**West Science Information System and Technology**
Vol. 4, No. 02, August 2026, pp. 270~287

# Modular Integration of Raw NASA-TLX and VisAWI-S in UIX-Probe: An Analytics Dashboard for Supporting UX Evaluation

**Anugrah Nur Rahmanto[1], Retno Damayanti[2], Muhammad Abel Rif'at Nandito[3]**
[1-3] Department of Information Technology, Politeknik Negeri Malang, Indonesia

**Article Info**

***Article history:***

Received Aug, 2026
Revised Aug, 2026
Accepted Aug, 2026

***Keywords:***

UI/UX Evaluation
Raw NASA-TLX
VisAWI-S
Analytics Dashboard
Multi-Method Evaluation Platform

**ABSTRACT**

Evaluation of User Interface (UI) and User Experience (UX) typically relies on several questionnaire instruments that are administered and scored separately, making score aggregation and cross-method triangulation labor-intensive. UIX-Probe is a web-based multi-method UI/UX assessment platform that already integrated the System Usability Scale (SUS), the Technology Acceptance Model (TAM), A/B testing and the Web Content Accessibility Guidelines (WCAG), but lacked instruments for two constructs central to user experience: subjective mental workload and perceived visual aesthetics. This paper reports the design, implementation and workflow of two new modules integrated into UIX-Probe Raw NASA-TLX and the Short Visual Aesthetics of Websites Inventory (VisAWI-S) with a focus on the analytics dashboard that turns raw responses into interpretable output for developers and interface designers. The dashboard provides a composite score with automatic classification, a per-dimension mean comparison, per-question response distributions, a per-respondent results table, respondent demographics, and automatically generated narrative descriptions, all exportable to Excel. The system was built using the Waterfall SDLC on a Laravel 10 + React.js (Inertia.js) architecture. Functionality was verified under real-world deployment: five independent developer teams created and distributed their own evaluation surveys on the platform, collecting 115 responses across five surveys. The four surveys using the new modules produced Raw NASA-TLX composite scores of 10.99–21.75/100 (all Moderate) and VisAWI-S composite scores of 4.23–5.81/7 (Positive to Very Positive). All three verification criteria were met: 5/5 teams configured a survey successfully, every survey obtained real respondents, and every dashboard rendered without error.



***Corresponding Author:***

Name: Anugrah Nur Rahmanto
Institution: Department of Information Technology, Politeknik Negeri Malang, Indonesia
Email: anugrahnur@polinema.ac.id

## 1. INTRODUCTION

The quality of User Interface (UI) and User Experience (UX) design has become one of the determining factors for the success of digital products today, since a product that is functionally correct but frustrating or aesthetically unconvincing can still fail to retain users in a competitive market. ISO 9241-210 (2019) defines the goals of human-centred design as encompassing effectiveness,

efficiency, and user satisfaction, with the satisfaction aspect covering both cognitive comfort and the perceived aesthetics of the interface [1]. These are not interchangeable dimensions: a system can be efficient yet mentally taxing, or aesthetically pleasing yet frustrating to operate, which is why evaluating only one dimension risks an incomplete picture of the overall experience. An extensive HCI literature review by Hartson and Pyla shows that UX evaluation is inherently multi-method in nature, since no single instrument captures all constructs relevant to an interactive system — a usability questionnaire says little about visual appeal, and an aesthetics instrument says nothing about mental workload — so that cross-method triangulation provides a more representative picture of the user experience than any single instrument alone [2].

In practice, however, the UI/UX evaluation process is still often fragmented. Different constructs such as usability, technology acceptance, accessibility, mental workload, and visual aesthetics are generally measured with separate questionnaires, whether through printed forms, statistical spreadsheets, or general-purpose survey applications. This fragmentation gives rise to three practical problems that frequently arise in research and product teams. First, manual score calculation is prone to error, especially when each instrument has a different scale (0-100 versus a 1-7 Likert scale) and different aggregation rules, so a reversed anchor or a simple transcription mistake can silently distort a composite score. Second, data collected through heterogeneous instruments becomes difficult to triangulate even when it originates from the same respondents, since incompatible scales stored in separate files cannot be directly compared without additional manual reconciliation. Third, the absence of integrated analytics makes it difficult to compare design iterations, since each evaluation cycle must be aggregated from scratch, leaving teams with no lightweight way to check whether a revision actually improved the metrics that mattered [2].

UIX-Probe is a web-based UI/UX assessment platform that has been developed incrementally through several research cycles to address this fragmentation, with each cycle typically adding one new evaluation method to the shared platform rather than building a separate tool for every individual study (Figure 1). In the previous development stages, the SUS and TAM modules were developed in earlier undergraduate projects at the authors' institution and are documented in unpublished theses at the institutional repository; the platform had already integrated the System Usability Scale (SUS) [3], [4], the Technology Acceptance Model (TAM) [5], A/B Testing, and the Web Content Accessibility Guidelines (WCAG) [6] into a single web-based environment. Taken together, these four instruments cover four related but distinct concerns: SUS captures a respondent's overall perceived usability of an interface on a single composite score, TAM explains whether users judge the system as useful and easy to use enough to warrant continued adoption, A/B Testing lets developers compare two competing design variants directly against real user behavior rather than relying on internal judgment alone, and WCAG checks compliance with recognized accessibility standards so that the interface remains usable for people with disabilities. Consolidating these four methods inside one platform is what allows a developer to triangulate across constructs for the same interface without manually reconciling four separately scored questionnaires, directly addressing the three practical problems described above.

