An online A/B testing Decision Support System for web usability assessment based on a linguistic decision-making model

Noe Zermeño[a], Cristina Zuheros[b], Lucas Daniel Del Rosso Calache[c], Francisco Herrera[b], Rosana Montes[a,*]

[a]*Department of Software Engineering, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071, Granada, Spain.*
[b]*Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071, Granada, Spain.*
[c]*Sao Paulo State University (Unesp), School of Engineering, Campus of Sao Joao da Boa Vista, São Paulo, Brazil.*

Abstract

Attention is increasingly directed toward enhancing user satisfaction with user interfaces, spanning both mobile applications and websites. A fundamental aspect of human-machine interaction is the concept of software usability. To assess website usability, the A/B testing technique facilitates the comparison of data between two designs. Expanding the scope of tests to include the designs being evaluated, along with the involvement of both real and fictional users, presents a challenge for which few online tools offer support. First, we conceptualize the previous ideas as a linguistic decision-making problem for usability assessment named LDM4USE model. We explore several usability tests with people engaged in role-playing scenarios. Building upon the widely recognized System Usability Scale (SUS), we merge collective opinions to present results on the usability of each alternative website in an understandable manner. Secondly, we integrate LDM4USE in an A/B testing tool named USE-AB-testing DSS (Decision Support System).

*Corresponding author.
E-mail addresses: nzermeno@correo.ugr.es (N. Zermeño), czuheros@ugr.es (C. Zuheros), lucas.calache@unesp.br (L. D. D. R. Calache), herrera@decsai.ugr.es (F. Herrera), rosana@ugr.es (R. Montes).

Preprint submitted to Information Sciences *February 8, 2024*

Thirdly, we use a real case study to assess three Moodle platforms at the University of Guadalajara, Mexico. Conducting these usability tests is an intensive task that has been significantly streamlined with the support of this DSS tool.

Keywords: A/B testing, Usability assessment, role playing, SUS scale, Decision Support System, Linguistic 2-tuples, FAHP, TOPSIS

1. Introduction

With the increasing use of websites and online platforms, such as Learning Management Systems (LMS) in the educational context, there is a growing need to enhance the quality of software usability to ensure satisfactory User eXperiences (UX). Software engineers specializing in UX methodologies, design and evaluate interfaces using user-centered techniques, such as *Design for all*. Web accessibility has been in development for decades [1] and in this regard, systems engineers can follow many guidelines, including ISO standards (ISO9216-11 [2], ISO25000 [3]) and also WCAG 2.2. guidelines from the World Wide Web Consortium (W3C) [4]. Despite this, there is a lack of clear guidance on how to validate web usability compliance.

Web usability has been explored under several evaluation methods, such as expert-driven inquiry, inspection, and testing. However, none of these methods has become a standard since they integrate a multitude and variety of factors in addition to the end-user's own opinions. Furthermore, there is no software tool that assists the entire process, which can range from being as simple or as complex as desired—undertaken solely by experts, with experts and *personas* (fictitious), or with actual end-users (the most suitable but also the most expensive option). In this regard, a standardized and cost-effective solution is the application of the System Usability Scale (SUS) questionnaire [5]. Its application is recommended in several phases of software design to maintain a focus on user-centered design [6].

A/B testing is a method very used by marketing, communication, and design professionals. It assists to compare different versions of an interface to find out what works best. To our knowledge, there is no complete and comprehensive online tool that facilitates usability evaluation and incorporates end-users as a central part of the evaluation process itself.

To place the focus on the users requires methodologies closer to human reasoning, such as Computing with Words (CW) [7]. CW is a methodology which operates with the perceptions of people instead of numerical measures. In fact, this could bring flexibility in the interpretation of the results, because they are expressed in natural language and not by numbers. If we consider usability assessment as the comparison of two, three or more versions of the same site, or even different related sites (alternatives) with respect to a set of attributes (or criteria), it is possible to be considered as a decision making (DM) problem [8]. Several studies have proposed to evaluate usability by survey [9], and by using a Multi-Criteria Decision Making with a total of 31 sub-criteria [10]. Specially aligned with WCAG 2.0. recommendations, Agrawal et al. [11] proposed an usability-accessibility-based MCDM approach using the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), the Analytic Hierarchy Process (AHP), and TAW tool for the evaluation of several Indian airline websites. Although the present proposal is concordant in the use of a similar MCMD approach, the usability parameters used by the authors are merely analytical data of the software and they do not take into account desirable aspects for the user (memorability, satisfaction, ease of use, etc.). Though the use of AHP is a trend [12], we find more suitable to operate with Fuzzy AHP [13] to allow uncertainty in the assessments given by the moderator user.

