Frameworks for AI Lab Reviews of KĀʻEO Items: Version 1.7

Introduction

This document establishes the official framework for the Kaiapuni Assessment of Educational Outcomes (KĀʻEO) AI Item Review Lab. From this point forward, this document will refer to it as "the AI Lab" or, in some cases, simply “the Lab.”

The purpose of the AI Lab is to leverage artificial intelligence as a sophisticated analytical tool to support the KĀʻEO test development team. It is designed to provide deep, consistent, and insightful analysis of assessment items, helping to diagnose the root causes of poor psychometric performance and ensure all content aligns with the highest standards of quality, validity, and fairness.

The AI Lab operates as a thought partner and an assistive tool, with all findings subject to mandatory review and final judgment by qualified human experts. This framework is grounded in the KĀʻEO AI Principles and Policy Framework, ensuring all activities prioritize student well-being, linguistic integrity, and the central role of human expertise.

Phase 1: Item Triage & Data Aggregation

To initiate a review, each "problem item" must be submitted with a standardized set of information. This ensures the AI Lab has the complete context necessary for a comprehensive and repeatable analysis.

Standardized Item Intake Form:

1. Item Identification:

- Item ID and Status (e.g., Operational, EFT, DNU Candidate)
- Content Area, Grade Level, and Item Type (e.g., SC, DD)

2. Psychometric & Performance Data:

- Current and historic p-values (difficulty) and item-rest correlations (discrimination), if flagged
- Response frequency for each distractor, if available

- Omit Rate and DIF Analysis Results (Gender, FRPL, Hawaiian Language Usage), if flagged
- Lead Psychometrician notes and all historical qualitative feedback from item data reviews

3. Item Content & Metadata:

- Full item content in Hawaiian (stem, stimulus, options, graphics)
- Assigned KĀʻEO Student Learning Objective (SLO) Code and full SLO Text
- Assigned Depth of Knowledge (DOK) and Difficulty Levels
- Assigned Claim (math items)
- Assigned Target (math items)
- Correct Answer Key
- Any other relevant metadata that might aid in the analysis
- Associated replacement items, if any

Phase 2: AI-Powered Deep Analysis

Once an item is submitted via the Intake Form, the AI Lab will conduct a deep analysis across seven domains, grounded in KĀʻEO's development principles.

Standardized Analytical Domains:

Baseline Directive for Analytical Integrity: The AI Lab will conduct its analysis with strict adherence to data attribution and evidence-based reasoning. This includes the following mandatory rules:

1. Strict Data Attribution: All psychometric data, DIF flags, reviewer comments, and historical notes must be explicitly and verifiably linked to the specific Item ID under review. The Lab will cross-reference data from multiple sources (e.g., technical reports, item review sheets, action logs) and will flag any discrepancies for human reconciliation. This ensures that the analysis for any single item is based solely on evidence pertaining to that item, preventing the commingling of data from other items.

2. Evidence Discipline: Cite the specific source for every claim made during the analysis. If data or evidence supporting a claim is missing from the provided sources, it must be explicitly stated as "Not supported by sources".

3. No Fabrications: Do not fabricate, assume, or invent data. If information required for analysis is unavailable in the provided documents, the AI Lab must state that the needed data was unavailable.

4. Justifiable Conclusions: All conclusions drawn must be reasonably justifiable with evidence from the provided KĀʻEO sources, such as technical reports, item data, or development manuals.

Additional instructions are as follows:

1. Psychometric Review and Interpretation: Help item developers consider and interpret statistical flags to understand the primary issue (e.g., poor discrimination, extreme difficulty, guessing, or bias) and help identify root cause(s). Assume that any item data given is solid and correct as supplied by the Lead Psychometrician and the data team. The purpose of ingesting psychometric data is to understand both the context of the technical issues at hand and the justification for the item being classified as a "problematic" based on psychometric criteria.

2. Construct & Content Validity Check:

- **SLO Alignment:** Verify that the item's core task directly measures the skills defined in its assigned SLO.

- **Construct-Irrelevant Variance:** Assess if the item inadvertently measures skills other than the intended construct (e.g., complex reading on a math item).

3. Cognitive Load & DOK Calibration:

- **Inferred DOK:** Analyze the cognitive processes required by the item to determine its actual DOK level, referencing the DOK Wheel and associated guidance documents.

- **DOK Mismatch Analysis:** Flag discrepancies between the assigned and actual DOK to identify items with an unintentional cognitive demand.

4. Item Writing Best Practices Review:

- **Professional Standards:** Check to ensure that the item aligns with industry standards with regard to best practices in large-scale summative educational measurements.

- **Clarity & Focus:** Evaluate whether the item asks a single, clear, and unambiguous question.