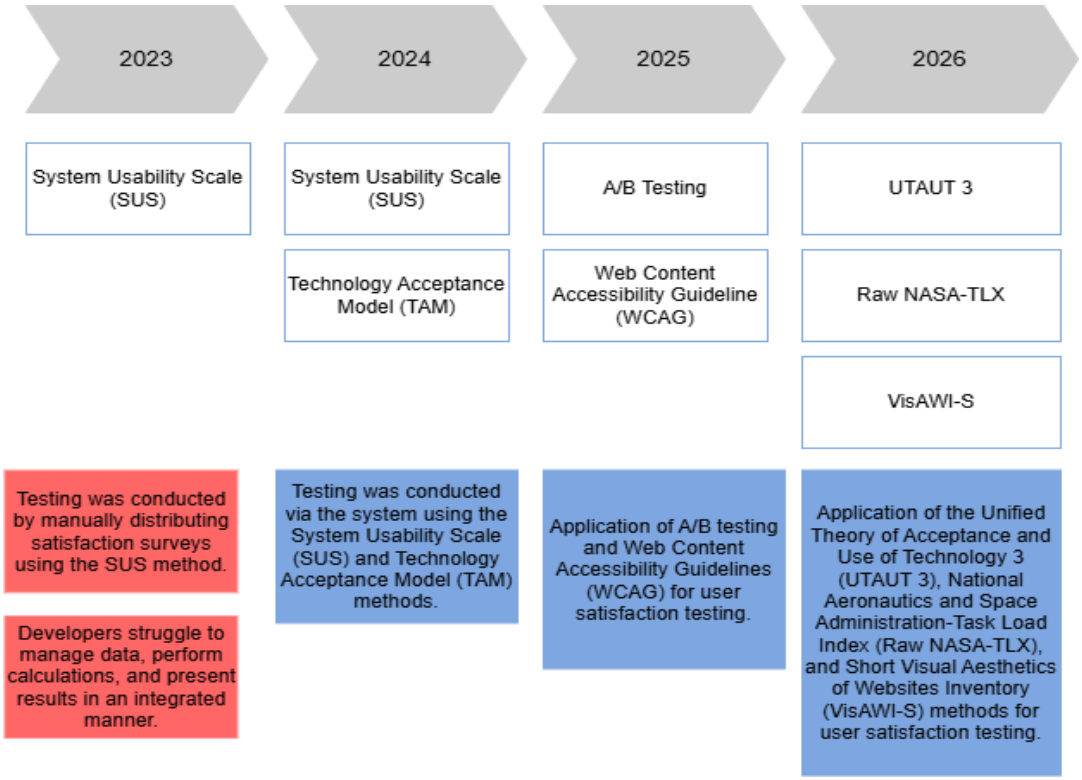


Figure 1. Timeline of evaluation-module development on the UIX-Probe platform (2023–2026). (UTAUT 3 also appears in Figure 1 as part of the platform's full 2026 development roadmap; its evaluation is outside the scope of this paper, which reports only the two modules Raw NASA-TLX and VisAWI-S also introduced that year.)

## 2. LITERATURE REVIEW

### *2.1 Raw NASA-TLX*

The NASA Task Load Index (NASA-TLX) was developed by Hart and Staveland as a multi-dimensional subjective rating scale for measuring perceived workload in operator tasks [7]. This instrument decomposes workload into six dimensions, namely Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. In the original NASA-TLX procedure, the six dimensions are combined through pairwise weighting in which respondents determine the relative importance of each dimension.

Hart, in a two-decade review of NASA-TLX usage, observed that the weighting step is frequently omitted without significantly reducing measurement quality [8]. This unweighted version, known as Raw NASA-TLX, has now become a widely applied alternative, particularly in software and web evaluation contexts where the cost of weighting is high. The validity of Raw NASA-TLX in digital contexts has been empirically confirmed through a number of recent studies. Said et al., in a pooled analysis of digital patient-monitoring tasks, reported that Raw NASA-TLX has adequate sensitivity to differentiate between different levels of workload [9]. Ramkumar et al. combined NASA-TLX with GOMS analysis for an interactive medical-image segmentation task [10]. Lowndes

et al. used NASA-TLX to measure workload variation across surgical specialties [11], and Al Madi, Peng, and Rogers applied NASA-TLX to introductory computer-science programming projects [12]. In the Indonesian industrial context, Azizah, Nurwinata Rinaldi, and Wafiq Al-Muqaffa compared Raw NASA-TLX with the Rating Scale Mental Effort (RSME) to assess the mental workload of automotive-factory employees, and reported consistent rankings across work stations between the two methods [13]. This range of contexts illustrates the adaptability of Raw NASA-TLX to both digital and non-digital tasks.

Regarding scale conventions, the original Hart and Staveland instrument uses a continuous 0-100 scale that, in printed administration, is divided into 20 five-point segments [7]. The choice of a 5-point interval in the UIX-Probe implementation is consistent with that original administration. The Performance dimension uses the anchors Perfect (0) and Total Failure (100) in accordance with the original instrument's convention [7], so that a higher value directly indicates a worse perceived performance and contributes consistently to the overall workload score.

The composite score is calculated as the arithmetic mean of the six dimensions as follows, where MD, PD, TD, P, E, and F respectively denote Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration.

$$\text{Raw NASA-TLX Score} = (MD + PD + TD + P + E + F) / 6 \quad (1)$$

For cross-respondent studies, the aggregate value per dimension is calculated as the mean across respondents:

$$MD\_Final = (MD\_1 + MD\_2 + \ldots + MD\_n) / n \quad (2)$$

The Raw NASA-TLX composite score produced by formula (1) falls within the range of 0–100. UIX-Probe classifies this score into five workload categories, as shown in Table 1, dividing the scale into bands that align with the instrument's native 0–100 range [7]. Two properties of this choice are worth making explicit. First, the same five thresholds are applied both to the composite score shown on the KPI card and to each individual dimension in the automatically generated narrative (Subsection 3.3), so that the headline label and the per-dimension narrative can never contradict one another. Second, this finer-grained scheme was chosen over a coarser three-way split (low/moderate/high) because it distinguishes between a workload that is merely elevated (Moderate, Somewhat High) and one that is clearly problematic (High, Very High), which is useful when comparing design iterations that fall within a similar general range.

Table 1. Raw NASA-TLX Score Classification

| Score Range | Class. | Interpretation |
|---|---|---|
| 0–9 | Low | Low mental workload |
| 10–29 | Moderate | Moderate workload |

| | | |
|---|---|---|
| 30–49 | Somewhat High | Somewhat high workload |
| 50–79 | High | High mental workload |
| 80–100 | Very High | Very high workload |

### 2.2 *VisAWI and VisAWI-S*

The Visual Aesthetics of Websites Inventory (VisAWI) was developed by Moshagen and Thielsch based on a factor analysis that identified four aspects of website aesthetic perception, namely Simplicity (clarity and orderliness of layout), Diversity (richness and visual dynamics), Colorfulness (quality and harmony of color), and Craftsmanship (perceived skill and diligence of the design) [14]. The full version of VisAWI consists of 18 items spread across four sub-scales. The background of the original German-language development of VisAWI can be traced to Thielsch and Moshagen [15].

Moshagen and Thielsch subsequently introduced VisAWI-S as a shortened version designed for research contexts with a limited respondent burden [16]. It is important to note that VisAWI-S consists of four items, one item per aspect, and is intended to capture the overall aesthetic factor as a brief approximation of the full VisAWI. This is confirmed by the VisAWI Manual [17], which states that VisAWI-S "measures the general factor of website aesthetics" with one item per aspect. Accordingly, the four-item implementation on UIX-Probe is consistent with the instrument's specification. The psychometric consequence of one item per aspect the fact that a per-subscale Cronbach's $\alpha$ cannot be computed is an inherent characteristic of the short-form instrument itself, not of the UIX-Probe implementation; the reliability of VisAWI-S as reported by Moshagen and Thielsch was calculated on the general factor score based on all four items [16].

The VisAWI-S items are rated on a 7-point Likert scale ranging from "strongly disagree" to "strongly agree". The score for each aspect is the mean of the related item across respondents, and the VisAWI-S composite score is the mean of the four aspect scores, with S, D, C, and Cr respectively denoting Simplicity, Diversity, Colorfulness, and Craftsmanship.

$$S\_Simplicity, Final = (S\_1 + S\_2 + \dots + S\_n) / n \quad (3)$$

$$VisAWI\text{-}S\ Score = (S\_Final + D\_Final + C\_Final + Cr\_Final) / 4 \quad (4)$$

Hirschfeld and Thielsch established a ROC-based cut-point of 4.5 as a commonly used threshold for judging a website as aesthetically pleasing overall [18]. The VisAWI-S composite score produced by formula (4) falls within the range of 1–7 and is classified into the four categories shown in Table 2. As with Raw NASA-TLX, the same thresholds are applied to the composite score and to each individual aspect. The scale is divided into four bands of approximately equal width (1.0 – < 2.6, 2.6 – < 4.1, 4.1 – < 5.6, and 5.6–7.0), giving a finer-grained distinction between a merely acceptable perception (Positive)

and a strongly favorable one (Very Positive) than a single Positive/Negative split would allow. The Hirschfeld and Thielsch ROC cut-point of 4.5 [18] falls within the Positive band and is displayed alongside the score on the dashboard so that survey owners can additionally compare their result with the international literature.