This paper proposes an an online A/B testing Decision Support System (DSS) for web usability assessment based on a linguistic decision-making model. It is composed of two main elements, the decision making model and the DSS, as we describe bellow:

- A Multi Expert Multi Criteria Linguistic Decision-Making model for web USability Evaluation (called LDM4USE model) based on user perceptions and user-centered methodologies such as Design Thinking (DT) [14], which allows a significant cost reduction in software maintenance. To consider the qualitative aspects of usability assessments, we proposed to manage them through fuzzy linguistic information. Fuzzy AHP technique is used to solve the assignment of criteria weights, being the criteria a set of common approaches for usability testing. We choose the 2-tuple linguistic model [15] –which operates with linguistic variables– as the computational model, given that it copes with unbalance information and data aggregation. We follow the Fuzzy TOPSIS [16] steps to generate various rankings that can provide insights particularly in target user groups (e.g., users with special needs). Thus, to summarize:

– **LDM4USE uses people perceptions.** The opinions are obtained from two groups of users –UX experts and website end-users– and managed with the 2-tuple linguistic computational model [15]. A weighting is assigned to their opinions according to group membership. It acknowledges varying expertise levels, perspectives, and contributions, enabling more informed and balanced decision-making processes.

– **LDM4USE design is user-centered.** It relies on the Design Thinking paradigm for *universal design* by using techniques such as *personas* and *role playing*. Evaluating a website from the perspective of specific roles (e.g., visually impaired, elderly, stressed) helps to observe compliance with the *design for all* principle.

– **LDM4USE uses many tests.** The proposal combines custom and standardized usability tests with accessibility tests from a linguistic perspective. Results are reported in the same domain of significance, qualifying the website's usability with the Adjective SUS Scale [17].

– **LDM4USE is configurable**. It is a model that allows the application of as many tests as necessary. Additionally, is configurable with personalized usability tests because we incorporate the concept of *usability testing*, for which we can define the most appropriate number of tasks and find the most suitable task-estimation time.

- **USE-AB-testing DSS**. The overall scheme for decision-making solving (definition, data elicitation, and exploitation) can be assisted by a Decision Support System (DSS) [7]. The main contribution of a DSS is to provide users with tools, information, and analytical capabilities to make better decisions efficiently and effectively [18]. Thus, USE A/B testing DSS is a free online A/B testing tool for UX engineers that helps in the process involving the evaluation of one or several website interfaces, using one or several tests, and using one or several users (experts or end-users). They can set the project and get two separate URLs to share with UX experts and end-users for online data elicitation. Later, the tool generates a final report with rankings per role and a general ranking of alternatives.

The remainder of this paper is structured as follows. In Section 2, we introduce usability tests and representations that enable the resolution of a linguistic decision-making problem for website usability assessment. Section 3 introduces LDM4USE, designed and implemented as an A/B testing facility –named USE A/B testing DSS– for comparing alternatives and managing user-generated data. The case study is presented in Section 4, where we follow all procedures to decide which of the three Moodle platforms has the best usability for students and teachers. Finally, Section 5 provides our conclusions and outlines future work.

2. Preliminaries

Our proposal is highly attached to concepts associated with the evaluation of web sites, such as web accessibility (explained in Section 2.1) and web usability (defined in Section 2.2). In order to understand what kind of tests are applied to assess the usability and apply the DT paradigm, we explain the technique known as *rol-playing* in Section 2.3. Section 2.4 presents basic aspects of the 2-Tuple Linguistic Representation model that helps to understand LDM4USE. Finally in Section 2.5, we describe TOPSIS which is the algorithm that solves the exploitation phase of LDM4USE.

2.1. Web accessibility assessment

Web accessibility is the practice of ensuring access to information on websites, especially for people with disabilities, whether visual, hearing, physical, or cognitive. The purpose of web accessibility evaluation is to measure the possibility of being used by people with disabilities, but also by any other person, such is the paradigm of *design for all*.

Given that standards and recommendations for ensuring correct software accessibility have been established for years, especially in the context of websites, various tools now facilitate the automatic evaluation of accessibility. The World Wide Web Consortium (W3C) plays a key role in this standardization effort. Their website features a comprehensive list of 166 accessibility testing tools[1], although only a few can effectively monitor compliance with the recently published WCAG 2.2 standard [4], released in July 2023. The ultimate aim is to assess a website's accessibility and assign it an A, AA, or AAA label, typically displayed as an image in the footer of the site.

[1] W3C list of assessment tools https://www.w3.org/WAI/ER/tools/

One of the most used tool to evaluate web accessibility is WAVE[2], which is able to give a detailed report of error and warnings over the HTML page. It helps to keep doing changes in the HTML / CSS tags, reevaluate and improve the final accessibility label.

2.2. Usability assessment methods

At the past century, usability was understood as a user-friendly product which is a highly subjective concept very difficult to measure. From a practical point of view, usability is about the experience of an user and the fact of being able to operate a system in the minimum time possible, without neglecting aesthetics and site content. Classical approaches for usability evaluation are commented below.

- **Inspection**. A panel of experts plays an important role to measure the degree of usability of a system by testing the user interface. There are practical checklists such as Nielsen's heuristic [19] or a Cognitive Walk-through for a given task flow [20].

- **Inquiry**. This method focuses on data acquisition mainly by the observation of people during the processes of software usage. This can be done as: (1) Focus Groups, discussion with users and recording facilities [21], (2) Interviews, usually letting the user to think out loud [22], (3) Questionnaires specifically designed, (4) Activity Logs [23], that can be further explored even with Data Mining techniques [24].

