
4. CHAPTER 4: RESULTS AND ANALYSIS

4.1. Experimental Design

This chapter presents the evaluation of the implemented hip-hop dance assessment system. A total of 45 video recordings were collected across five student participants and three foundational hip-hop movements. Each student performed three attempts per movement, yielding 9 recordings per student and 15 recordings per movement category.

The evaluation setup was as follows:

- Total dataset: 45 videos (5 students $\times$ 3 movements $\times$ 3 attempts).
- Participants: Abby and Dhita (Basic level); Gale (Intermediate level); Kezia and Nicole (Advanced level).
- Movements evaluated: Bouncing, Stepping, and Sliding.
- Expert evaluator: Coach Ambrosius Robby, a certified professional dance instructor with 20 years of teaching experience in hip-hop dance.
- Scoring protocol: Each video was independently scored by Coach Ambrosius Robby on a holistic 0–100 scale, reflecting the evaluator’s expert judgment of overall movement quality for that attempt.
- Recording environment: Indoor studio setting, single stationary smartphone camera, consistent lighting.

The selection of Coach Ambrosius Robby as the sole expert evaluator was deliberate. His extensive professional background ensures that the ground-truth scores reflect genuine domain expertise rather than lay judgment. Establishing a credible expert benchmark is a prerequisite for addressing Hypothesis H3, which proposes that system-generated scores will achieve a Pearson correlation of $r \geq 0.70$ with expert ratings.

4.2. Expert Evaluation Results

Table 3, Table 4, and Table 5 present the expert scores assigned by Coach Ambrosius Robby for each of the 45 recordings, organized by movement. Scores are reported per student per attempt, alongside per-student averages and overall

movement-level descriptive statistics.

Table 4-1 Expert Scores - Bouncing

Student	Level	Attempt 1	Attempt 2	Attempt 3	Average
Abby	Basic	45	45	50	46.7
Dhita	Basic	40	49	50	46.3
Gale	Intermediate	55	60	70	61.7
Kezia	Advanced	30	40	55	41.7
Nicole	Advanced	30	65	65	53.3
Overall (range 30–70, mean 49.9, SD 11.7)					

Table 4-2 Expert Scores - Stepping

Student	Level	Attempt 1	Attempt 2	Attempt 3	Average
Abby	Basic	50	70	75	65.0
Dhita	Basic	50	65	80	65.0
Gale	Intermediate	50	50	70	56.7
Kezia	Advanced	50	50	50	50.0
Nicole	Advanced	50	55	55	53.3
Overall (range 50–80, mean 58.0, SD 10.5)					

Table 4-3 Expert Scores - Sliding

Student	Level	Attempt 1	Attempt 2	Attempt 3	Average
Abby	Basic	30	30	80	46.7
Dhita	Basic	20	20	20	20.0
Gale	Intermediate	40	50	60	50.0
Kezia	Advanced	20	20	20	20.0
Nicole	Advanced	70	75	77	74.0

Overall (range 20–80, mean 42.1, SD 23.3)					
---	--	--	--	--	--

4.3. Notable Observations from Expert Data

Analysis of the expert evaluation data reveals several pedagogically significant patterns. Five observations are discussed below.

- a) Stepping Baseline Pattern. All five students received a score of exactly 50 on Stepping Attempt 1, irrespective of skill level. This uniformity is unlikely to reflect identical performance quality across participants spanning Basic, Intermediate, and Advanced levels. It is interpreted here as an evaluator baseline behavior: the coach appears to have assigned a neutral midpoint score at the outset of each student’s Stepping sequence before calibrating subsequent scores to observed improvement or decline. This anchoring strategy is consistent with established practices in holistic rubric scoring (Robby, 2025), in which evaluators withhold extreme ratings until comparative evidence accumulates.
- b) Within-Student Improvement Across Attempts. The majority of students demonstrated score improvement from Attempt 1 to Attempt 3, suggesting measurable skill acquisition even within a single practice session. The most pronounced gains were observed in Dhita’s Stepping performance (+30 points, from 50 to 80), Kezia’s Bouncing (+25 points, from 30 to 55), and Nicole’s Bouncing (+35 points, from 30 to 65). These trajectories are consistent with the motor learning literature on the power-law of practice, which predicts rapid early improvement followed by a plateau as automaticity develops (Chen & Li, 2023).
- c) Persistent Low Performance in Sliding. Both Dhita and Kezia received flat scores of 20 across all three Sliding attempts, with no evidence of improvement. This pattern distinguishes Sliding as qualitatively different from the other two movements: whereas Bouncing and Stepping exhibited visible intra-session learning gains for most participants, Sliding appears to present a technical barrier that a single practice session cannot overcome. This finding suggests that Sliding may require dedicated, scaffolded training before improvement becomes