Table 2. VisAWI-S Score Classification

| Score Range | Class. | Interpretation |
|---|---|---|
| 1.0 – < 2.6 | Negative | Negative aesthetics |
| 2.6 – < 4.1 | Fair | Fair aesthetics |
| 4.1 – < 5.6 | Positive | Positive aesthetics |
| 5.6 – 7.0 | Very Positive | Very good aesthetics |

## 3. METHODS

### 3.1 Development Cycle and Justification

Both modules were developed using the Waterfall SDLC, consisting of the stages of requirements analysis, design, implementation, testing, and maintenance. Sommerville, in a comparative review of development methodologies, states that Waterfall remains a valid choice for projects with three characteristics: (a) stable and well-defined requirements from the outset, (b) predetermined technical constraints, and (c) a limited scope of work [19]. All three characteristics apply to this project: the module requirements were fully determined by the original instruments (the six NASA-TLX dimensions, the four VisAWI-S items), which have been standardized for several decades and did not evolve during the course of the project [8], [16]; the technical constraints were determined by the already-established conventions of the UIX-Probe platform; and the scope was limited to adding two modules without modifying any other features.

### 3.2 System Architecture

UIX-Probe is built on Laravel 10 (PHP) as the backend and React.js as the frontend, bridged by Inertia.js. Persistence is served by a relational database. The new modules are implemented as additions within this stack, following the controller-model-view convention and the routing scheme of the parent platform.

Two new tables were added to store per-respondent scores, namely `nasa_tlx_scores` and `visawi_s_scores`. Each table is linked to the survey, the user, and the parent `survey_responses` record through foreign keys with the `cascadeOnDelete()` rule. The `nasa_tlx_scores` table contains six integer columns for per-dimension ratings (range 0-100) and a `final_score` column of type `decimal(5,2)`. The `visawi_s_scores` table contains four integer columns for the aesthetic aspects (Likert range 1-7) and a `final_score` column of type `decimal(5,2)`. Both tables are created through the `create_nasa_tlx_scores_table` and `create_visawi_s_scores_table` migrations. An additional migration, `update_method_type_enum_in_survey_ai_recommendations_table`, extends the `method_type` enumeration on the `survey_ai_recommendations` table to accept the values 'NASA-TLX' and 'VisAWI-S'.

### 3.3 Form Design and Score-Calculation Procedure

Two React components were implemented for response capture. The `NasaTlxForm.jsx` component displays the six Raw NASA-TLX dimensions in separate cards, each containing a 0-100 range slider with a 5-point interval and a real-time value badge. The `VisawiSForm.jsx` component displays the four VisAWI-S items, one per aspect, in accordance with the original

instrument [16], in separate cards, each containing seven radio inputs representing the Likert scale points. All radio inputs include the HTML `required` attribute. Both forms save their current state to `localStorage` through a `useEffect` hook to allow recovery after a page reload.

Score calculation is performed server-side in the `FormController::store()` method, within a single database transaction (`DB::beginTransaction` / `DB::commit`). The procedure runs in three steps. First, a record is inserted into `survey_responses`, which holds the demographic data and the raw response payload in the `response_data` JSON column. Second, if the payload contains a `nasa_tlx` block, the controller calculates `final_score` using formula (1) and inserts a record into `nasa_tlx_scores`. Third, if the payload contains a `visawi_s` block, the controller calculates `final_score` using formula (4) and inserts a record into `visawi_s_scores`. If any step fails, the transaction is rolled back (`DB::rollBack`); if all steps succeed, the transaction is committed.

On the dashboard side, `NasaTlxController` and `VisawiSController` calculate the mean `final_score` across all respondents through a `foreach` iteration over a score collection retrieved from `Cache::remember()` with a 2-hour TTL, using a formula equivalent to formulas (2) and (3) for the cross-respondent context. In addition to the aggregate score and overall classification (Table 1 and Table 2), both controllers also generate a per-dimension narrative description through the `getResumeDescription()` method. This method classifies the mean of each individual dimension using the same thresholds as the composite classification 0–9, 10–29, 30–49, 50–79, and 80–100 for Raw NASA-TLX (Table 1), and 1.0 – < 2.6, 2.6 – < 4.1, 4.1 – < 5.6, and 5.6 – 7.0 for VisAWI-S (Table 2). Applying one threshold set at both levels is a deliberate design decision: it guarantees that the categorical label on the KPI card and the wording of the per-dimension narrative are always mutually consistent. The per-dimension results are combined into a single descriptive paragraph that helps developers locate the interface's strengths and weaknesses on each individual aspect, complementing the overall score on the KPI card.

### 3.4 Functional Verification under Real-World Deployment

The two modules and their analytics dashboard were verified through real-world deployment rather than through a controlled acceptance study. Five independent developer teams, none of them members of the core development team, created and ran their own UI/UX evaluation surveys on the UIX-Probe platform for their respective applications. The purpose of this procedure is deliberately bounded: it establishes that the complete workflow survey configuration, respondent-facing delivery, score computation, and dashboard rendering executes correctly outside the development environment and on data the research team did not generate. It does not, and is not intended to, measure perceived usability, user acceptance, or the decision-support effectiveness of the dashboard; those are addressed as follow-up work in Subsection 4.7.

A summary of the five developer samples together with the number of respondents and the modules used is shown in Table 3. Four of the five surveys used the Raw NASA-TLX and/or VisAWI-S modules, which are the main contribution of this research, while one survey (Redesign of the SKKM Polinema Website UI/UX) used the SUS, TAM, and A/B Testing modules that were already available on the system, serving as evidence of the stability of the existing modules after the addition of the new modules.

Table 3. Summary of the Five Developer Samples (Functional Verification)

| No. | Survey Name | Respondents | Modules Used |
|---|---|---|---|
| 1 | MONSI – Monitoring System | 20 | Raw NASA-TLX |
| 2 | Coco Apps | 22 | Raw NASA-TLX, VisAWI-S |
| 3 | LabForLapt | 21 | Raw NASA-TLX, VisAWI-S |
| 4 | Web Quest | 21 | Raw NASA-TLX, VisAWI-S |
| 5 | Redesign of the SKKM Polinema Website UI/UX | 31 | SUS, TAM, A/B Testing |
| | Total | 115 | |

The number of respondents per survey (20–31) was determined independently by each developer team according to the reach of their own application. These volumes are reported here as a description of the deployment conditions rather than as a basis for statistical inference about the evaluated products; the verification criteria below concern system behavior, not the precision of any individual product's score.

The functional verification procedure is as follows: (1) developer teams create a survey on the UIX-Probe platform by selecting the required evaluation modules; (2) the survey is distributed to real respondents independently by each team; (3) the system automatically generates a results dashboard after respondents complete the survey; and (4) system functionality is evaluated against three verification criteria. Each developer sample carried out this sequence independently, without direct guidance from the research team.

