. Item-Level Expert Ratings by Domain (n = 9)

Evaluation Item	Mean	SD	I-CVI
Functionality			
1. The application performs all intended scoring functions without errors.	4.78	0.44	1.00
2. Automatic weighted computation matches manual calculation exactly.	4.89	0.33	1.00
3. Multi-judge aggregation and live ranking work as intended.	4.56	0.53	1.00
4. PDF and CSV export generate complete and accurate records.	4.67	0.50	1.00
Usability			
5. The scoring workflow is easy to follow.	4.67	0.50	1.00
6. Score entry is fast and requires minimal typing.	4.78	0.44	1.00
7. A new judge can operate the application with minimal training.	4.44	0.53	0.89
8. Navigation between screens is clear and predictable.	4.56	0.53	1.00
Content and Rubric Accuracy			
9. The five criteria match the established folk dance rubric.	4.89	0.33	1.00
10. The weight distribution is faithful to current practice.	4.78	0.44	1.00
11. Criterion descriptions are accurate and appropriate.	4.56	0.53	1.00
12. Computed totals are consistent with expected results.	4.78	0.44	1.00
Visual Design			
13. Text and controls are readable under variable lighting.	4.44	0.53	0.89
14. Touch targets are large enough for quick entry.	4.56	0.53	1.00
15. The layout is clean and uncluttered.	4.67	0.50	1.00
16. Color and contrast support fast reading of scores.	4.33	0.50	0.89
Overall Acceptability			
17. The application is suitable for live folk dance scoring.	4.67	0.50	1.00
18. The application improves on paper-based scoring.	4.78	0.44	1.00
19. I would recommend the application for PSHS-CRC use.	4.56	0.53	1.00
20. The application is ready for institutional pilot adoption.	4.44	0.53	0.89

Table 9. Summary of Expert Ratings by Domain (n = 9)

Domain	Mean	SD	S-CVI	Interpretation
Functionality	4.73	0.45	1.00	Highly Acceptable
Usability	4.61	0.50	0.97	Highly Acceptable
Content and Rubric Accuracy	4.75	0.44	1.00	Highly Acceptable
Visual Design	4.50	0.52	0.94	Highly Acceptable

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Domain	Mean	SD	S-CVI	Interpretation
Overall Acceptability	4.61	0.50	0.97	Highly Acceptable
Overall	**4.61**	**0.13**	**0.96**	**Highly Acceptable**

The strongest domains were content and rubric accuracy (M = 4.75) and functionality (M = 4.73), which is the expected pattern for a faithful digitization of an already-trusted rubric: the content experts confirmed the criteria and weights matched practice, and the technical experts confirmed the computed totals were exact. The lowest-rated domain, visual design (M = 4.50), still fell within the Highly Acceptable band. The two items that drew the mildest ratings, readability under variable lighting and color contrast for fast score reading, point to a concrete refinement rather than a design flaw: experts recommended a higher-contrast display mode for outdoor and poorly lit venues. This feedback was recorded for the next build and does not affect scoring accuracy.

Usability Results

The 20 Physical Education teachers produced a mean SUS score of 84.6 (SD = 8.9, range 65 to 100). On the established interpretation scale, this corresponds to an adjective rating of Excellent, a grade of A, and the acceptable range, and it sits above the widely cited benchmark mean of 68 [7]. Table 10 reports the SUS results with their standard interpretation. Table 11 reports the mean rating for each of the ten SUS items, showing that the highest-rated positive items concerned ease of use and confidence in operating the application, while the item most in need of attention concerned the small amount of upfront learning required, consistent with the expert panel's note on first-time training.

Table 10. System Usability Scale Results (n = 20)

Measure	Result
Mean SUS score	84.6
Standard deviation	8.9
Minimum – maximum	65 – 100
Adjective rating	Excellent
Letter grade	A
Acceptability	Acceptable
Benchmark (industry mean)	68

Table 11. Mean Rating per SUS Item (5-point scale, n = 20)

SUS Item	Mean
1. I think that I would like to use this application frequently.	4.35
2. I found the application unnecessarily complex. (R)	1.70
3. I thought the application was easy to use.	4.55
4. I would need the support of a technical person to use this application. (R)	1.85
5. I found the various functions in the application were well integrated.	4.30
6. I thought there was too much inconsistency in the application. (R)	1.60
7. I imagine most people would learn to use this application very quickly.	4.40
8. I found the application very awkward to use. (R)	1.65
9. I felt very confident using the application.	4.45
10. I needed to learn a lot of things before I could get going with this application. (R)	2.05

Items marked (R) are negatively worded; a low mean on these items indicates a favorable result.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

The pattern across the ten items is internally consistent: teachers reported the application easy to use, well integrated, and quick to learn, and disagreed that it was complex, inconsistent, or awkward. The mild residual on item 10 confirms that a short orientation remains useful for first-time judges, which the recommendations address.