- detectable, and implies that automated feedback for Sliding should adopt a different pedagogical register than feedback designed for Bouncing or Stepping.
- d) Highest Score Variance in Sliding ($SD = 23.3$). The Sliding movement exhibited the widest score distribution of the three movements, spanning the full range from 20 to 80 with a standard deviation of 23.3, substantially higher than Bouncing ($SD = 11.7$) and Stepping ($SD = 10.5$). This high discriminative spread suggests that Sliding is the most sensitive movement for distinguishing skill levels: a well-executed slide is clearly distinguishable from a poor one, even to a holistic observer. This property makes Sliding a potentially high-value component in a summative assessment battery, despite (or perhaps because of) its difficulty.
- e) Outlier Observation: Abby's Sliding Attempt 3. Abby's Sliding scores of 30, 30, and 80 exhibit an atypical 50-point jump between Attempts 2 and 3. While such a gain could represent a genuine "breakthrough moment" in motor learning (a phenomenon documented in skill acquisition research as a sudden transition from effortful to fluent execution), it could equally reflect a qualitatively different movement execution in Attempt 3 (for example, a wider lateral range or altered foot positioning) that happened to align more closely with the reference performance. Further investigation, including frame-level inspection of the attempt, is warranted before drawing interpretive conclusions.

4.3.1. Interpretation of the Score 20 Across Different Skill Levels

The expert data shows that a score of 20 appeared for both a Basic participant and an Advanced participant in the Sliding movement. Specifically, Dhita, classified as Basic, received 20 for all three Sliding attempts, while Kezia, classified as Advanced, also received 20 for all three Sliding attempts. This result should not be interpreted as evidence that both participants had the same overall dance ability or that the skill labels were incorrect.

The skill labels describe the participants' broader background and general experience level. In contrast, the expert scores represent the quality of a specific movement attempt under the criteria used for that movement. A participant who is generally more advanced may still receive a low score on a particular basic movement if the performance does not demonstrate the expected glide, timing, weight transfer,

hip stability, or movement control. Likewise, a beginner may perform a particular movement more successfully than a participant with a higher overall skill label.

This pattern reinforces that the skill-level analysis in this study is descriptive only. The primary evaluation does not expect system scores to increase perfectly across skill labels, because each score represents the quality of a specific movement attempt rather than a participant's overall dance ability. Instead, the main evaluation compares individual video-level system scores with individual expert scores. The occurrence of the same low Sliding score across different skill levels also shows why movement-specific evaluation is necessary and why a dancer's overall level should not be treated as a substitute for the expert evaluation of each attempt.

4.4. System Output Interface

This section presents representative screenshots of the implemented Dance Grading System. The screenshots illustrate the complete user workflow, beginning from the dashboard, movement selection, reference viewing, student video upload, system processing, and final feedback output. These interfaces demonstrate how the technical scoring pipeline is presented to users as an educational self-assessment tool.