The system is declared to have passed functional verification if it meets three criteria: (1) all developer samples (5/5) successfully created a survey with the selected method configuration; (2) each survey successfully obtained real respondents; and (3) the system generated a results dashboard including total score, classification, per-dimension charts, demographic data, and automatic narrative descriptions without errors for every survey created. These three criteria were chosen because they correspond to the three points in the workflow where the modules could plausibly fail in practice: survey configuration (criterion 1), respondent-facing delivery (criterion 2), and result computation and rendering (criterion 3). A failure at any single stage would have been sufficient to indicate the module was not yet ready for real-world use regardless of whether the other two stages succeeded, which is why all three are treated as jointly necessary rather than merely indicative. Because the dashboard's score cache carries a 2-hour TTL (Subsection 3.3), each dashboard check against criterion 3 was performed either after the TTL had elapsed or after a manual cache clear, so that the outputs inspected reflected freshly computed values rather than a potentially stale cache read.

## 4. RESULTS AND DISCUSSION

### 4.1 System Implementation

This subsection describes the implementation of the Raw NASA-TLX and VisAWI-S modules on the UIX-Probe platform in sequence, starting from survey creation by the owner through to the presentation of aggregate results via the analytics dashboard. The discussion is divided into three parts. First, the flow of survey creation and form completion by respondents, covering method selection in the admin interface and the display of the corresponding questionnaire to respondents. Second, the data submission and persistence mechanism, explaining the JSON payload structure, the transactional procedure in FormController::store(), and the governance of the local data footprint. Third, the complete anatomy of the analytics dashboard the part that is the main focus of this paper together with an explanation of how each component helps developers and interface designers interpret evaluation results and determine design-improvement priorities.

#### 4.1.1 Survey Creation and Form-Completion Flow

Authenticated users with the appropriate access rights can access

`/account/surveys/create`, which renders the `Create.jsx` component. The form captures the title, theme, a rich-text description (Quill.js), the URL of the system to be evaluated, one or more category checkboxes, and one or more method checkboxes. Among the method checkboxes are `#check-methods-5` for Raw NASA-TLX and `#check-methods-6` for VisAWI-S. When the form is submitted, the survey record is saved to `surveys`, and the method association is saved to the `survey_has_methods` pivot table with the corresponding `method_id`.

The complete flow involving the three actors in this module the system developer (survey owner), the system itself, and the respondent is summarized in Figure 2. The diagram shows the branching based on the selected method (Raw NASA-TLX, VisAWI-S, or both) as well as the data-persistence points in the database before the results are displayed back to the survey owner.

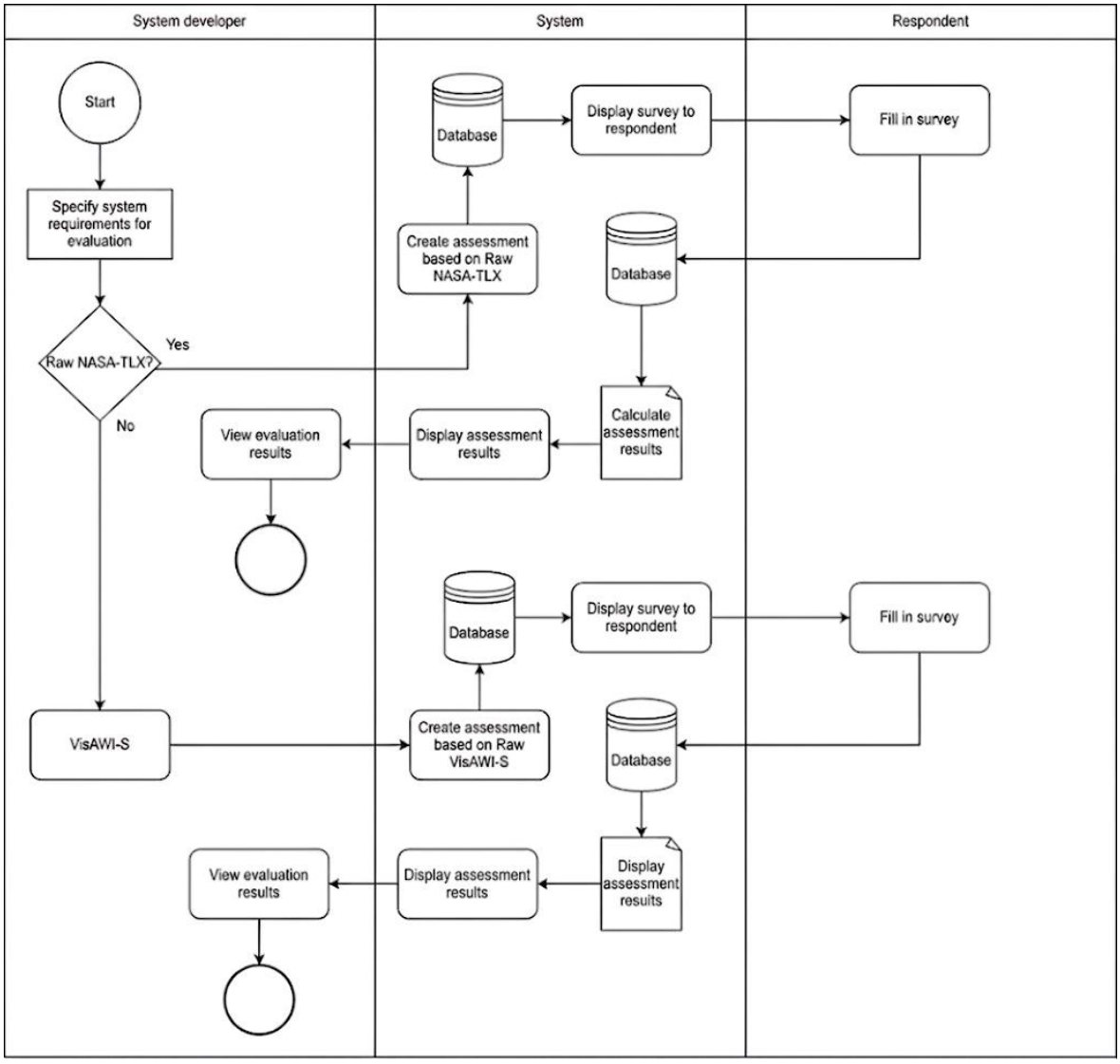


Figure 2. Business-process flow diagram of the Raw NASA-TLX and VisAWI-S modules on the UIX-Probe platform.

#### *4.1.2 Submission and Persistence*

When the "Submit Survey Response" button is pressed, the system constructs a `responseData` object containing the raw ratings and demographic data. For NASA-

TLX, the object contains a nested block as follows:

```
nasa_tlx: {
  mental_demand: <0-100>,
  physical_demand: <0-100>,
  temporal_demand: <0-100>,
  performance: <0-100>,
  effort: <0-100>,
  frustration: <0-100>
}
```

For VisAWI-S, the object contains the following block:

```
visawi_s: {
  simplicity: <1-7>,
  diversity: <1-7>,
  colorfulness: <1-7>,
  craftsmanship: <1-7>
}
```

The object is stringified as JSON and sent via `Inertia.post()` to `/form/{id}/{slug}`. After the commit succeeds, the `localStorage` key for that response is deleted to minimize the local data footprint.