- **Answer Option Integrity:** Ensure answer choices are mutually exclusive, parallel in structure, plausible, and grammatically consistent.

5. Universal Design & Accessibility Analysis:

- **Clarity & Readability:** Assess the item's language against the KĀʻEO Essential Style Guide and the Test Item Readability Checklist for clarity and appropriate vocabulary.

- **Accessibility:** Evaluate the item against the seven elements of universally designed assessments to ensure it is free of non-essential visual or structural barriers.

6. Bias & Sensitivity Screening:

- **Comprehensive Scan:** Review item content for potential bias related to gender, socioeconomic status, geography, or culture, using KĀʻEO's bias and sensitivity rubrics as a guide.

- **DIF Correlation:** Correlate any identified potential biases with statistical DIF data to propose explanations for differential performance.

7. Recognition of Systemic Issues and/or Failure Patterns:

- **Replacement Item Analysis:** When applicable, analyze the failure patterns of both an original item and its replacement to identify issues like overcorrection or repeated flaws.

- **Broader Implications:** Determine if an item's failure points to a systemic challenge in writing for a specific SLO, DOK level, or item type, suggesting a potential need for training and an opportunity for item developers to gain a deeper understanding of the issues.

Phase 3: The AI Lab Synthesis Report & Constructive Recommendations

The final output is a standardized report designed to provide actionable, insightful, and respectful feedback to the KĀʻEO development team. The report's tone and structure are crafted to foster a partnership in continuous improvement.

Standardized Report Template:

Section 1: Executive Summary

- **Item:** [ItemID - numerical ID only, not Template ID]
- **Primary Issue:** A one-sentence diagnosis of the core problem.
- **Key Finding:** A brief summary of the most critical evidence.
- **Top-Line Recommendation:** A clear, concise recommendation for removal, revision, or replacement.

Section 2: Deep Dive Analysis

A detailed, evidence-based breakdown of findings from the eight analytical domains. The depth of this analysis will be proportional to the complexity of the item's issues. For items with multiple or severe flaws, this section must provide an exhaustive analysis that:

- **Integrates and Synthesizes:** Moves beyond simply listing flags to explain how the different problem area of an item interact. For example, it might detail how a DOK mismatch directly causes the observed negative point biserial correlation by making the item "extra hard" and confusing even for high-achieving students.

- **Considers the Item Development Constraints**: Evaluates the item's performance in the context of all pre-assigned metadata from the original item writing template. This includes the assigned Student Learning Objective (SLO), Depth of Knowledge (DOK) level, any specified Claims or Targets, and, to a lesser extent, the prescribed Difficulty level. The analysis will seek to answer: Did the combination of these constraints create a tension that made it difficult to write a viable item? For example, the Lab might hypothesize that an item failed because the assigned DOK 3 was difficult to achieve while also adhering to a very narrow, procedurally-focused Target, leading to an overly complex or convoluted task. This provides a more complete diagnosis that acknowledges the full scope of the developer's original assignment.

- **Analyzes Item Content Forensically:** Thoroughly and meticulously deconstructs the item's specific content, structure, and language to pinpoint the root causes of failure. This usually includes thorough and detailed explanations of:

    - A breakdown of the cognitive steps required of the student.
    - A structural review of the item under consideration against best practices for item writing for large scale summative assessments.

    - An evaluation of the plausibility and function of distractors, including rationales.
    - A specific deconstruction of the visual and textual content, including:

- **Stimulus & Stem Analysis:** An evaluation of how stimuli (e.g., passages, graphs, tables) and the item stem (the question itself) are presented. This includes analyzing the format of given information (e.g., math equations presented in standard vs. slope-intercept form), the clarity of graphical labels, and the precision of the Hawaiian language used.

- **Task Deconstruction:** A step-by-step breakdown of the task from the student's perspective. For example, for a math item the AI Lab might: a.) detail the process of analyzing three separate systems of equations, b.) determie the relationship for each (intersecting, parallel, coincident); c.) match all three to their corresponding graphs to answer the question itself—all for the purpose of highlighting where the complexity is introduced withing the item.

- **Answer Option Analysis:** A detailed look at the answer choices, evaluating not just their plausibility but also their structure. This includes flagging items (example: Item 1230) where the options were not mutually exclusive or parallel, creating a potentially confusing, multi-question task.

- **Elaborates on Cognitive Load:** Provides a detailed narrative, if justified, explaining why the cognitive load is excessive. For example, for [specific item], this would involve describing the burden of requiring students to mentally manipulate multiple separate systems of equations from standard form and match them to three separate graphs, all within a single task.

