AI Lab Project Instructions

This document serves as the project brief for the initial AI Lab reviews and analysis of “problematic” KĀ‘EO items. All analytical work and reporting will be conducted in strict accordance with the *Frameworks for AI Lab Reviews of KĀ‘EO Items* to ensure consistency, rigor, and quality of the final outputs.

AI Lab Project: Analysis of Poorly Performing KĀ‘EO Assessment Items

Project Purpose

To leverage AI as a tool to identify inherent flaws in the construction of previously flagged "problem" assessment items (written in Hawaiian) based on established practices for large-scale, standards-based item development. The analysis will consider factors such as actual vs. assigned cognitive complexity (Depth of Knowledge - DOK), reading load, adherence to universal design principles, and potential biases, to pinpoint reasons for psychometric red flags or item failure.

Core AI Principles in Practice

This project will strictly adhere to the KĀ‘EO AI Principles and Policy Framework, ensuring:

- **Stewardship of Student Data & Linguistic Integrity:** KĀ‘EO data, especially in Hawaiian, will be protected with the highest standards and used only within a closed system for this project, without external model training or public release of inputs/outputs.

- **Prioritizing Human Expertise:** AI will serve as an assistive tool, with qualified educators retaining all final authority and providing critical human oversight for cultural nuance and overall judgment.

- **Bias Mitigation:** The AI itself will be reviewed for cultural sustainability, absence of bias (gender, geographic, linguistic, socioeconomic, stereotypical representations), and alignment to universal design principles. Identified biases must be addressed promptly.

Required Inputs for Analysis (AI & Human Experts)

Problem Item Data:

- The specific "problem" assessment items (in Hawaiian)
- Associated psychometric data for each item; for example:
    - Point biserial correlations (less than .15 considered poor/marginal)
    - Difficulty p-value (less than .20 or greater than .95 flagged)
    - Omit rate (higher than 10% flagged)
    - Rasch item fit statistics (infit/outfit mean square values, RMSR)
    - Differential Item Functioning (DIF) analysis results, particularly for gender, Free or Reduced-Price Lunch (FRPL) eligibility, and Hawaiian language usage
    - Rating scale functionality for open-ended items (e.g., non-distinct peaks, lack of score point separation)

- Assigned KĀʻEO Student Learning Objectives (SLOs) and SLO codes
- Assigned Depth of Knowledge (DOK) levels
- Assigned Claims and Targets (math items)
- Assigned Difficulty Levels
- Item Type metadata (e.g., SC, MC, DD, FTG, SA, ER)
- Any available qualitative feedback from item writing workshops, internal reviews, or cognitive interviews

KĀʻEO Reference Documents (where applicable):

- **KĀʻEO AI Policy Draft (v4.0):** Outlines principles, acceptable/prohibited uses, and oversight
- **KĀʻEO Item Development Manual (2023):** Provides guidelines for planning, developing, and managing KĀʻEO test content
- **KĀʻEO Technical Report (2024):** Details development processes, validity, reliability, fairness, and technical analyses
- **KĀʻEO Parent Information Booklet (2024-2025):** Context for public communication and visual examples of some item types

- **SLO Tables (Excel files):** User-provided document containing the detailed Hawaiian Language Arts, Mathematics, and Science SLOs
- **Crosswalk Documents:** Available for Hawaiian Language Arts and Science (not Math) for links to CCSS/NGSS
- **DOK Resources:** "DOK wheel" diagram, DOK question stems (Appendix D), Depth of Knowledge: Potential Activities (Appendix E), and "Applying Webb's Depth-of-Knowledge (DOK) Levels in Science" (Appendix F)
- **Universal Design & Readability:** "Creating Better Tests for Everyone Through Universally Designed Assessments" (Appendix H) and "Test Item Readability Checklist" (Appendix G)
- **Bias and Sensitivity Rubrics:** (e.g., Figure 19, 20 from Item Development Manual)
- **Scoring Materials:** Generic and item-specific rubrics for open-ended items, including Hawaiian language writing conventions
- **Well-performing Exemplars**: a curated set of at least 3-5 exemplar items for the relevant grade level and content area that are performing well psychometrically (e.g., positive item-rest correlation, acceptable p-value, no adverse DIF flags). These items serve as a critical reference point for the AI Lab. They are not necessarily aligned to the same SLO as the problem item, but they provide a baseline for what successful, high-quality items look like in terms of structure, linguistic clarity, and cognitive demand. By analyzing these exemplars alongside problematic items, the AI can generate more concrete, viable, and contextually appropriate recommendations for revision and replacement

AI Lab Process Workflow

Phase 1: Data Ingestion & Setup (Human-Led)

Secure Environment: Ensure the AI analysis is conducted within a secure, closed system, compliant with KĀʻEO data protection protocols.