#### 4.1.3 Anatomy of the Analytics Dashboard

Survey owners can access aggregate results via `/account/nasa-tlx` and `/account/visawi-s`. The `index()` method redirects to the user's first survey that uses the relevant method. The `show($id)` method performs authorization (`nasa_tlx.index.full` or `visawi_s.index.full`, with an ownership-based access path), retrieves the list of surveys for the navigation dropdown, and retrieves the related score data. This retrieval is cached via `Cache::remember()` with a 2-hour TTL.

This subsection describes each component of the Raw NASA-TLX and VisAWI-S results dashboard in sequence, from the aggregated summary to per-question detail, together with an explanation of the usefulness of each component for developers and interface designers who wish to act on the evaluation results.

**1) KPI Card and Automatic Classification**

The dashboard's interpretation rules are those defined in Table 1 and Table 2, applied to the composite score. For Raw NASA-TLX (range 0–100): low workload (0–9), moderate workload (10–29), somewhat high workload (30–49), high workload (50–79), very high workload (80–100). For VisAWI-S (range 1–7): negative aesthetics (1.0 – < 2.6), fair aesthetics (2.6 – < 4.1), positive aesthetics (4.1 – < 5.6), very positive aesthetics (5.6 – 7.0). As a complement, the dashboard also presents the empirical ROC threshold of 4.5 [18] so that survey owners can position their score relative to the international literature.

**2) Average-per-Dimension Chart**

Below the KPI card, the dashboard displays an "Average per Dimension" pie chart that divides the composite score into the proportional contribution of each constituent dimension. Figure 3 shows an example of this chart for VisAWI-S results, with the four aspects Simplicity, Diversity, Colorfulness, and Craftsmanship each shown together with their mean value in the legend. Although the composite score is an arithmetic mean rather than a sum so the dimensions do not stand in a strict part-to-whole relationship the proportional display is used here as a comparative aid to identify relative dimension magnitudes at a glance, complementing the precise per-dimension means already available on the KPI card and in the per-respondent results table (point d).

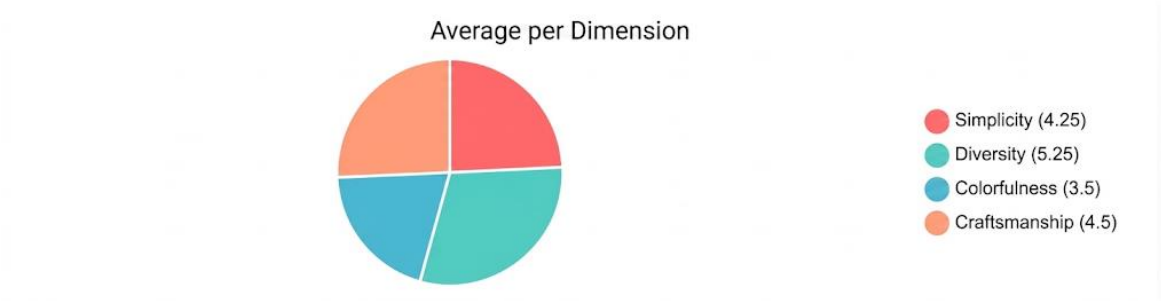


Figure 3. Average-per-dimension chart on the VisAWI-S results dashboard (illustrative; synthetic data, n = 4).

This visualization is useful for developers because the single composite score on the KPI card does not show which dimension contributes the most or is the weakest. By looking at the proportions in Figure 3, developers can immediately identify the dimension with the lowest mean value as a priority candidate for improvement for example, if Colorfulness shows a mean value far lower than the other three aspects, interface designers can focus their design revisions on the color palette and visual contrast without needing to trawl through the entire raw dataset. The same chart is available for all six Raw NASA-TLX dimensions (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration), enabling developers to identify which workload dimension is most dominantly felt by respondents.

### 3) Per-Question Response Distribution Chart

The "Chart of Results for Each Dimension" accordion presents one pie chart for every question/dimension, showing the full distribution of respondents' answers across all scale points. For VisAWI-S, each pie chart shows the proportion of respondents who chose each of the seven Likert points (1 = Strongly Disagree to 7 = Strongly Agree) for a given aspect, as shown in Figure 4 for the Simplicity, Diversity, and Colorfulness aspects.


Figure 4. Per-question response distribution chart for VisAWI-S (illustrative; synthetic data, n = 4).

This question-level granularity complements the average-per-dimension chart in Figure 3. As an illustration, two dimensions can have the same mean value yet

very different distribution patterns: the first dimension might receive uniform answers clustered around the middle of the scale (indicating a consistently neutral perception), while the second might be polarized some respondents choosing the low extreme and others the high extreme yet happen to share the same mean. Without the distribution chart, these two situations cannot be distinguished from the composite score or the per-dimension mean alone. For developers, a polarized distribution on a given question can be a signal that the related interface element is perceived very differently by different user groups (e.g., new users versus long-time users), so that design improvements may need to be tailored per user segment rather than applied uniformly. The same pattern applies to all six Raw NASA-TLX dimensions on the 0-100 scale.

**4) Per-Respondent Results Table**

The results-table accordion presents one row of data per respondent, containing the raw values for the six dimensions (NASA-TLX) or four aspects (VisAWI-S) together with the `final_score` calculated from formula (1) or (4). This table allows developers to perform case-by-case inspection for example, tracing respondents with an extreme `final_score` (very high or very low) to understand the underlying dimension profile, something that is not visible in the overall aggregate. This table is also the data unit exported when the "Export to Excel" action is used (see point f).

**5) Respondent Demographic Data**

The demographics accordion displays pie charts for the gender, profession, educational background, and age range of respondents, as shown in Figure 5.

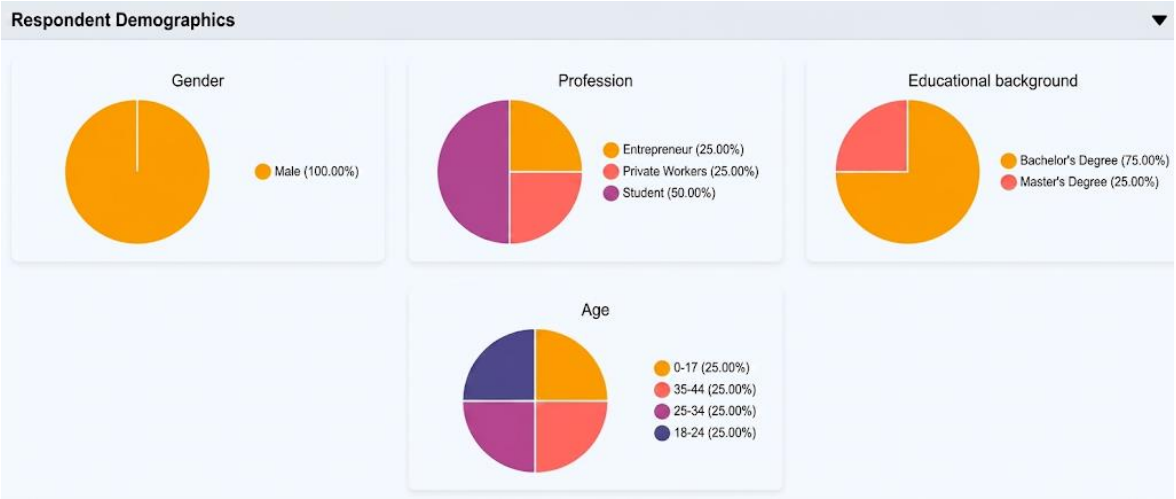


Figure 5. Respondent demographic charts on the results dashboard (illustrative; synthetic data, n = 4; actual samples were dominated by the 18–24 age group).