- **Test**. This method collects information in real time in sessions called *Usability Tests* (UT), with participants that are real users who will use the developed software. Therefore, this information has a high degree of reliability [6]. In addition, this test can detect some omissions derived from the heuristic evaluations. If the UT is well defined by *the conductor*, and explained to *the participants* (number of task, expected starting and ending time), it is possible to run it through an online session, collecting afterwards the information with a Google form [25]. Other tests that fall into this category are: (1) Eye Tracking, a solution related to neuromarketing that tracks the eyes to know the point where the gaze is fixed. It can help to better understand what attracts the

[2]Wave Accessibility testing tool https://wave.webaim.org/

customer's attention; (2) Specific tests that applies to real people with disabilities and Assistive Technology enabled browsers.

- **Standardized questionnaires**. Some usability evaluation questionnaires have been developed and accredited [26] such as: (1) System Usability Scale (SUS), (2) Questionnaire for Users Interfaces Systems (QUIS) and (3) Web site Analysis and MeasureMent Inventory (WAMMI). It is also very frequent to make use of a single question for product / service satisfaction inquiry, called Net Promoter Score (NPS) [27]. These tests evaluate different criteria or dimensions of usability, such as: software performance, design, ease of use, user satisfaction, among others. Our model is able to linguistically incorporate the results from SUS and NPS questionnaires.

  1. SUS developed by John Brooke in 1996 [5, 28], is still quite frequent in the literature [29, 30, 31]. Main features are: 1) easy to fill, only 10 items with Likert scale, 2) free and multi language and, 3) applicable with various types of user interface such as websites, mobile, TVs. From the answers, a formula compute a number in 0-100 domain. More importantly SUS also applies to many linguistic scales as is shown in Figure 1 [17].
  2. NPS [27, 32] is used for service evaluation of companies and institutions. It comprises a single and easy question that fall to categorize the customer as: 1) Promoters, 2) Passive users or 3) Detractors. The user answers with a number between 0 and 10, but the NPS score is a value within the range $[-100, 100]$. There is a factor of conversion from NPS to SUS scale as can be seen in [33].

- **A/B testing**. This type of testing is used extensively in business marketing. For instance, it is used to improve performance metrics [34] such as conversion and click-through rates. A/B testing compares two versions of the same product or service in an attempt to identify which version or features are considered better by the user. The importance of this technique is to recognize the value of user feedback in order to design better user experiences [35]. In the UX field, A/B testing technique has been incorporated to compare two versions of the same site [36], for instance, to test satisfaction with dark vs light themes.

To conduct an A/B test, there are some payment solutions[3], and some simple calculators[4] or spreadsheets[5] but the usability evaluation area is open for comprehensive free online tool that facilitates the task of identifying the best response from a range of alternative responses. Thus is not just A or B compared, it could be more.

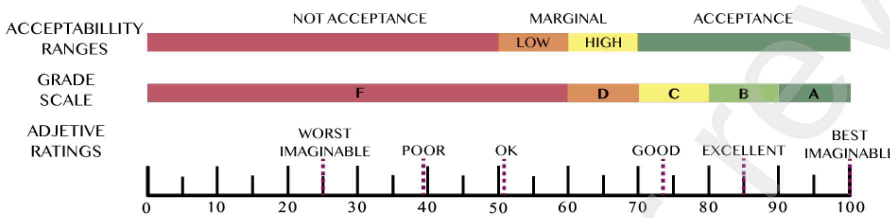


Figure 1: Several linguistic scales can be derived from the numerical SUS score. Source: [17]

2.3. Empathy and role-playing

Design Thinking (DT) is defined as "a human-centered innovation process that emphasizes observation, collaboration, rapid learning, visualization of ideas, rapid prototyping, and simultaneous business analysis" [14]. It is a methodology with great ability to understand people's needs through the establishment of well defined phases, and by applying many tools for conceptual analysis of user needs and the identification of software development requirements. Given its user-centered nature, this type of methodology focuses on the collection of user characteristics and needs.

We apply *role playing* in our model to let people to express some temporary situation in their lives such as a broken arm o even states of mood. By linking a role to the assessment, the UX expert who conducts the A/B testing can also apply several assessments for each of the defined *personas* (an archetype of a user that helps designers and developers empathize with people with especial needs [37]). For instance, they can play the role of a foreign student visiting the university website or empathize with a visually impaired person. Numerous possibilities exist for defining these roles, as illustrated in Figure 2 [38], which provides some examples.