Document Ingestion: Ingest all relevant KĀʻEO reference documents and the Excel files containing the actual SLOs into the AI system.

Problem Item Identification: Input the specific "problem" items, their assigned metadata (SLO, DOK, difficulty, item type), and all available psychometric data (point biserial, p-value, omit rate, fit statistics, DIF results, rating scale issues).

Phase 2: AI-Assisted Item Analysis

Note: all analytical work and reporting will be conducted in strict accordance with the *AI Lab Frameworks for Reviewing KĀ‘EO Items* which provides more detailed instructions for analysis based on the outline below.

The AI system will perform the following analyses for each flagged item, generating preliminary findings:

1. Psychometric Flag Interpretation:

- **For items with low point biserial:** Suggest if the item is not discriminating well between high- and low-performing students
- **For items with extreme p-values (too easy/difficult):** Indicate if the item's actual difficulty deviates significantly from its intended difficulty or DOK level
- **For items with poor Rasch item fit:** Suggest potential issues like confusing wording, unnecessary complexity, or construct mismatch
- **For items flagged with DIF:** Highlight the specific demographic group(s) (gender, FRPL, language usage) and suggest potential underlying reasons, considering Hawaiian language nuances
- **For open-ended items with rating scale issues:** Identify score points that lack differentiation or logical progression

2. Content Alignment Analysis:

- **SLO Match (primary):** Compare the item's content directly against its assigned KĀ‘EO SLO (from the Excel files and manual) to verify if it measures the stated objective
- **Overall Alignment Match (secondary)**: Compare the item's content to any secondary content-based prescribed metadata, such as Claim and Target (math)
- **Construct Relevance:** Assess if the item introduces construct-irrelevant characteristics (e.g., linguistic, cultural, cognitive barriers) that might affect student performance

3. Cognitive Complexity (DOK) Analysis:

- **Inferred DOK:** Analyze the item's inherent cognitive demands using the Webb DOK model. The AI will cross-reference the item's question structure, required

cognitive processes, and expected student activities with the DOK Wheel, DOK Question Stems, and Potential Activities documents

- **DOK Discrepancy:** Compare the inferred DOK level with the item's assigned DOK level, flagging any mismatches. This helps determine if the item is unintentionally easier or harder in its cognitive demand than intended

4. Difficulty Level Assessment:

- **Perceived vs. Actual:** Evaluate the item's perceived difficulty based on its structure and content, and compare this against the actual performance (p-value)

- **DOK-Difficulty Relationship:** Analyze if an item's difficulty is appropriate for its DOK level (e.g., a DOK 2 item is not always "harder" than a DOK 1 item)

5. Readability and Comprehensibility Analysis:

- **Hawaiian Language Review:** Apply guidelines from the KĀʻEO Essential Style Guide (Appendix A) and Test Item Readability Checklist (Appendix G) to assess clarity, conciseness, sentence length, and vocabulary in Hawaiian

- **Flagging Issues:** Identify potentially confusing or overly burdensome language, use of academic vocabulary that might not be universally accessible across Kaiapuni schools/islands, or irregular grammatical patterns

6. Universal Design & Accessibility Review:

- **Principle Adherence:** Evaluate the item's design against the seven elements of universally designed assessments (Appendix H), such as simple/clear instructions, maximum legibility, and amenability to accommodations

- **Construct-Irrelevant Barriers:** Flag visual elements, graphics, or response formats that might create barriers for the widest range of students, especially if not central to the construct being measured

7. Bias and Sensitivity Analysis:

- **Comprehensive Scan:** Review for potential biases, including cultural insensitivity, stereotypes, and representation issues, drawing on the bias/sensitivity rubrics

- **Hawaiian Context:** Pay specific attention to subtle gender bias (given gender-neutral pronouns/names in Hawaiian) and socioeconomic bias, especially in word problems or figurative language, that might impact understanding based on life experiences or exposure to language variations

- **DIF Correlation:** Correlate identified biases with DIF analysis results to understand potential causes of differential performance

Phase 3: Human Oversight, Interpretation & Recommendations (Mandatory)

Expert Review: Qualified program staff and content experts will review all AI-generated analyses and flagged issues for each "problem" item.

Refined Diagnosis: Human experts will interpret the AI's findings, applying their deep understanding of Hawaiian language, culture, pedagogical practices, and student characteristics to confirm or refine the identified flaws. This is where cultural nuance, which AI cannot fully grasp, is critically applied.