This demographic information helps developers interpret the evaluation results in context for example, a high Raw NASA-TLX composite score among a group of respondents with a particular educational background or profession can indicate that the interface is specifically less friendly to that user segment, so design improvements can be targeted at that segment without changing the experience of other users for whom it is already good. Demographics are also useful for assessing the extent to which the evaluation results can be generalized to the target user population of the system being evaluated.

**6) Automatic Narrative Description and Excel Export**

The summary card presents a per-dimension narrative description generated by `getResumeDescription()`, as explained in Subsection 3.3. For example, for a Mental

Demand dimension whose mean falls within the Low band (0–9), the dashboard displays the sentence "The mental workload of [survey theme] is considered low by most users"; for a VisAWI-S Simplicity dimension whose mean falls in the Very Positive band (5.6–7.0), the dashboard displays the sentence "The layout of [survey theme] is very simple and clearly understood by users". This narrative description translates the numbers in Figures 3 and 4 into ready-to-use sentences that can be directly included in an evaluation report or a design-iteration planning document, without developers needing to formulate their own interpretation. The LLM recommendation card displays follow-up recommendations based on the cached aggregate score (see Subsection 4.5 for a critical discussion of this component).

For further analysis beyond what the dashboard presents for instance additional statistical analysis, custom visualizations, or merging with data from other evaluation methods on UIX-Probe (SUS, TAM, A/B Testing, WCAG) the "Export to Excel" action is available to users with the appropriate permissions. This export produces a file containing the raw per-respondent data (as shown in the results table in point d), so that developers can perform further data processing in their statistical software or spreadsheet tool of choice, without being limited to the visualizations provided by the dashboard by default.

### 4.2 Functional Verification Results

Five developer samples independently used the UIX-Probe system to create UI/UX evaluation surveys for their respective applications, as described in Subsection 3.4. This subsection presents the test results descriptively; a discussion of the meaning of these results is presented in Subsection 4.3.

#### 4.2.1 Usage Summary Across the Five Developer Samples

Based on Table 3, the four surveys using the Raw NASA-TLX and/or VisAWI-S modules obtained a total of 84 respondents (20+22+21+21), while the overall total across all five surveys reached 115 respondents. The variation in the number of respondents between surveys (20 to 31 respondents) shows that the five developer samples used the system independently and in different application contexts, so that this data represents real-world usage conditions.

The system is declared to have passed the usage test if it meets three criteria: (1) all developer samples (5/5) successfully created a survey; (2) every survey obtained real respondents; and (3) the results dashboard displayed without errors. Based on the detailed results in Subsections 4.2.2 and 4.2.3, all three criteria were met for all five developer samples.

#### 4.2.2 Results of the Raw NASA-TLX and VisAWI-S Surveys (4 Samples)

The first four surveys used the Raw NASA-TLX and/or VisAWI-S modules. A synthesis of the main scores is shown in Table 4. All Raw NASA-TLX composite scores fall in the range 10.99–21.75 out of 100 and are classified as Moderate according to Table 1, indicating that the evaluated applications imposed a moderate perceived mental workload on their users. VisAWI-S composite scores range from 4.23 to 5.81 out of 7, with one survey classified as Very Positive and two as Positive according to Table 2.

Table 4. Synthesis of the Main Scores for the Four Raw NASA-TLX and VisAWI-S Surveys

| Survey Name | Resp. | NASA-TLX Score | NASA-TLX Class. | VisAWI-S Score | VisAWI-S Class. |
|---|---|---|---|---|---|
| MONSI | 20 | 17.79 / 100 | Moderate | – | – |
| Coco Apps | 22 | 10.99 / 100 | Moderate | 5.81 / 7 | Very Positive |
| LabForLapt | 21 | 21.75 / 100 | Moderate | 4.23 / 7 | Positive |
| Web Quest | 21 | 15.24 / 100 | Moderate | 5.42 / 7 | Positive |

As a detailed illustration, the "Web Quest" survey (n = 21) obtained a Raw NASA-TLX composite score of 15.24/100 (Moderate), with all six dimensions also classified as

Moderate (per-dimension means 13.57–17.38), and a VisAWI-S composite score of 5.42/7 (Positive), with Simplicity as the highest-scoring aspect (5.57) and Diversity as the lowest (5.29). The "LabForLapt" survey (n = 21) obtained the lowest VisAWI-S composite score among the four surveys (4.23/7, Positive), with per-aspect means ranging from 3.90 (Colorfulness) to 4.62.

#### 4.2.3 Results of the SUS, TAM, and A/B Testing Survey (1 Sample)

The fifth developer sample, the "Redesign of the SKKM Polinema Website UI/UX" survey (n=31), used the SUS, TAM, and A/B Testing modules that were already available on the system. This data is included as evidence that the pre-existing modules continued to run stably after the addition of the Raw NASA-TLX and VisAWI-S modules, not as primary evidence for the two new modules.

On the SUS module, the total score obtained was 82.02 out of 100, which falls into the Acceptable category on the Brooke scale [3], with the majority of respondents strongly disagreeing with the negatively worded items (item 2, "this system is complex": 45.16% strongly disagree). On the TAM module, the system produced three inter-construct regression coefficients, all of which were positive: PEU→PU of 0.57; PEU→ATU of 0.63; and PU→ATU of 0.82 showing positive associations consistent with the TAM framework. On the A/B Testing module, Design A was consistently chosen by the majority of respondents across all three tested pages (Homepage 67.7%, Dashboard 90.3%, Poinku 96.8%).

### 4.3 Discussion of Functional Verification Results

The functional verification results from the five developer teams in Subsection 4.2 show that all three verification criteria were met: all 5/5 teams successfully created a survey without technical obstacles, every survey obtained real respondents (115 respondents in total across 5 surveys), and every results dashboard displayed without errors including the KPI card, per-dimension charts, demographic data, and automatic narrative descriptions in accordance with Subsection 4.1.3.

The composite scores obtained across the four surveys varied both within a construct (Raw NASA-TLX: 10.99–21.75/100) and across classification bands (VisAWI-S: 4.23/7 Positive for LabForLapt versus 5.81/7 Very Positive for Coco Apps). Two observations follow. First, the dashboard returned differentiated values for products evaluated in different application contexts rather than a constant output; the present design, however, does not include a reference condition with a known difference in workload or aesthetics, so no claim about the instruments' discriminant sensitivity can be made from these data establishing that would require a controlled comparison and is set as follow-up work. Second, the automatic per-dimension classification and narrative descriptions (Subsection 4.1.3, point f) allowed the developer teams to locate the aspects requiring attention directly for example Colorfulness (3.90) and Diversity (4.19) in the LabForLapt VisAWI-S result without performing any additional manual statistical analysis.