[3]VWO https://vwo.com/

[4]https://www.kissmetrics.com/growth-tools/ab-significance-test/

[5]https://offers.hubspot.com/ab-testing-kit

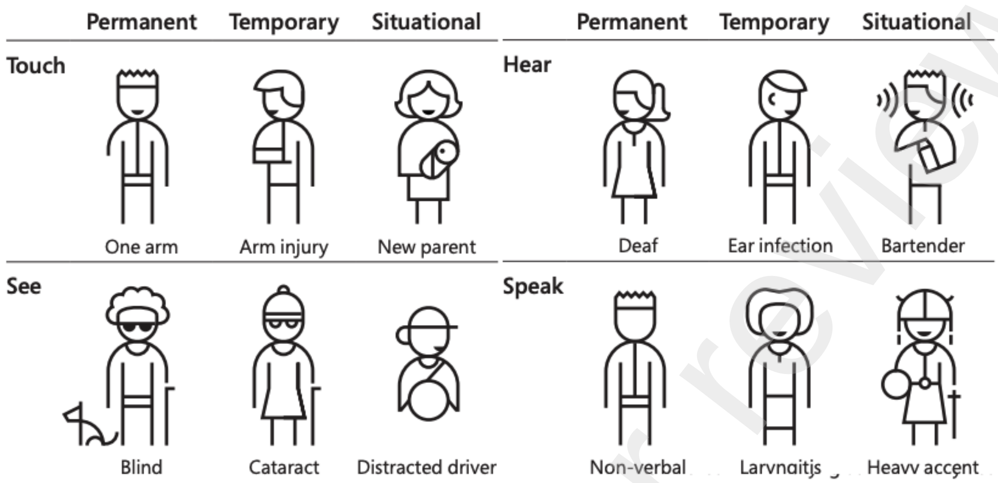


Figure 2: Decision makers can reflect about self situation or can roleplaying to empathize with real user needs. Source: [38]

2.4. A linguistic representation model for Decision Making

Decision Making (DM) models are essential to computationally solve decision situations in many areas such as healthcare, business and education. They require effective management of the expert evaluations to rank the alternatives with quality. There are multiple different representations to solve DM, among which the 2-Tuple Linguistic Representation has a great impact [39]. It handles uncertainty through linguistic terms and prevents the loss of information by means of a continuous domain. Rooted on the fuzzy set theory, Herrera *et al.* defines the functions Δ and Δ^{-1} to transform numerical values into 2-tuples and vice-versa.

In our case, for usability assessments where individuals have different needs, consensus-reaching processes –usually embedded in large group decision-making (LGDM) [40]– may not be sought. The focus lies more on understanding and accommodating diverse perspectives rather than striving for uniform agreement among stakeholders.

Proposition 1. *[41] Let $S^{g+1} = \{s_0, \ldots, s_g\}$ be a linguistic term set and $\beta \in [0, g]$ the result of an aggregation operation. The function Δ transforms* *a β value to an equivalent information 2-tuple by Equation 1.*

$$\begin{aligned} &\Delta : [0, g] \rightarrow S^{g+1} \times [-0.5, 0.5) \\ &\Delta(\beta) \;=\; (s_i, \alpha),\ \mathit{with} \left\{ \begin{array}{l} s_i \quad i = round(\beta), \\ \alpha = \beta - i \end{array} \right. \end{aligned} \tag{1}$$

where round$(\cdot)$ *is the usual round operation.*

Proposition 2. *[41] The inverse function Δ^{-1} transforms a 2-tuple to its equivalent numerical value $\beta \in [0, g]$ by Equation 2.*

$$\Delta^{-1}(s_i, \alpha) \;=\; i + \alpha = \beta \tag{2}$$

2.5. A ranking method for Decision Making

The Technique for Order of Preference by Similarity to Ideal Solution, known as TOPSIS [42], stands out as a multi-criteria method that facilitates the selection of the most optimal alternative. The main idea of TOPSIS is based on the assumption that the selected alternative is as close as possible from the positive ideal solution and contains the farthest distance from the negative ideal solution.

Let D be a normalized matrix composed of 2-tuple values (s_{ij}, α_{ij}), where $i = 1, \ldots, n$ represent the evaluated alternatives and $j = 1, \ldots, m$ the associated criteria. We assume the weights of the criteria are equal. The TOPSIS procedure [43] determines the positive ideal solution A^+ and negative ideal solution A^- as vectors of 2-tuples formed by:

$$A^+ = [(r_1^+, \alpha_1^+), (r_2^+, \alpha_2^+), \ldots, (r_m^+, \alpha_m^+)] \tag{3}$$

$$A^- = [(r_1^-, \alpha_1^-), (r_2^-, \alpha_2^-), \ldots, (r_m^-, \alpha_m^-)] \tag{4}$$

where

$$(r_j^+, \alpha_j^+) = \left\{ \underset{i}{max}\{(s_{ij}, \alpha_{ij})\} \right\}, j = 1, \ldots, m \tag{5}$$

$$(r_j^-, \alpha_j^-) = \left\{ \underset{i}{min}\{(s_{ij}, \alpha_{ij})\} \right\}, j = 1, \ldots, m \tag{6}$$

Next, the separation measures D_i^{+l} and D_i^{-l} of each alternative from the positive ideal solution and the negative ideal solution are computed based on the Euclidean distance:

$$D_i^+ = \Delta \sqrt{\sum_{j=1}^{m} \left(\Delta^{-1}(s_{ij}, \alpha_{ij}) - \Delta^{-1}(r_j^+, \alpha_j^+) \right)^2} \tag{7}$$

$$D_i^- = \Delta\sqrt{\sum_{j=1}^{m}\left(\Delta^{-1}(s_{ij},\alpha_{ij}) - \Delta^{-1}(r_j^-,\alpha_j^-)\right)^2} \tag{8}$$

The coefficient of relative closeness to the ideal solutions of each alternative A_i with respect to the positive ideal solution A^{+l} is then calculated:

$$RC_i = \Delta\left(\frac{\Delta^{-1}(D_i^-)}{\Delta^{-1}(D_i^+) + \Delta^{-1}(D_i^-)}\right) \tag{9}$$

where $0 \leq \Delta^{-1}(RC_i) \leq 1$.