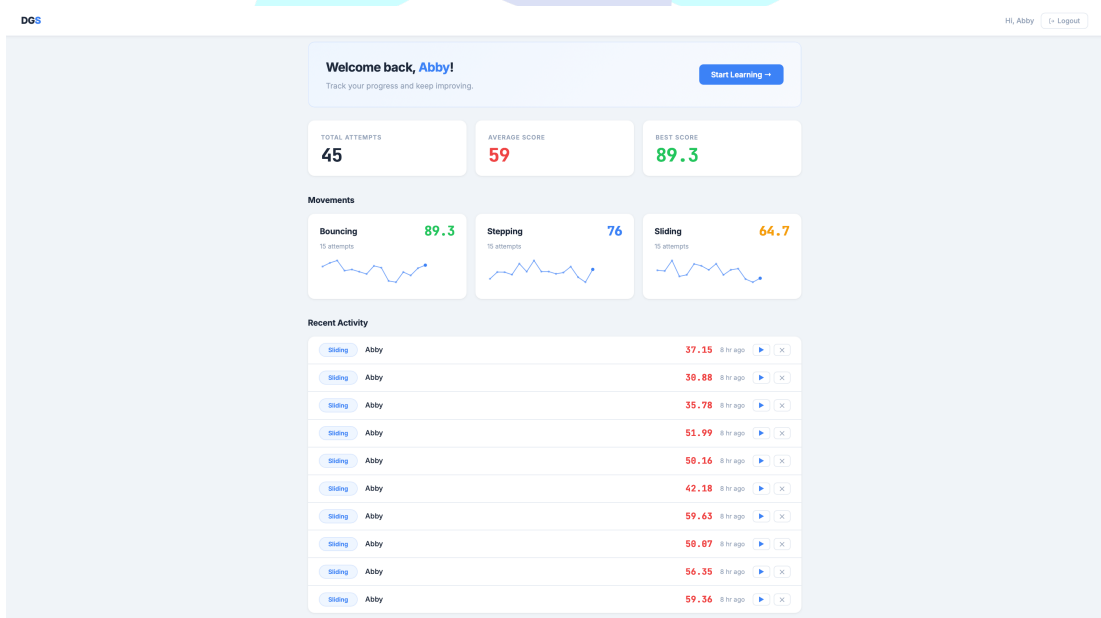


Figure 4-1 Dashboard Interface

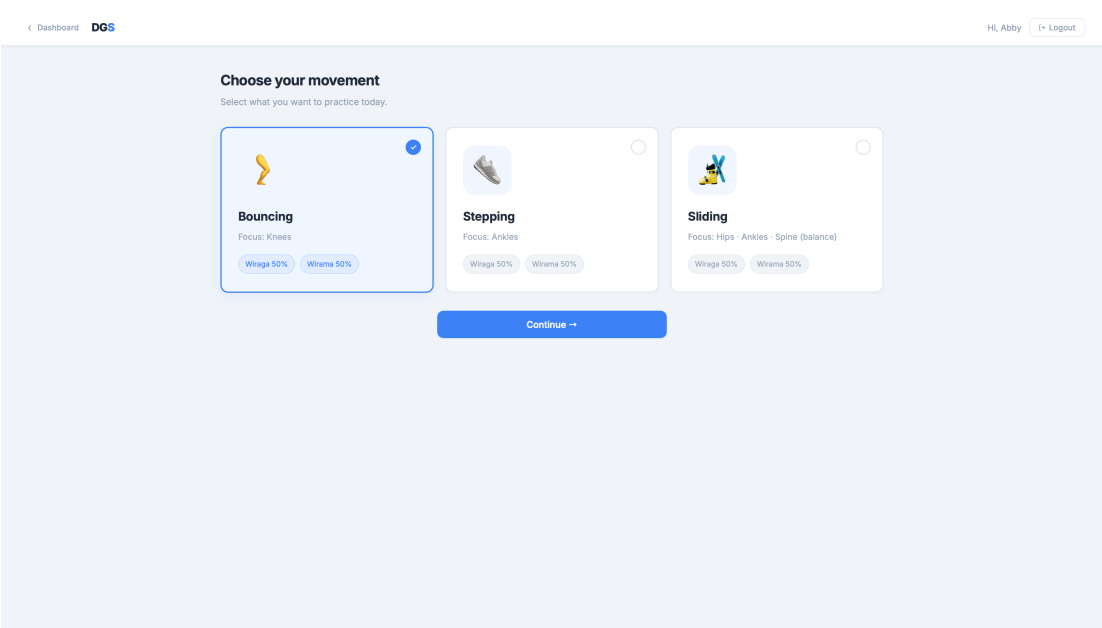


Figure 4-2 Movement Selection Interface

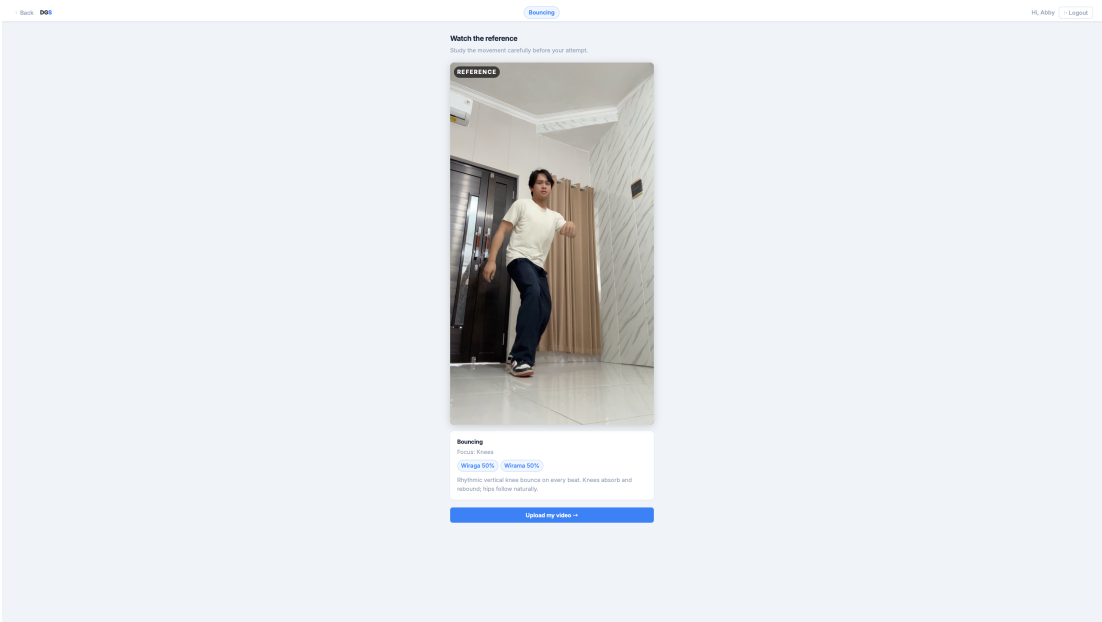


Figure 4-3 Reference Video Interface

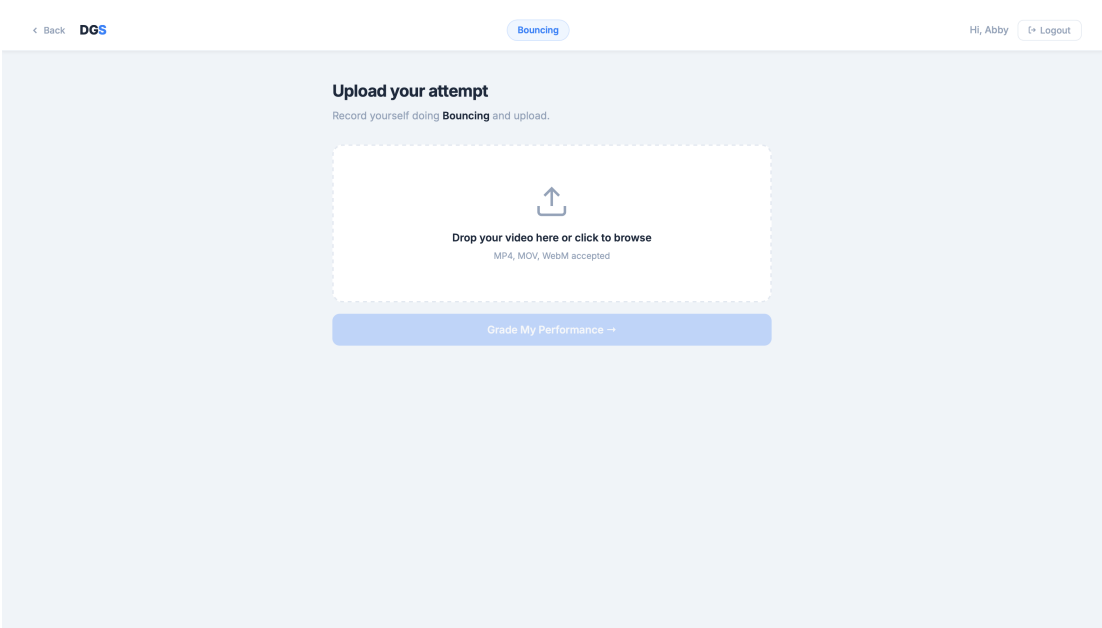


Figure 4-4 Upload Interface

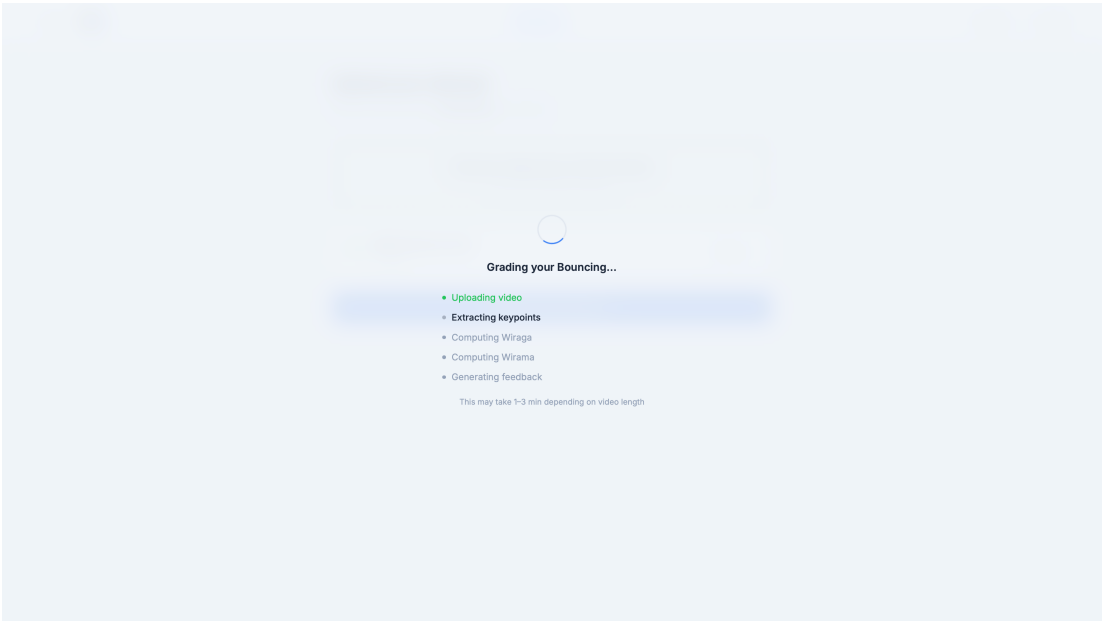


Figure 4-5 Processing/Loading Interface