The fifth survey (SKKM Polinema, n=31), with a SUS score of 82.02, demonstrates the stability of the existing SUS/TAM/A-B modules after the addition of the Raw NASA-TLX and VisAWI-S modules the additive modular enhancement did not disrupt the functionality of the prior modules. The entirely positive TAM regression coefficients in this survey are consistent with the positive associations between TAM constructs reported in the literature [5].

Demographics of respondents across the five surveys were consistently dominated by the 18-24 age group (71-90%) with a student background (87-91%). This demographic concentration reflects the user population of the UX evaluation system in an academic context and shows that the evaluation results from this platform are representative for that user segment; generalization to a more diverse professional population requires further testing with

broader demographic representation, as noted in Subsection 4.7.

***4.4 Alignment with the Underlying Instruments***

The Raw NASA-TLX implementation follows the unweighted formulation validated by Hart [8]. The Performance dimension uses the anchors Perfect (0) and Total Failure (100) in accordance with the original instrument [7], so the implementation is already consistent without requiring any additional transformation. Note that some implementations reverse this dimension so that higher values indicate better perceived performance; the present implementation follows the original anchor direction, so no additional transformation is applied and a higher Performance rating consistently signals worse perceived performance, as with the other five dimensions. The 0-100 slider range with a 5-point interval is consistent with the original administration of Hart and Staveland [7].

The VisAWI-S form consists of four items, one item per aspect, exactly matching the specification of the original instrument [16]. As affirmed by the VisAWI Manual, VisAWI-S "measures the general factor of website aesthetics" with one item per aspect [17]; the implementation therefore does not deviate from the original instrument's specification. The psychometric consequence of one item per aspect that a per-subscale Cronbach's $\alpha$ is undefined with one item per factor is an intrinsic characteristic of the short-form instrument, not a characteristic of the UIX-Probe implementation. Studies requiring detailed per-facet psychometrics should use the full 18-item VisAWI, and expansion to the full VisAWI is proposed as one direction for future development.

As for the SUS and TAM instruments available as modules on the UIX-Probe platform and used in the "Redesign of the SKKM Polinema Website UI/UX" survey as reported in Subsection 4.2.3 they follow the standard instruments without any modification to the scoring formula, as documented in the earlier research that built the multi-method foundation of this platform.

**4.4 Data Ethics, Privacy, and Discussion of the LLM Component**

The form component saves the in-progress response state to the browser's `localStorage` under the key `surveyData_{survey_id}_{user_id}`. This approach provides the benefit of recovery after a page reload, but raises privacy implications that need to be disclosed openly. The stored data includes mental-workload ratings and aesthetic perceptions, which are personal in nature. In line with the data-minimization principle in the GDPR framework [20] and Indonesia's Personal Data Protection Law [21], the implementation is designed so that: (a) the `localStorage` key is deleted immediately after the transaction commit succeeds fulfillment of this behavior is currently verified through manual exploratory testing and is set as the next priority for automation; (b) the form page displays a notice that the form state is stored locally; and (c) the platform's privacy policy contains an explanation of this data processing. Respondents participated voluntarily and were informed of the purpose of data collection before beginning the survey. The respondents' demographic data was collected with consent and is used only in aggregate form without individual identities.

Ethics and consent. Respondents were recruited independently by each developer team and participated voluntarily. Before beginning a survey, each respondent was shown a notice describing the purpose of the data collection, the categories of data collected (demographic attributes and instrument ratings), the local storage of in-progress responses in the browser, and the exclusively aggregate use of the results; participation proceeded only after the respondent gave consent. No directly identifying information was collected, and no individual-level result is reported in this paper. Institutional ethics review was not required for this category of study under the applicable policy at the authors' institution (Politeknik Negeri Malang).

### 4.5 Comparison with Similar Systems

A number of web-based UX evaluation platforms have developed well in both commercial and academic circles, including UsabilityHub / Lyssna [22], Maze [23], and Lookback [24]. Most of these platforms focus on task-based usability testing, session recording, or rapid preference testing rather than on aggregating a battery of validated psychometric questionnaire instruments into a single analytics dashboard; a detailed feature-by-feature comparison against each platform's current documentation is left as follow-up work rather than asserted here without independent verification. UIX-Probe, with its newly integrated modules, combines SUS, TAM, A/B testing, WCAG, NASA-TLX, and VisAWI-S within a single platform with an integrated per-dimension analytics dashboard.

### 4.6 Implementation Limitations

A number of technical and methodological limitations are openly noted. (a) Cache invalidation is not automatic when a new response comes in; a cache with a 2-hour TTL means the dashboard can display slightly stale data, and the proposed fix is to call `Cache::forget()` in `FormController::store()` after a successful commit. (b) The use of `method_id` hardcoded as constants 5 and 6; refactoring to a PHP enum with a single reference to the `methods` table is set as follow-up work. (c) The VisAWI-S implementation (4 items) already conforms to the original instrument, but users who need deeper per-facet analysis should wait for the full VisAWI module (18 items). (d) The functional verification in Subsections 4.2 and 4.3 verifies the system's functionality under real-world usage conditions, but is not a controlled user-acceptance study; a more formal acceptance test using standardized psychometric instruments is set as follow-up work. (e) The demographics of respondents across the five surveys were dominated by students aged 18-24, so generalization to a more diverse user population requires further research. (f) The dashboard's effectiveness in supporting design decisions is described based on the anatomy of its components, but has not yet been measured directly through a task-based study comparing the speed or accuracy of design decision-making with and without the dashboard. (g) The deployment volumes used in this study (20–31 respondents per survey) were not sized against a statistical-precision target; a future study making inferential claims about individual products' scores should instead size its sample using established continuous-metric guidance [25], [26].

## 5. CONCLUSION

This paper reports the design, implementation, and workflow of two new evaluation modules, Raw NASA-TLX and VisAWI-S, integrated as modular extensions into the UIX-Probe platform, with a focus on how the resulting analytics dashboard makes it easier for developers and interface designers to interpret evaluation results. The extension is realized through three database migrations, two Eloquent models, two controllers, two React form components, and two React results pages, together implementing survey creation, response capture, automatic score calculation, transactional persistence, aggregate analytics, and data export. The Raw NASA-TLX module calculates the unweighted mean of six dimensions on a 0-100 scale, classified into five workload categories (Table 1); the VisAWI-S module calculates the unweighted mean of four items on a 7-point Likert scale, classified into four visual-aesthetics categories (Table 2), consistent with the original instruments [16].

The analytics dashboard designed for both modules (Subsection 4.1.3) provides six main components: a KPI card with automatic classification (Table 1 and Table 2), an average-per-dimension chart, a per-question response-distribution chart, a per-respondent results table, demographic data, and an automatic narrative description with an Excel-export option for further analysis. The acceptability and correct operation of both

modules together with this dashboard were verified through real-world deployment by five independent developer teams, totaling 115 responses across five surveys. All verification criteria were met: 5/5 teams successfully created a survey, every survey obtained real respondents, and every dashboard rendered without error. The composite scores obtained across the surveys (Raw NASA-TLX: 10.99-21.75/100, all Moderate; VisAWI-S: 4.23-5.81/7, Positive to Very Positive) show that the dashboard returned differentiated output across products evaluated in different application contexts; a controlled study with a known reference difference would be required before any claim about the instruments' discriminant sensitivity could be made.