3. A Multi Expert Multi Criteria Linguistic Decision-Making model for usability evaluation

LDM4USE is a model for usability evaluation through a Multi Expert Multi Criteria LDM problem that merges several sources of information. It incorporates the use of standardized tests and takes into account the approximation to human reasoning by means of the linguistic transformation of users' judgements. It is implemented as USE A/B testing DDS that is used as online tool for A/B testing. Indeed, our LDM4USE problem resembles A/B testing, with the distinct advantage of maintaining the same degree of complexity when evaluating more than two alternatives. This is in contrast to the classical A/B testing technique.

Section 3.1 describes the steps for solving a linguistic decision making problem. Later, we follow each step: problem definition at Section 3.2, elicitation and data gathering at Section 3.3, aggregation at Section 3.4 and exploitation at Section 3.5.

3.1. General description and flowchart of LDM4USE model

Web usability evaluation from a user-centered approach considers mainly satisfaction with the use of the site, as well as data gathered from the UX expert or answered by the end-user. Given the extensive use of the Adjective SUS scale [17], we can rely in linguistic information and the paradigm of Linguistic Decision Making [44]. For that reason, we adapt to our case the following steps to have a solving schema.

- Problem structuring.Define the alternatives, criteria, and their respective weights. In our case, include the setting of user weights, roles, and their relative weights.

- Elicitation of user information. The decision makers, playing particular roles, evaluate the alternatives based on the test criteria. Lets define S^g as a term set of g linguistic elements $\{s_0, \ldots, s_{g-1}\}$. A single linguistic opinion could be noted as $s_i \in S^g$. Under the 2-tuple representation, the same term is noted as $(s_i, 0)$.

- Unification. All assessments are brought into a common linguistic opinion space.

- Collective aggregation. In our case, a 2-tuple operator (arithmetic mean, harmonic mean, minimum value per criterion, etc.) can compute the linguistic opinion of all the evaluators on the same alternative on each criteria [15]. Subsequently, with a second aggregation by criteria, an overall 2-tuple assessment per alternative is obtained.

- Exploitation. In order to rank the alternatives from the best to the worst assessed, we apply Fuzzy TOPSIS algorithm [43].

- Retranslation. For convenience, in addition to having a ranking, it is possible to present linguistic output information on a specific scale. In our case, we perform this step before generating the report using the Adjective SUS scale.

Figure 3 illustrates the twelve steps to solve the proposed usability linguistic evaluation model. Further details about each step are provided below.

3.2. Phase 1. Problem structuring

The set up of the model determines the essentials elements required to carry out the web usability evaluation. Particularly, we establish a moderator figure who builds the following elements: (1) alternatives, which are websites or online platforms to evaluate, (2) criteria, which are instruments of accessibility and usability tests, (3) decision makers whom we call users, consisting of both end-users and experts, who evaluate the alternatives based on the criteria, 4) roles, moderator could select all or the subset of the roles shown in Figure 2, and 5) weights, the model can incorporate criteria weights, user weights and roles weights.

Step 1. Definition of the alternative set.
We define a set of alternatives $A = \{A_1, \ldots, A_n\}$ as websites, website

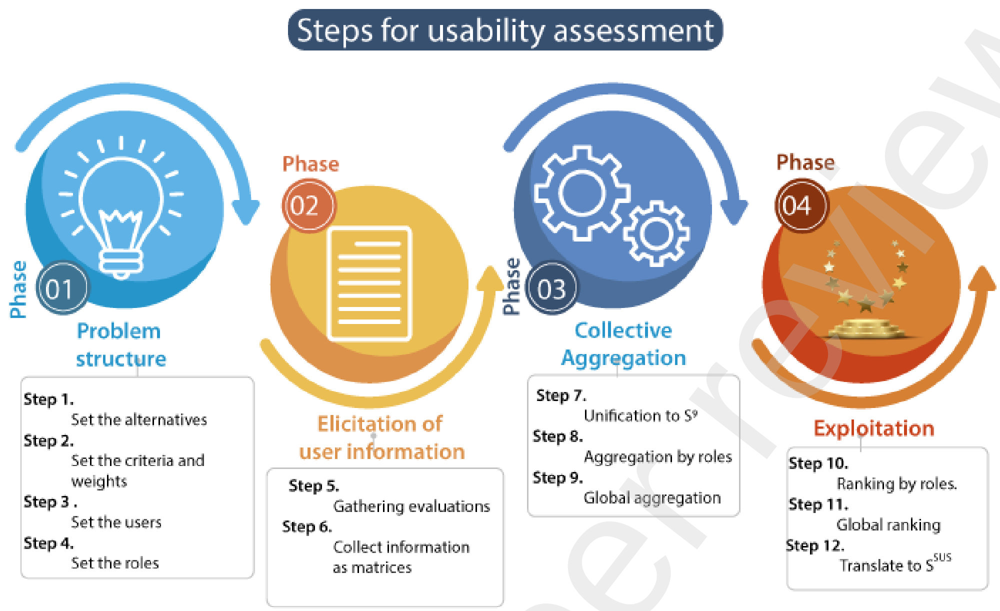


Figure 3: Flowchart of the Multi Expert Multi Criteria Linguistic Decision-Making model for web USability Evaluation.

versions, web pages, or online platforms that have similar objectives. We refer to each alternative by $A_i, (i = 1, \dots, n)$.