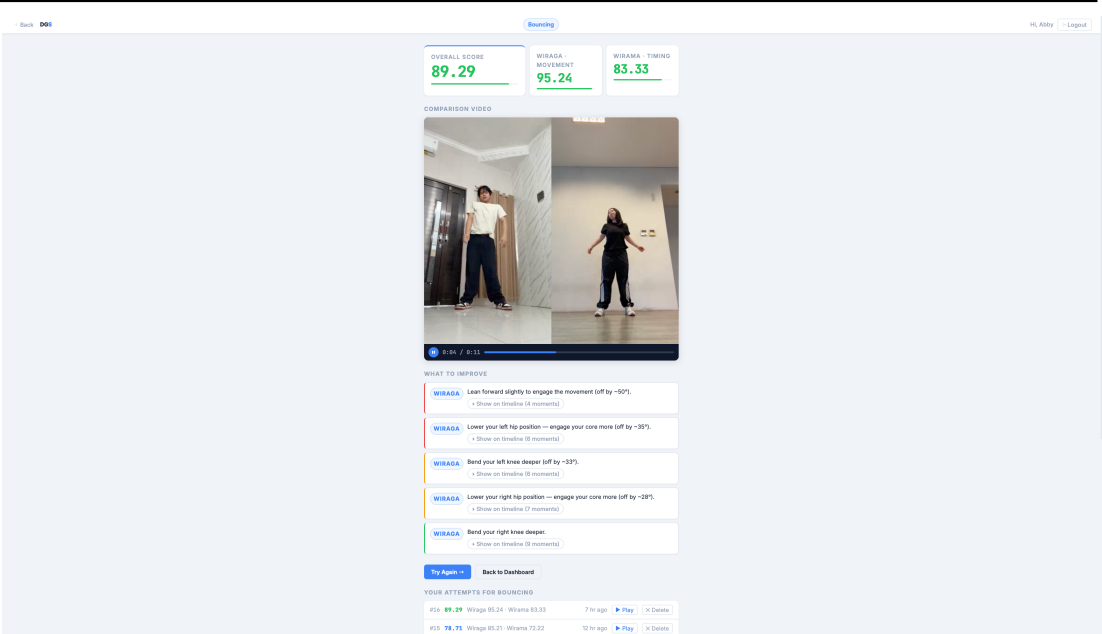


Figure 4-6 Result Interface

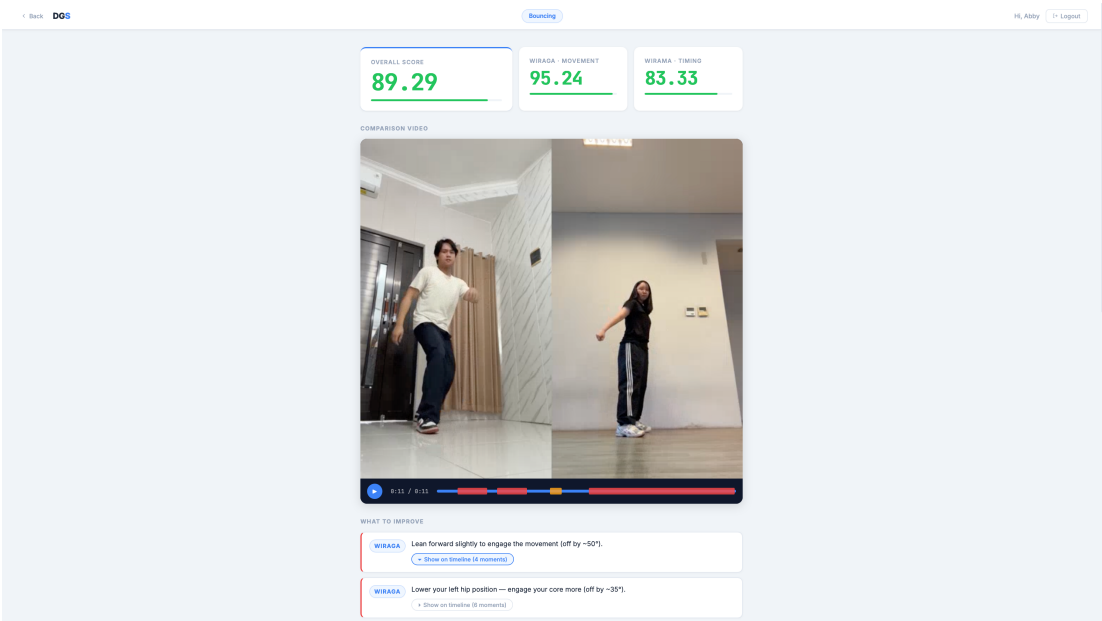


Figure 4-7 Timeline Marker Feedback Interface

4.5. System Evaluation Results

System scores for all 45 videos were computed using the implemented Wiraga + Wirama scoring engine described in Chapter 3. The scoring pipeline processes each student video against the pre-computed reference pose cache for the corresponding movement and outputs a composite overall score on a 0–100 scale, where Wiraga (physical form) and Wirama (timing precision) each contribute 50 percent of the final

result. All 45 videos were processed successfully with no extraction failures, yielding a complete dataset for subsequent correlation analysis.