Actionable Recommendations: Based on the combined AI analysis and human expert judgment, for each "problem" item, develop concrete recommendations for:

- **Revision:** Specific linguistic, structural, DOK, or bias-related changes
- **Replacement:** Guidance on what kind of item should replace it, addressing the identified flaws and gaps in item quality when comparing “problematic” items to others with more favorable psychometric properties
- **Removal:** Justification for permanently removing the item from the item bank

Output of the AI Lab Project

For each "problem" item, the output will be a comprehensive report including:

- **Original Item Details:** Item text (in Hawaiian), assigned SLO, DOK, difficulty, item type

- **Psychometric Summary:** Key psychometric data and any associated flags (e.g., low point biserial, high omit rate, DIF for specific groups)

- **AI Analysis Findings:** Detailed breakdown of AI-identified issues related to content alignment, DOK (inferred vs. assigned), difficulty, readability, universal

design, and bias

- **Categorization of Flaws:** A categorization of the primary reasons for the item's poor performance (e.g., DOK mismatch, unclear phrasing in Hawaiian, cultural insensitivity, visual accessibility issue)

- **Human Expert Consensus:** A summary of the human experts' interpretation of the AI findings, including cultural nuances and confirmation/refinement of flaws

- **Specific Recommendations:** Clear, actionable steps for item revision, replacement, or removal, with a rationale based on the analysis.

Key Reference Information

RE: Translations

Directive on Bilingual Modeling and Linguistic Partnership:

The AI Lab operates in a bilingual context where Hawaiian is the primary language of assessment and English is the functional language of this analytical partnership. The following principles govern the AI's use of both languages:

1. Authority of Input Language: The Hawaiian language in all KĀʻEO source materials must be treated as the authoritative and definitive version for analysis. The AI Lab's role is not to evaluate or critique the quality of the provided Hawaiian language, which is the product of expert Kaiapuni educators.

2. Function of English Analysis: The AI's deep analysis, root cause hypotheses, and structural recommendations will be generated in English. This is to ensure maximum clarity and precision in communicating complex technical and psychometric concepts.

3. Standard for Output Language (Proactive Modeling): When the AI Lab provides proactive models for new or revised items as part of its recommendations, it must do so bilingually and with cultural respect. The standard is as follows:

- Structural Model (English): The English version of a suggested item will serve as the structural and conceptual model. It will demonstrate the recommended DOK, format, clarity, and adherence to item writing best practices.

- Viable Draft (Hawaiian), when applicable, in generating any model item: Alongside the English model, the AI will generate a high-quality draft of the item in Hawaiian. This draft is not intended to be a final, perfect product, but a viable and respectful starting point for Hawaiian-language-expert item developers. It must adhere to the vocabulary, phrasing, and grammatical patterns documented in the KĀʻEO Essential Style Guide (within the Item Development Manual) and observed in existing high-quality exemplar items.

- Explicit Deference: Every Hawaiian language suggestion will be accompanied by a note of deference, clarifying its purpose. For example: "Below is a conceptual model in English and a draft in Hawaiian, provided as a viable starting point for review and refinement by Hawaiian language content experts."

This approach ensures that the AI Lab's output is not only technically sound but also immediately useful and respectful to the Kaiapuni community and its master educators who are authorities on the Hawaiian language. It positions the AI as a partner that provides fully formed, bilingual "thought starters," rather than English-only solutions that would require developers to perform the extra work of translation and cultural adaptation from scratch

RE: Item IDs:

- **Formal item IDs** are always a unique number (example: 2303)
- **Item Template IDs** often contain unique alphanumeric strings (example: 2MW81713)

RE: Item Types (referenced in item templates):

Item_Type	Description	Scoring
SC	Single Choice	computer scoring
MC	Multiple Choice(s)	computer scoring
STV	Sort the Values	computer scoring
DD	Drag & Drop	computer scoring
TIDD	Text-Image Drag & Drop	computer scoring
DMD	Drag & Multi-Drop	computer scoring
MP	Multi-Part	computer scoring
FTG	Fill In the Gap	hand scoring
SA	Short Answer	hand scoring

ER	Extended Response	hand scoring
PT	Performance Task (HLA only)	hand scoring

Next Step Suggestion

One way to further refine this process is to consider developing specific, quantitative rubrics for evaluating the AI's output. These rubrics could be used by humans to assess the AI's accuracy in identifying DOK discrepancies, readability issues, or potential biases, and general item development helpfulness while allowing for continuous improvement of the AI tool itself and a more standardized "go/no-go" decision for its application—especially in sensitive areas where cultural nuance may override AI accuracy or otherwise put an alternate perspective on AI's observations, strategies, and recommendations.