Step 2. Definition of the criteria set.

A set of criteria $C = \{C_1, \dots, C_m\}$ to be evaluated is defined, where each one is referenced by $C_j, (j = 1, \dots, m)$. In this proposal, we set $m = 4$. With our software, the *moderator* (usually the site developer or designer) can create an A/B testing with m or a subset of tests. Depending on the requirements of the LDM4USE problem, different importance of criteria may be considered. In the following, we describe the method to obtain the vector of criteria weights, which we denote by $WC' = \{WC'_1, \dots, WC'_m\}$. This vector is normalized by generating the criteria weights normalized vector $WC = \{WC_1, \dots, WC_m\}$ which verifies $\sum_{j=1}^{m} WC_j = 1$. The core tests are:

- $C_1 \cong$ SUS is filled with 10 items in Likert scale.
- $C_2 \cong$ NPS is filled with a single number between 0 and 10.
- $C_3 \cong$ UT is filled for each task with a boolean *task-done*, *task-time* in seconds, and a linguistic term for satisfaction doing the task. This term

is within $S^5 = \{$Unsatisfying, A little satisfying, Indifferent, Satisfying, Very satisfying$\}$.

- $C_4 \cong$ ACC is filled with the report value for Accessibility. This term is in $S^3 = \{A, AA, AAA\}$.

To precisely determine the criteria weights, we suggest applying the Fuzzy Extended Analytic Hierarchy Process (FAHP) [45] method. This method facilitates organizing criteria into a hierarchy and enables the detection of inconsistencies through a consistency analysis between judgments.

- **Step 2.1. Obtain pairwise judgments regarding the importance of criteria**. From moderator we collect the pairwise judgements of decision-makers to derive priority scales regarding criteria. We complete the CP matrix with the Triangular Fuzzy Numbers (TFN) as presented in Equation 10.

  Moderator gives answers from the set $S^5_{CP} = \{$*Equally important*, *Moderately important*, *Very important*, *Strongly important*, *Absolute*$\}$, where each linguistic term has an associated membership function $(low, medium, upper) = (l, m, u)$. A $m \times m$ matrix of Criteria Preferences (CP) is created from the membership functions of the linguistic ratings provided by the moderator.

$$CP_{j,j'} = (l_{j,j'}, m_{j,j'}, u_{j,j'}), \forall j \leq j'; \quad j, j' = 1, \ldots, m \tag{10}$$

  and their values are completed by means of Equation 11:

$$CP_{j',j} = (l_{j',j}, m_{j',j}, u_{j',j}) = \left(\frac{1}{u_{j',j}}, \frac{1}{m_{j',j}}, \frac{1}{l_{j',j}}\right), \forall j > j'. \tag{11}$$

- **Step 2.2. Compute fuzzy synthetic extension**.

  The fuzzy synthetic extension for criterion C_j is calculated using Equations 12-14:

$$s_j = (l_j, m_j, u_j) = \sum_{j'=1}^{m} M^{j'}_{C_j} \otimes [\sum_{j=1}^{m} \sum_{j'=1}^{m} M^{j'}_{C_j}]^{-1}, \forall j = 1, \ldots, m \tag{12}$$

being

$$\sum_{j'=1}^{m} M_{C_j}^{j'} = \left(\sum_{j'=1}^{m} l_{j,j'}, \sum_{j'=1}^{m} m_{j,j'}, \sum_{j'=1}^{m} u_{j,j'}\right), \forall j = 1, \ldots, m, \tag{13}$$

$$[\sum_{j=1}^{m}\sum_{j'=1}^{m} M_{C_j}^{j'}]^{-1} = \left(\frac{1}{\sum_{j=1}^{m}\sum_{j'=1}^{m} u_{j,j'}}, \frac{1}{\sum_{j'=1}^{m}\sum_{j'=1}^{m} m_{j,j'}}, \frac{1}{\sum_{j=1}^{m}\sum_{j'=1}^{m} l_{j,j'}}\right). \tag{14}$$

- **Step 2.3. Degree of possibility**.
  We obtain the degrees of possibility of the elements s_j and $s_{j'}$, $\forall j, j' = 1, \ldots, m$, $j \neq j'$ using Equation 15:

$$V(s_j \geq s_{j'}) = \begin{cases} 1, & if\ s_j \geq s_{j'} \\ 0, & if\ l_{j'} \geq u_j \\ \dfrac{l_{j'} - u_j}{(m_j - u_j) - m_{j'} - l_{j'}}, & in\ other\ case \end{cases} \tag{15}$$

- **Step 2.4. Obtaining the vectors of weights for the set C**.
  Finally, we compute the weight of each criterion $C_j, (j = 1, \ldots, m)$ by means of Equation 16:

$$WC'_j = min[V(s_j \geq s_{j'})], \forall j, j' = 1, \ldots, m, j \neq j', \tag{16}$$

  and we normalize them, obtaining the final weight for each criterion $C_j, (j = 1, \ldots, m)$ by applying Equation 17:

$$WC_j = \frac{WC'_j}{\sum_{j=1}^{m} WC'_j}. \tag{17}$$

Step 3. Definition of the user set.