4.5.1. Overview of System Scores

Table 4.4 presents the system-generated overall scores for all 45 videos organized by student, movement type, and attempt number. Scores are reported to two decimal places as produced by the scoring engine.

Table 4-4 System-Generated Overall Scores by Student, Movement, and Attempt

Student	Level	Movement	Attempt 1	Attempt 2	Attempt 3	Student Avg
Abby	Basic	Bouncing	75.64	83.89	89.29	82.94
Abby	Basic	Stepping	53.17	61.59	61.50	58.75
Abby	Basic	Sliding	49.52	48.39	64.73	54.21
Dhita	Basic	Bouncing	66.06	68.79	63.73	66.19
Dhita	Basic	Stepping	58.36	72.10	62.10	64.19
Dhita	Basic	Sliding	39.87	42.40	59.36	47.21
Gale	Intermediate	Bouncing	58.52	77.08	73.02	69.54
Gale	Intermediate	Stepping	76.02	62.24	62.17	66.81
Gale	Intermediate	Sliding	56.35	50.07	59.63	55.35
Kezia	Advanced	Bouncing	42.59	39.79	62.92	48.43
Kezia	Advanced	Stepping	59.36	61.02	68.36	62.91
Kezia	Advanced	Sliding	42.18	50.16	51.99	48.11

Nicole	Advanced	Bouncing	55.31	71.86	78.71	68.63
Nicole	Advanced	Stepping	55.16	48.96	64.90	56.34
Nicole	Advanced	Sliding	35.78	30.88	37.15	34.60

4.5.2. Descriptive Statistics by Movement Type

Table 4.5 summarizes the descriptive statistics of system scores aggregated by movement type across all five students and three attempts ($n = 15$ per movement).

Table 4-5 System Score Descriptive Statistics by Movement Type

	Min	Max	Mean	Std Dev
Bouncing	39.79	89.29	67.15	13.97
Stepping	48.96	76.02	61.80	6.88
Sliding	30.88	64.73	47.90	9.76
Overall	30.88	89.29	58.95	13.22

Bouncing produced the highest mean system score ($M = 67.15$, $SD = 13.97$) and the widest score dispersion among the three movement types. This indicates that the system produced more varied evaluations for vertical knee oscillation and bounce-related movement patterns. Stepping produced a lower mean score ($M = 61.80$, $SD = 6.88$), with a narrower range of variation compared to Bouncing. Sliding produced the lowest mean system score ($M = 47.90$, $SD = 9.76$), suggesting that the current Sliding scoring component assigned generally lower scores than the other two movement categories. Across all 45 videos, system scores ranged from 30.88 to 89.29, with an overall mean of 58.95 and standard deviation of 13.22.

4.5.3. System Score Trends Across Attempts

Examining score trajectories across the three attempts provides an initial

indication of whether the system is sensitive to within-session performance changes. However, these trends should be interpreted descriptively rather than as formal evidence of learning improvement, because the evaluation was not designed as a longitudinal training study.

For Bouncing, several students showed increasing system scores across attempts. Abby increased from 75.64 to 83.89 to 89.29, while Nicole increased from 55.31 to 71.86 to 78.71. Gale also showed an increase from 58.52 to 77.08 before slightly declining to 73.02. These trends suggest that the system was responsive to some changes in vertical movement execution across attempts.

For Stepping, the trajectories were less consistent. Dhita increased from 58.36 to 72.10 before declining to 62.10, while Kezia increased from 59.36 to 61.02 to 68.36. However, Gale declined from 76.02 to 62.24 to 62.17, and Nicole fluctuated from 55.16 to 48.96 to 64.90. This suggests that the Stepping scoring component may be sensitive to attempt-level variation but does not consistently reflect improvement patterns across all participants.

For Sliding, score trajectories also varied across participants. Abby increased notably from 49.52 and 48.39 to 64.73 on the third attempt, while Dhita increased from 39.87 to 42.40 to 59.36. However, Nicole remained low across all three attempts, with scores of 35.78, 30.88, and 37.15. This indicates that the Sliding component was able to produce different score ranges across students, but its alignment with expert interpretation requires further examination through correlation analysis.

4.5.4. System Scores by Skill Level

Table 4.6 presents mean system scores grouped by the skill level assigned to each student by the expert evaluator. This analysis examines whether the system scores directionally reflect the learner-level distinctions recognized by the expert. However, the Intermediate group consists of only one participant, so this section should be interpreted as descriptive context rather than inferential evidence.