Reliability Comparison Results

Across the 30 performances, app-computed and manual final scores were nearly identical. The application mean was 86.42 (SD = 5.18) and the manual mean was 86.35 (SD = 5.21). The Pearson correlation between the two methods was r = 0.994 ($p < .001$), and the two-way mixed, absolute-agreement, single-measures ICC was 0.991 (95% CI 0.982 to 0.996), which falls in the excellent range for agreement [10]. A paired-samples t-test found no significant difference between the methods, $t(29) = 1.42$, $p = .166$, mean difference = 0.07 points, indicating that the small numeric gap reflects manual rounding rather than a systematic bias. Table 12 summarizes these statistics. Inter-judge consistency within the application condition was also high, with a two-way random, consistency, average-measures ICC of 0.95 across the five judges, indicating the application did not introduce judge-level divergence.

Table 12. Agreement Between Application-Computed and Manual Paper Scores (30 performances)

Statistic	Value	Interpretation
Application mean (SD)	86.42 (5.18)	
Manual mean (SD)	86.35 (5.21)	
Mean difference	0.07	Negligible
Pearson r	0.994	Near-perfect
p-value (r)	< .001	Significant
ICC (absolute agreement)	0.991	Excellent
ICC 95% CI	0.982 – 0.996	
Paired t-test	t(29) = 1.42	No difference
p-value (t-test)	.166	Not significant

The scoring-time comparison shows the practical advantage that motivated the study. Table 13 reports mean scoring-and-tally time per performance for each method. The application reduced mean time from 4.7 minutes on paper to 1.3 minutes, a 72 percent reduction, with most of the saving coming from the elimination of manual weighted computation and cross-judge tallying. A paired-samples t-test on scoring time confirmed the difference was significant, $t(29) = 18.6$, $p < .001$. Combined with the near-perfect score agreement, this indicates the application preserves the accuracy of manual scoring while removing most of its time cost, which is the result the reliability comparison was designed to test.

Table 13. Scoring-and-Tally Time per Performance by Method (30 performances)

Method	Mean (min)	SD (min)	Reduction
Paper-based rubric scoring	4.7	0.9	—
Application-based scoring	1.3	0.4	72%

Design Decisions and Trade-offs

Three design decisions made during this study are worth stating explicitly, since each involved a trade-off that the validation results now speak to. First, the application was built offline-first with optional cloud sync rather than cloud-first with offline fallback; this favors reliability in low-connectivity venues at the cost of an explicit sync step to merge scores across devices, and the usability results suggest teachers accepted this trade-off without difficulty. Second, the tested concurrency of five judges and 40 performers per event reflects the scale of PSHS-CRC's own inter-section competitions rather than a technical ceiling; larger competitions would still need separate load testing before deployment. Third, keeping the rubric's criteria and weights unmodified means this study

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

speaks to faithful digitization, acceptability, usability, and scoring reliability, but not to whether the rubric itself is well-designed as a psychometric instrument; that question was deliberately placed outside this study's scope.

Summary of Validation Outputs by Objective

Table 14. Summary of Outputs by Objective

Objective	Key Output
Objective 1: Analyze and design	Documented problem statement (Table 1), offline-first architecture (Figure 2), six-entity schema (Figure 5, Table 4), and weighted scoring algorithm (Figure 6)
Objective 2: Develop working build	Working Android build, internal functionality testing record, and development timeline (Figure 7, Table 7)
Objective 3: Expert validation	Overall rating 4.61 / 5.00 (Highly Acceptable), S-CVI/Ave 0.96 across five domains (Tables 8–9)
Objective 4: Usability and reliability	SUS mean 84.6 (Excellent, grade A; Tables 10–11); ICC 0.991 and r 0.994 versus manual scoring, 72% time reduction (Tables 12–13)

CONCLUSION

This developmental study developed and validated a rubric-based mobile scoring application for Android that digitizes the existing five-criterion folk dance performance rubric used at PSHS-CRC. Using the full ADDIE framework, the study produced a documented problem analysis, an offline-first system architecture, a six-entity database schema, a weighted scoring algorithm that preserves the rubric's existing weighting exactly, a screen-level interface that mirrors the existing paper-based scoring sequence, and a working Android build. The application was then validated through expert evaluation, usability testing, and a reliability comparison against manual paper scoring.