Two groups of users are considered, named experts and end-users. Let $E = \{E_1, \ldots, E_p\}$ be a set of experts with knowledge in some area of technology, interfaces, or user experience, where p is the total number of experts.

Let $D = \{D_1, \ldots, D_q\}$ be a set of end-users, where q is the total number of non-expert users. Therefore, the set of users $U = E \cup D$ is the union of experts and end-users, and each user is referenced by $U_k, (k = 1, \ldots, u)$ where $u = p+q$. Users have associated a weight corresponding of their group: $WE \in [0,1]$ if their are experts, or $WD \in [0,1]$ for end-users. Both values are set directly by the moderator. Then, vector $WU = \{WU_1, \ldots, WU_u\}$ is filled with values according to the decision maker's membership in one of two groups. For example, if we have one expert and two end users this vector is $WU = \{WE, WD, WD\}$. Note that we do not impose that $WE + WD = 1$ but we have to compute the normalization of WU.

Step 4. Definition of the set of roles.

The model proposes a methodology focused on the end user, therefore, it relies on Design Thinking with the role-playing technique (see section 3.2) to capture the end-user's needs. Let $R = \{R_1, \ldots, R_r\}$ be the set of roles determined by the moderator where the possible roles are $R_l, (l = 1, \ldots, r)$. Each user plays at least one role in which they evaluate all alternatives, and each role has different importance according to the requirements of the problem. Let $WR' = \{WR'_1, \ldots, WR'_r\}$ be the vector of weights associated to the roles set directly by the moderator figure. This vector is normalized by obtaining the vector of role-playing weights $WR = \{WR_1, \ldots, WR_r\}$ that verifies $\sum_{l=1}^{r} WR_l = 1$.

3.3. Phase 2. Elicitation of user information

All the user evaluate criteria C_1, C_2, C_3 (common for U) and C_4 (only for E) in a particular way according to the test to which it refers. And each single assessment is attached to the role selected. Each user U_k, playing a role R_l, provides assessments of each alternative A_i and according to criteria C_j. Overall, it configures the Individual Decision matrix $ID^{k,l} = (ID_{ij}^{k,l})_{n \times m}$. Following, we explain in detail steps 5 and 6 for each instrument or criterion to give a clear procedure for constructing the individual decision matrices, respectively.

Step 5. Gathering user evaluations.

The user U_k playing the role R_l evaluates an alternative A_i by performing one or more tests C to qualify the usability of A_i. The information provided by each possible test is shown below:

- $C_1 \cong$ **System Usability Scale (SUS)**. The user answers ten questions following a Likert scale with five response options for each. Odd questions have a positive connotation, while even questions have a negative tone. For each alternative A_i, we denote the ten responses provided by the user U_k playing the role R_l as $x_h^{k,l,i}, h = 1, \ldots, 10$. We obtain the SUS score of user U_k playing the role R_l for each alternative $A_i, (i = 1, \ldots, n)$, by:

$$SUS\_score_i^{k,l} = 2.5 \times \sum_{h=1}^{5} [(x_{2h-1}^{k,l,i} - 1) + (5 - x_{2h}^{k,l,i})], \forall i = 1, \ldots, n \quad (18)$$

  For each user U_k playing the role R_l, the value $SUS\_score_i^{k,l} \in [0, 100]$ is available for each alternative $A_i, (i = 1, \ldots, n)$. This implies that we have for criterion C_1 a linguistic 2-tuple in the Adjective SUS Scale. Detailed computations are given in Step 9. at Section 3.4.

- $C_2 \cong$ **Net Promoter Score (NPS)**. The user faces a single question known as Likelihood to Recommend (LTR): *How willing are you to recommend the website represented by A_i?*. The answer must be an integer value belonging to the interval $[0, 10]$ so that values close to 0 represent that you would not recommend the website while values closer to 10 represent that you would recommend it. Thus, the user's LTR score U_k is obtained by playing the role R_l for each alternative $A_i, (i = 1, \ldots, n)$ ($NPS\_LTR\_score_i^{k,l}$) directly through the answer to the previous question. Users who answer with the values 10 or 9 are known as promoters, those who answer with 8 or 7 are called passive, and if they answer with another value they are known as detractors. In a complementary way, the NPS value associated with the evaluation of all users can be obtained as the percentage of promoters minus the percentage of detractors. It is frequent to get this computation done with online support[6].

  In summary, for each user U_k playing the role R_l, there is a value $NPS\_LTR\_score_i^{k,l} \in [0, 10]$ for each alternative $A_i, (i = 1, \ldots, n)$ that allows us to evaluate the criterion C_2 with a linguistic 2-tuple in the

[6]NPS calculator https://npscalculator.com/

Adjective SUS Scale. Detailed computations are provided in Step 9. of Section 3.4.