The contribution of this paper covers two aspects. First, the integrated operationalization of two constructs previously unavailable on UIX-Probe subjective mental workload and visual aesthetics into the existing multi-method platform, complete with calculation formulas, automatic score classification, and narrative descriptions. Second, the anatomy of an analytics dashboard specifically designed so that developers and interface designers can directly interpret results, from the single composite score down to the per-question response distribution, without requiring additional manual statistical analysis, with an Excel-export option for those who need further analysis. The functional-verification exercise with five developer teams under real-world deployment, with 115 total respondents and all three verification criteria fully met, provides empirical evidence that the system operates correctly, although a more formal acceptance test using standardized psychometric instruments is set as follow-up work, as noted in Subsection 4.7.

The main limitations include: (a) the functional verification exercise confirms the system's operation under real-world deployment but is not a controlled, formal acceptance test; (b) the demographics of respondents across the five surveys were dominated by students aged 18-24, so generalization requires caution; (c) the LLM recommendation component has not yet been evaluated systematically; (d) the dashboard's effectiveness in supporting design decisions has not yet been measured directly through a task-based study.

Four directions for future development are proposed. First, a task-based study measuring how quickly and accurately developers identify improvement priorities with the help of the dashboard versus manual Excel export. Second, a formal acceptance test using standardized psychometric instruments (such as SUS or TAM) with a larger, controlled sample, to complement the functionality evidence already obtained from the functional verification exercise. Third, replication of the functional verification exercise with more diverse demographic representation to measure the internal reliability of VisAWI-S and the convergent correlation of NASA-TLX. Fourth, a systematic evaluation of the LLM recommendation component through comparison with expert recommendations, as well as expansion to the full VisAWI module (18 items) to support per-facet analysis.

## REFERENCES

[1] International Organization for Standardization, Ergonomics of human-system interaction Part 210: Human-centred design for interactive systems, ISO 9241-210:2019, 2019.

[2] R. Hartson and P. Pyla, The UX Book: Agile UX Design for a Quality User Experience, 2nd ed. Burlington, MA, USA: Morgan Kaufmann, 2018.

[3] J. Brooke, “SUS: A 'quick and dirty' usability scale,” in Usability Evaluation in Industry, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189-194.

[4] J. R. Lewis, “The System Usability Scale: Past, present, and future,” Int. J. Hum.-Comput. Interact., vol. 34, no. 7, pp. 577-590, Mar. 2018, doi: 10.1080/10447318.2018.1455307.

[5] F. D. Davis, “Perceived usefulness, perceived ease of use, and user acceptance of information technology,” MIS Quart., vol. 13, no. 3, pp. 319-340, Sep. 1989, doi: 10.2307/249008.

[6] World Wide Web Consortium, Web Content Accessibility Guidelines (WCAG) 2.1, W3C Recommendation, 2018. [Online]. Available: https://www.w3.org/TR/WCAG21/
[7] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in Human Mental Workload, P. A. Hancock and N. Meshkati, Eds. Amsterdam, The Netherlands: Elsevier, 1988, pp. 139-183, doi: 10.1016/S0166-4115(08)62386-9.
[8] S. G. Hart, "NASA-Task Load Index (NASA-TLX): 20 years later," in Proc. Hum. Factors Ergon. Soc. Annu. Meeting, vol. 50, no. 9, 2006, pp. 904-908, doi: 10.1177/154193120605000909.
[9] S. Said et al., "Validation of the Raw National Aeronautics and Space Administration Task Load Index (NASA-TLX) questionnaire to assess perceived workload in patient monitoring tasks: Pooled analysis study using mixed models," J. Med. Internet Res., vol. 22, no. 9, p. e19472, Sep. 2020, doi: 10.2196/19472.
[10] A. Ramkumar et al., "Using GOMS and NASA-TLX to evaluate human-computer interaction process in interactive segmentation," Int. J. Hum.-Comput. Interact., vol. 33, no. 2, pp. 123-134, 2017, doi: 10.1080/10447318.2016.1220729.
[11] B. R. Lowndes et al., "NASA-TLX Assessment of surgeon workload variation across specialties," Ann. Surg., vol. 271, no. 4, pp. 686-692, Apr. 2020, doi: 10.1097/SLA.0000000000003160.
[12] N. Al Madi, S. Peng, and T. Rogers, "Assessing workload perception in introductory computer science projects using NASA-TLX," in Proc. 53rd ACM Tech. Symp. Comput. Sci. Edu. (SIGCSE), vol. 1, 2022, pp. 668-674, doi: 10.1145/3478431.3499406.
[13] F. N. Azizah, D. Nurwinata Rinaldi, and F. Wafiq Al-Muqaffa, "Assessing mental workload of automotive factory employees using NASA-TLX and RSME methods," J. Ilm. Tek. Ind., vol. 24, no. 1, pp. 107-116, Jun. 2025, doi: 10.23917/jiti.v24i1.8363.
[14] M. Moshagen and M. T. Thielsch, "Facets of visual aesthetics," Int. J. Hum.-Comput. Stud., vol. 68, no. 10, pp. 689-709, Oct. 2010, doi: 10.1016/j.ijhcs.2010.05.006.
[15] M. Thielsch and M. Moshagen, "Measuring visual aesthetics with the VisAWI" [originally published in German as "Erfassung visueller Ästhetik mit dem VisAWI"], in Usability Professionals. Stuttgart, Germany: German UPA, 2011, pp. 260-265.
[16] M. Moshagen and M. T. Thielsch, "A short version of the Visual Aesthetics of Websites Inventory," Behav. Inform. Technol., vol. 32, no. 12, pp. 1305-1311, Dec. 2013, doi: 10.1080/0144929X.2012.694910.
[17] M. Thielsch and M. Moshagen, VisAWI Manual (Visual Aesthetics of Websites Inventory) and the Short Form VisAWI-S. Münster, Germany: Univ. Münster, 2014.
[18] G. Hirschfeld and M. T. Thielsch, "Establishing meaningful cut points for online user ratings," Ergonomics, vol. 58, no. 2, pp. 310-320, Feb. 2015, doi: 10.1080/00140139.2014.965228.
[19] I. Sommerville, Software Engineering, 10th ed. Boston, MA, USA: Pearson, 2016.
[20] European Parliament and Council, Regulation (EU) 2016/679 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation), Official J. Eur. Union, L 119, pp. 1-88, 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679
[21] Republic of Indonesia, Law No. 27 of 2022 on Personal Data Protection [Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi], State Gazette of the Republic of Indonesia No. 196, 2022.
[22] Lyssna, "Lyssna: user research and testing platform (formerly UsabilityHub)," [Online]. Available: https://www.lyssna.com. [Accessed: Aug. 2026].
[23] Maze, "Maze: user research platform," [Online]. Available: https://maze.co. [Accessed: Aug. 2026].
[24] Lookback, "Lookback: user research platform," [Online]. Available: https://www.lookback.com. [Accessed: Aug. 2026].
[25] R. Budiu and C. Moran, "How many participants for quantitative usability studies: A summary of sample-size recommendations," Nielsen Norman Group, Aug. 2021. [Online]. Available: https://www.nngroup.com/articles/summary-quant-sample-sizes/
[26] J. Sauro and J. R. Lewis, Quantifying the User Experience: Practical Statistics for User Research, 2nd ed. Amsterdam, The Netherlands: Morgan Kaufmann, 2016.