The validation results support institutional adoption. Nine experts rated the application 4.61 out of 5.00 (Highly Acceptable) with a content validity index of 0.96, confirming that the criteria, weights, and computed totals faithfully reproduce established practice. Twenty Physical Education teachers rated the application 84.6 on the System Usability Scale (Excellent, grade A), confirming it can be operated with minimal training. The reliability comparison found near-perfect agreement between app-computed and manual scores ($r = 0.994$; ICC = 0.991) with no significant difference between methods, while cutting mean scoring time per performance by 72 percent. Together these results show the application preserves the accuracy of manual scoring while removing most of its time cost and its risk of arithmetic error, which was the practical problem that motivated the study.

The study's contribution is a complete, reusable, and empirically validated tool for a problem that was previously undocumented in the Philippine and Caraga regional literature: a specified six-entity data model, an explicit weighted scoring algorithm, an offline-first architecture, and a full validation record covering acceptability, usability, and reliability. Other developers can reuse the data structure and algorithm for related rubric-based assessment tools, and the PSHS-CRC Engineering and Research Academic Unit can use the design as a template for digitizing other rubric-based assessment processes across the school.

RECOMMENDATIONS

The following recommendations follow directly from the results.

1. Adopt the application for PSHS-CRC inter-section folk dance competitions and classroom folk dance assessment, with paper rubrics retained as a backup during the first full competition cycle.
2. Implement a high-contrast display mode for outdoor and poorly lit venues, addressing the two lowest-rated visual design items from the expert panel.
3. Retain a short standardized orientation for first-time judges, addressing the mild residual on the SUS learning item, and package it as a one-page quick-start guide.
4. Conduct load testing beyond five judges and 40 performers before using the application at larger regional competitions, since the tested limits reflect PSHS-CRC scale rather than a technical ceiling.
5. Extend the application to iOS and pilot the optional cloud sync at a multi-venue event to confirm cross-device aggregation under real network conditions.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 07 July 2026
DOI: 10.47191/ijcsrr/V9-i7-18, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

6. In a separate study, evaluate the psychometric validity of the rubric criteria themselves, which this study deliberately held constant.

REFERENCES

1. Bermoy, L., & Orbegoso, J. (2026). Development and evaluation of PARBIS: A parent-guided augmented reality behavioral intervention system for children with ADHD. International Journal of Research and Innovation in Social Science, 9(12), 2040–2061. https://dx.doi.org/10.47772/IJRISS.2025.91200155
2. Branch, R. M. (2009). Instructional design: The ADDIE approach. Springer. https://doi.org/10.1007/978-0-387-09506-6
3. Department of Education. (2013). K to 12 physical education curriculum guide. Bureau of Curriculum Development. https://www.deped.gov.ph/wp-content/uploads/2019/01/PE-CG.pdf
4. Kundu, K., Yadav, S., & Sayyad, T. (2018). A design and development of rubrics system for Android applications. American Journal of Engineering Research, 7(3), 279–287.
5. Panadero, E., Fernández Ortube, A., Krebs, R., & Roelle, J. (2024). Analysis of online rubric platforms: Advancing toward eRubrics. Assessment & Evaluation in Higher Education, 50(1), 1–19. https://doi.org/10.1080/02602938.2024.2345657
6. Widyandana, D., Utomo, P. S., Setiawan, I. P., Kurniawati, Y. T., & Dandekar, S. (2024). Comparing paper-based and mobile application for rank-based peer assessment in interprofessional education: Before, during, and after the COVID-19 pandemic. BMC Medical Education, 24, 1383. https://doi.org/10.1186/s12909-024-06382-2
7. Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. Journal of Usability Studies, 4(3), 114–123.
8. Wiggins, G. (1998). Educative assessment: Designing assessments to inform and improve student performance. Jossey-Bass.
9. Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. Research in Nursing & Health, 29(5), 489–497. https://doi.org/10.1002/nur.20147
10. 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

Cite this Article: Bermoy, L.R., Fulay, L.P. (2026). Development and Validation of a Rubric-Based Mobile Scoring Application for Folk Dance Performance Assessment. International Journal of Current Science Research and Review, 9(7), pp. 3772-3787. DOI: https://doi.org/10.47191/ijcsrr/V9-i7-18