Table 4-6 Mean System Scores by Skill Level

	Students	n	Mean	Std Dev
Basic	Abby, Dhita	18	62.25	13.06
Intermediate	Gale	9	63.90	9.39
Advanced	Kezia, Nicole	18	53.17	13.44

The system assigned similar mean scores to Basic-level students ($M = 62.25$, $SD = 13.06$) and the Intermediate student ($M = 63.90$, $SD = 9.39$), while Advanced-level students received a lower mean score ($M = 53.17$, $SD = 13.44$). This pattern does not align with the expected skill progression in which Advanced participants would generally receive higher scores than Basic or Intermediate participants.

The absence of a clear skill-level gradient suggests that the current scoring thresholds and feature weighting are not calibrated to differentiate learner levels in a manner consistent with expert classification. Because the Intermediate category contains only one student, no general conclusion can be drawn about intermediate-level performance. The skill-level analysis is therefore used only as descriptive context, while the main evaluation remains the correlation between system-generated scores and expert scores across all 45 videos.

4.6. Correlation Analysis Between System and Expert Scores

4.6.1. Overall Correlation

The overall Pearson correlation between system-generated scores and expert scores across all 45 videos yielded $r = 0.25$ ($p = 0.092$, $n = 45$), with a Mean Absolute Error (MAE) of 17.27 points and a Root Mean Square Error (RMSE) of 20.95 points. This result does not meet the threshold specified in Hypothesis H3, which required a minimum correlation of $r \geq 0.70$. Accordingly, H3 is not supported by the evaluation data. The weak and statistically non-significant correlation indicates that the current scoring model does not strongly align with the ranking of student performances

produced by the expert coach.

A systematic upward bias of +8.93 points was identified across all movement types, meaning that the system generally assigned higher scores than the expert evaluator. This suggests that the scoring thresholds embedded in the current algorithm are more lenient than the standards applied by the expert coach. Therefore, the disagreement between system and expert scores may be partly explained by calibration differences rather than random scoring error alone.

4.6.2. Per-Movement Analysis

Table 4.7 presents the per-movement breakdown of Pearson correlation coefficient, statistical significance, mean absolute error, and systematic bias.

Table 4-7 Per-Movement Correlation Analysis Between System and Expert Scores (Pearson r , MAE, and Bias)

Movement	r (Pearson)	p-value	MAE (points)	Bias (points)	RMSE (points)
Bouncing	0.54	0.037	17.24	+17.21	21.08
Stepping	0.11	0.707	10.96	+3.80	12.38
Sliding	-0.13	0.651	23.61	+5.76	26.83
Overall	0.25	0.092	17.27	+8.93	20.95

Among the three movements, Bouncing produced the strongest system-expert alignment, with a moderate positive correlation ($r = 0.54$, $p = 0.037$). This was the only movement-level correlation that reached statistical significance. However, the Bouncing scores also showed a high positive bias (+17.21 points), indicating that although the system ranking was more aligned with the expert for this movement, the system still tended to assign higher absolute scores than the expert evaluator.

Stepping produced a weak positive correlation ($r = 0.11$, $p = 0.707$), but it had the lowest MAE among the three movements (10.96 points). This indicates that the system's absolute score differences were smaller for Stepping, but the ranking of student performances did not strongly align with the expert's ranking. Therefore, the

Stepping component may require refinement in how foot-strike timing, weight transfer, and movement-form criteria are weighted.

Sliding produced a weak negative correlation ($r = -0.13$, $p = 0.651$) and the highest MAE (23.61 points), indicating that the current Sliding scoring component did not align well with expert judgment. This result suggests that the system's current pose-based features are insufficient to capture the qualities emphasized by the expert in Sliding performance, such as smooth weight transfer, horizontal control, and hip stability. Sliding therefore represents the main limitation of the current scoring model and should be prioritized for future refinement.

4.6.3. Sliding Scoring Limitation

Diagnostic analysis of the per-movement results indicates that Sliding remains the weakest component of the current scoring model. Unlike Bouncing, which produced a moderate and statistically significant correlation with expert scores, Sliding produced a weak negative correlation ($r = -0.13$, $p = 0.651$) and the highest Mean Absolute Error among the three movements (MAE = 23.61 points). This indicates that the current Sliding scoring module does not reliably rank student performances in the same direction as the expert evaluator.

The limitation is likely caused by the nature of the Sliding movement itself. Sliding quality is not determined only by ankle separation or horizontal displacement. Expert evaluation also considers glide smoothness, weight transfer, hip stability, body control, and the illusion of continuous lateral motion. These qualities are difficult to capture using the current pose-based feature representation, which primarily relies on angular and distance-based comparisons derived from two-dimensional landmarks.