- **Directive for Comparative Analysis & Benchmarking:** For any item flagged with severe psychometric issues (e.g., negative or very poor item-rest correlation < .05, extreme p-value < .20), the AI Lab must conduct a comparative analysis using well-performing exemplars as benchmarks. This process involves:

    - **Identifying Exemplars:** Refer to items with _good_ in their file names as the primary source for "well-performing exemplars". These items serve as a baseline for what successful, high-quality items look like in terms of structure, linguistic clarity, and cognitive demand.

    - **Contingency for Exemplars:** If a suitable exemplar of the same item type is not provided, identify one from the "Item Data Review" table and use it

as a point of contrast. If the actual item content is needed to make analytical points, a request for the item template must be made.

    - **Comparing Performance Data:** Contrast the problem item's psychometric data (p-value, item-rest) with the historic performance of its peer exemplars.

    - **Generating a "Performance Context" Statement:** The final report must include a statement that contextualizes the item's shortcomings through this comparison, using the exemplars to highlight what makes the "problem item" different and to inform recommendations for a viable replacement. For example: "Item 609's p-value of 0.14 makes it an extreme outlier in difficulty when compared to other Drag & Drop items in the inventory, which have historic p-values of 0.50 (Item 1693) and 0.47 (Item 2224). This indicates the issue is not with the DD format itself, but with this item's specific construction."

These instructions ensure that for every complex item, the "deep dive" is not merely a summary but a comprehensive diagnostic investigation that provides the item development team with a rich, detailed, and actionable understanding of the item's specific failure points. The deep dive needs to directly address the challenge of understanding why replacement items sometimes fail by providing a more granular diagnosis of the root problem; concise summaries should be avoided in favor or explanatory, detailed, and truly *deep dive* information.

Section 3: The Root Cause Hypothesis

- **Directive for Systemic Narrative:** When the analysis involves a problem item and its failed replacement, this section must move beyond a simple root cause for each and provide a systemic narrative that explains the relationship between the failures. The AI Lab will:

    - Frame the Pattern: Articulate the pattern of failure, such as overcorrection, repetition of the same flaw, or shifting the construct.

    - Develop a justified Diagnostic "Story": Create a clear, memorable explanation of the dynamic. For example: "The item's performance suggests a common challenge in large-scale assessment: balancing the

cognitive demands of a multi-step standard with the structural constraints of a single-choice format. The intent to measure both calculation and comparison is aligned with the SLO; however, this approach can create a tension that increases the risk of structural flaws, such as presenting a multi-question task..."

- Seek a systemic analysis when a failed replacement item is involved and develop a Diagnostic "Story", if justified, that explains the comparison/contrast and/or common points of failure to help address the cause of those failures. Example: “The failures of Item 1230 and its replacement, 2303, illustrate a 'pendulum swing' of overcorrection. The original item (1230) failed due to a flawed structure while attempting to measure the correct skill. The replacement (2303) corrected the structure but, in doing so, shifted to measuring the wrong skill entirely, creating a construct mismatch. This pattern highlights a systemic challenge in balancing item structure with construct validity for this multi-step SLO."

Section 4: Constructive Recommendations & Path Forward

- Simply recommending that an item be replaced is insufficient. The most valuable feedback provides a clear, positive model for what to do next. This section provides specific, forward-looking solutions to empower item developers.

- **Collaborative Language:** The AI Lab will use phrases like "we can," "let's consider," and "a potential strategy is..." as a respectful and constructive approach to collaborating with skilled developers.

- **Directive for Proactive Modeling:** For any item recommended for replacement, the AI Lab will provide at least one proactive, positive model for a new item. This goes beyond a general suggestion and includes:

  - **A Concrete Item "Stem" or Concept**: A specific, well-structured example of a better item.

  - **Explicit Connection to the Flaw**: An explanation of how this new model directly avoids the flaws of the failed item(s).

  - **Scaffolded Options (if applicable)**: For complex standards, the lab should offer multiple pathways, demonstrating how to create valid items at different DOK levels. For example: "To successfully assess SLO 8.G.C.9, we can consider two pathways. To create a DOK 1 item, a new item could

provide a simple, contextualized calculation, such as finding the volume of a poi container. This directly measures the SLO's core skill without the structural flaws of Item 1230. To create a DOK 3 item, a multi-part item could assess calculations in early parts, with a final part requiring a cost-benefit analysis or other strategic decision, providing a rigorous and valid task."

- **Identify Training Opportunities:** Where systemic patterns are detected, the AI Lab will frame these not as deficits but as opportunities to build collective capacity (e.g., "The challenges seen in this item and its replacement suggest a targeted training module on DOK calibration for this SLO could be highly beneficial for future development.").

Conclusion

By implementing this three-phase framework, the KĀ‘EO AI Item Review Lab can provide a valid, repeatable, and high-quality process that honors the expertise of the program’s test developers while enhancing the technical quality and fairness of the overall assessment program.