RaterSync: Inter-Rater Reliability & Calibration Suite

Documentation & User Guide

Release Date: November 2025

Platform: Universal (HTML5/JavaScript) - Client-Side Only

1. Executive Summary

RaterSync is a specialized psychometric tool designed for academic health professions (Pharmacy, Medicine, Nursing). It facilitates **Inter-Rater Reliability (IRR)** calibration workshops, allowing faculties to harmonize their grading standards for OSCEs (Objective Structured Clinical Exams) and other subjective assessments.

Unlike complex statistical software (SPSS/SAS) or static spreadsheets, RaterSync provides instant, visual feedback on rater bias ("Hawk" vs. "Dove" effects) and generates certification-ready reports entirely within the browser, ensuring zero data privacy risks.

2. Key Differentiators & Features

Feature	Description	Benefit
Dual Calibration Modes	**Peer-to-Peer:** Compares raters against the group average. **Gold Standard:** Compares raters against a designated Expert score.	Flexible for both group consensus workshops and expert-led training sessions.
Dual Scoring Patterns	**Continuous:** For rubric-based scoring (e.g., 0-10, 0-100). **Pass/Fail:** For checklist-based scoring.	Adaptable to any exam format.
Hawk/Dove Analysis	Visualizes which rater is too strict (Hawk) or too lenient (Dove) relative to the standard.	Provides actionable feedback to faculty to adjust their grading behavior.
Smart Reports	Generates two types of PDF exports: 1. **Analysis Report:** Full data breakdown. 2. **Certificate:** A formal certificate if Reliability > 0.7.	Ready for accreditation files and faculty development portfolios.
Privacy-First	**Client-Side Architecture:** All calculations happen on the user's device. No student data is sent to the cloud.	GDPR/FERPA compliant by design.

3. Methodology & Benchmarks

RaterSync utilizes gold-standard psychometric algorithms used in high-stakes testing.

A. For Continuous Data (Rubrics)

- **Test Used: Intraclass Correlation Coefficient (ICC)**.
- **Specific Model:** ICC(2,k) — Two-way random effects, absolute agreement, average measures.
- **Why:** This model accounts for both the correlation between raters and the systematic difference in their mean scores (bias).
- **Reference:** *Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86(2), 420.*

B. For Categorical Data (Pass/Fail)

- **Test Used: Fleiss' Kappa**.
- **Why:** Unlike simple percent agreement, Kappa calculates agreement correcting for chance (random guessing).
- **Reference:** *Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378–382.*

C. Bias Detection (Hawk/Dove)

- **Method: Z-Score Deviation**.
- **Logic:** Calculates the deviation of a specific rater's mean score from the "Reference Mean" (either the Group Mean or the Expert Mean).
- **Visualization:**
    - **Red Bars (Negative):** Indicates "Hawk" bias (Stricter than standard).
    - **Green Bars (Positive):** Indicates "Dove" bias (Lenient than standard).

4. User Guide

Section 1: Exam Metadata

- Located at the top of the interface.
- **Function:** Fields entered here (Course Code, Exam Name, Date) are purely cosmetic—they are auto-populated into the header of the **Printed Report** and the **Certificate**.

Section 2: Calibration Setup

This panel configures the logic engine before generating the grid.

1. **Calibration Mode:**
    - *Peer Comparison:* Use this when there is no "Expert." The app assumes the "Average of the Group" is the truth.
    - *Gold Standard:* Use this when a Course Director has pre-scored the student. The app compares everyone to this Expert score.
2. **Scoring Pattern:**
    - *Continuous:* Unlocks the "Max Possible Score" field.
    - *Pass / Fail:* Locks inputs to binary options.
3. **Max Possible Score:** (Continuous mode only). Sets the upper limit validation for the grid (e.g., 10, 20, 100).
4. **Students/Raters:** Define the matrix size (Rows x Columns).
5. **Generate Grid Button:** Locks these settings and builds the Data Entry Matrix.

Section 3: Data Entry Matrix

1. **Rater Headers:** The top row (e.g., "Rater 1") is **editable**. Click on the text to rename them (e.g., "Dr. Ibrahim").
2. **Expert Score Column:** Only visible in "Gold Standard" mode. This is the reference value.
3. **Input Cells:**
    - *Continuous:* Accepts numbers. Alerts if you exceed the Max Score.
    - *Pass/Fail:* Dropdown menu (Pass=1, Fail=0).
4. **Run Analysis Button:** Triggers the calculation engine.

Section 4: Analysis Dashboard

Once analysis is run, this right-hand panel populates:

1. **Reliability Coefficient:** The main score (0.0 to 1.0).
2. **Status Badge:**
    - **CALIBRATED (Green):** Score ≥ 0.70.
    - **NOT CALIBRATED (Red):** Score < 0.70.

3. **Chart:** Visualizes the specific bias of each rater.
4. **Print Buttons:**
    - *Print Analysis Report:* Opens print dialog formatted for a detailed data report.
    - *Print Certificate:* **Hidden** unless the Status is "CALIBRATED." Prints a ceremonial document.

5. Technical Specifications & Requirements

- **File Type:** Single HTML File (.html).
- **Dependencies:**
    - **Chart.js (v3.x):** Loaded via CDN (Internet connection required for first load, or can be downloaded for offline use).
    - **FontAwesome:** For UI icons.
- **Browser Support:** Google Chrome, Microsoft Edge, Safari, Firefox (Latest Versions).
- **Printing:** Optimized for A4 / Letter size paper using CSS @media print queries.
- **Data Persistence:** None. Refreshing the page clears all data (Security Feature).
- **Concept & Logic:** Dr. Muhammad AlShorbagy, Dean, College of Pharmacy, GMU.
- **Technical Implementation:** AI-Assisted Development (Code generation).
- **Methodology:** "This single-file HTML application demonstrates a 'No-Code/Low-Code' development approach. The domain expertise, algorithm logic, and user experience design were provided by Dr. Muhammad AlShorbagy, while the source code was generated via prompt engineering using Large Language Models (LLMs)."

6. Troubleshooting

- **"My Reliability Score is 0.00 but we all agreed perfectly!"**
    - *Cause:* If everyone gives the *exact same score* to *all students* (e.g., Student 1 gets an 8, Student 2 gets an 8), the variance is zero.
    - *Fix:* v4.0 has a patch for this. If variance is near-zero and error is near-zero, it defaults to 1.0 (Perfect Agreement).
- **"I can't see the Certificate button."**
    - *Reason:* The certificate is locked to maintain academic standards. It only appears if your reliability score is **0.70 or higher**.
- **"The Print preview is blank."**
    - *Fix:* Ensure "Background Graphics" is checked in your browser's Print Dialog settings to see the colors and watermarks.