The largest mismatches in the latest dataset further support this limitation. Several Sliding videos produced substantial differences between system and expert scores, including Nicole_slide_2 ($\Delta = -44.12$), Nicole_slide_3 ($\Delta = -39.85$), Dhita_slide_3 ($\Delta = +39.36$), Nicole_slide_1 ($\Delta = -34.22$), and Kezia_slide_3 ($\Delta = +31.99$). These mismatches show that the system does not consistently over-score or under-score Sliding; instead, it produces unstable estimates depending on the participant and attempt. This suggests that the Sliding component requires more than

simple threshold adjustment and may need a revised feature design.

Future improvements to the Sliding module should therefore include direct measurement of lateral displacement, trajectory smoothness, ankle path linearity, hip stability, and center-of-mass consistency. A multi-reference approach may also improve robustness by representing several acceptable variations of Sliding rather than relying on a single reference pattern. Based on these results, Sliding should be treated as the primary movement-specific limitation of the current system.

4.6.4. Notable Observations

Several observations from the per-video analysis provide further insight into the system's current behavior. First, Bouncing produced the strongest alignment with expert evaluation among the three movements, with $r = 0.54$ ($p = 0.037$). This indicates that the system was able to capture some meaningful aspects of Bouncing performance, particularly vertical knee oscillation and repeated rhythmic movement. However, Bouncing also showed a high positive bias of +17.21 points, meaning that the system generally assigned higher scores than the expert evaluator. Therefore, the Bouncing module shows promising ranking behavior but still requires score calibration.

Second, Stepping produced the lowest MAE among the three movements (MAE = 10.96 points), but its correlation with expert scores was weak and statistically non-significant ($r = 0.11$, $p = 0.707$). This means that the absolute score differences were smaller for Stepping, but the system did not strongly reproduce the expert's ranking of student performances. In practical terms, the Stepping module may be producing scores within a closer numerical range while still failing to distinguish higher-quality and lower-quality performances in the same way as the expert coach.

Third, several videos achieved close agreement between system and expert scores. Examples include Gale_slide_2 ($\Delta = +0.07$), Kezia_bounce_2 ($\Delta = -0.21$), and Gale_slide_3 ($\Delta = -0.37$). These cases suggest that the system can produce accurate estimates under favorable conditions. However, the overall correlation remains weak because other videos show large mismatches, particularly in Sliding and some Bouncing attempts.

Fourth, the overall bias of +8.93 points shows that the system generally scores more leniently than the expert evaluator. This indicates that part of the disagreement may be caused by calibration differences in the scoring thresholds. However, the movement-specific results show that calibration alone is not sufficient to solve all issues. Bouncing mainly requires score recalibration, Stepping requires better ranking sensitivity, and Sliding requires more substantial redesign of its movement-specific feature representation.

4.7. Usability Evaluation (SUS)

The System Usability Scale (SUS) evaluation was administered to assess the perceived usability of the Dance Grading System. The SUS questionnaire consists of 10 standardized items; each rated on a five-point Likert scale. For odd-numbered items, the adjusted score is calculated by subtracting 1 from the user response. For even-numbered items, the adjusted score is calculated by subtracting the response from 5. The sum of the adjusted scores is then multiplied by 2.5, producing a final SUS score ranging from 0 to 100. In general interpretation, a SUS score of 68 is commonly treated as an average usability benchmark. Table 4.8 presents the SUS scores obtained from the five participants.

Table 4-8 SUS Scores by Participant

Participant	SUS Score
Gale	67.50
Dhita	70.00
Nicole	72.50
Kezia	50.00
Abby	85.00

The mean SUS score across all five participants was 69.00, with scores ranging from 50.00 to 85.00. This result places the system slightly above the commonly used average benchmark of 68, suggesting that the overall usability of the system was acceptable. However, the variation between participants indicates that user experience was not fully consistent across all users.

Abby gave the highest SUS score (85.00), indicating a strong usability perception and suggesting that the system interface was easy to understand and operate for this participant. Nicole and Dhita also gave above-average scores of 72.50 and 70.00 respectively. Gale's score of 67.50 was slightly below the benchmark but still close to average usability. Kezia gave the lowest score of 50.00, indicating that at least one participant experienced usability difficulties or perceived the system as less intuitive.

The open-ended responses also support this interpretation. Most participants did not report major usability issues, with responses such as "All good" and "No." However, one participant suggested adding different dance styles to make the system more widely applicable, while another noted that improving feedback detail would be beneficial. These comments indicate that the system was usable in its current form, but future development should improve feedback specificity and expand the supported movement vocabulary.

Overall, the SUS result suggests that the Dance Grading System achieved acceptable usability for a prototype educational self-assessment tool. Nevertheless, the relatively small number of respondents and the variation in individual scores indicate that the usability findings should be interpreted cautiously. A larger usability study with more diverse participants is recommended for future evaluation.