- $C_3 \cong$ **Usability Test (UT)**. In this test, the end-user has to answer as many questions as the moderator figure poses by setting an UT of d tasks to be performed. These d tasks define our usability test $UT = \{q_1, \ldots, q_d\}$. The user U_k's responses playing the role R_l for alternative $A_i, (i = 1, \ldots, n)$ to each task $q_v, v = 1, \ldots, d$ (provided with a UT questionnaire), are the following three measures:
  1. Efficiency ($Efficiency_i^{k,l}(q_v)$). This measure establishes whether the user has managed to perform the requested task in an adequate amount of time. The user tracks the time expended ($time_i^{k,l}(q_v)$) and our system compares it with the moderator's estimate for maximum time ($MaxTime(q_v)$). This measure can take two values: 1 if the user complete q_v under the estimated time, *i.e.*, if $time_i^{k,l}(q_v) \leq MaxTime(q_v)$ and 0 otherwise.
  2. Success ($Success_i^{k,l}(q_v)$). The user indicates whether they were successful or unsuccessful in performing the task. This measure can take two values: 1 if successful and 0 if unsuccessful.
  3. Satisfaction ($Satisfaction_i^{k,l}(q_v)$). The user indicates the feeling they experienced while solving the task. It is expressed by one adjective out of five possible ones: *unsatisfying*, *a little satisfying*, *indifferent*, *satisfying*, and *very satisfying* which correspond to the five linguistic terms of $S^5 = \{s_0, s_1, s_2, s_3, s_4\}$.

We compute the success, efficiency, and satisfaction measures for each question based on all user's responses in order to uncover those tasks that are remarkably complex for them. Furthermore, we calculate these three metrics for each possible role played to identify if there are tasks that are more difficult for a specific type of user profile. From WU, we should get r weighted vectors WU^l with the normalized weights of all the users playing the role $R_l, (l = 1, \ldots, r)$. Then, we compute the success, efficiency, and satisfaction associated to the task $q_v, (v = 1, \ldots, d)$ and the given role R_l using following equations:

$$Efficiency_i^l(q_v) = 100 \times \frac{\sum_{k=1}^{u^l} Efficiency_i^{k,l}(q_v)}{u^l} \tag{19}$$

$$Success_i^l(q_v) = 100 \times \frac{\sum_{k=1}^{u^l} Success_i^{k,l}(q_v)}{u^l} \tag{20}$$

$$Satisfaction_i^l(q_v) = \Delta\left(\sum_{k=1}^{u^l} \Delta^{-1}(Satisfaction_i^{k,l}(q_v), 0) \times WU_k^l\right) \tag{21}$$

where u^l is the number of users playing the role R_l.

The previous success and efficiency metrics are percentages while the satisfaction metric is a 2-tuple linguistic value. All this information is collected in the usability report that is complementary to the rankings solutions.

Thus, for each user U_k playing the role R_l and for each task q_v, three values are available $Success\_score_i^{k,l}(q_v) \in [0,1]$, $Efficiency\_score_i^{k,l}(q_v) \in [0,1]$, and $Satisfaction\_score_i^{k,l}(q_v) \in S^5$. For each alternative $A_i, (i = 1, \ldots, n)$ and after aggregation, we can evaluate criterion C_3 according to the overall satisfaction (success and efficiency will pose secondary metrics for the report) as it has the possibility of be expressed as a linguistic 2-tuple in the Adjective SUS Scale.

- $C_4 \cong$ **Accessibility (ACC)**. This test should be only performed by expert users. First, experts test alternatives via WAVE online tool that list errors and warnings on possible areas for improvement of the website or alternative. Given this report, the expert user U_k playing the role of R_l assigns a rating ($Acc\_score_i^{k,l}$) for each alternative $A_i, (i = 1, \ldots, n)$. This score measures the accessibility of the alternative, so it corresponds to the A, AA, and AAA conformance criteria with current Web Content Accessibility Guidelines[7]. The A label indicates the lowest level, while the AAA label indicates the highest level and therefore the highest quality.

  In conclusion, for each expert user U_k playing the role R_l, a value $Acc\_score_i^{k,l} \in S^3 = \{A, AA, AAA\}$ is available for each alternative $A_i, i = 1, \ldots, n$. According to the unification Step 6. at Section 3.3, it is our aim to evaluate the criterion C_4 with a linguistic 2-tuple in the Adjective SUS Scale.

[7] WCAG 2.2 https://www.w3.org/TR/WCAG22/

Step 6. Construction of the individual decision matrices.

The results of the tests previously performed by the users present very varied information. We consider linguistic models in order to interpret all the information together. For this purpose, this phase builds a matrix for each user playing a role that collects the results of all the performed tests, that is, the evaluations they provide, represented by a linguistic approach. For each user U_k playing the role R_l the individual decision matrix $ID^{k,l} = (ID_{ij}^{k,l})_{n\times m}$ is constructed. Its elements correspond to the linguistic information generated from the original user evaluations, attending to each criterion $C_j, j = 1, \ldots, m$ as follows:

- $C_1 \cong SUS$. The $SUS\_score_i^{k,l}, (i = 1, \ldots, n)$ scores of user U_k playing the role R_l are available. These scores are values from 0 to 100. If that user with that role does not evaluate alternative A_i for this criterion, $SUS\_score_i^{k,l} = \{\emptyset\}$.

  Bangor [17] proposes an adjective-based ranking within the SUS scale. Specifically, Figure 1 shows the SUS score equivalence in the interval $[0, 100]$ across a set of unbalanced terms. Next we explain how this equivalence is carried out. Let define the Adjective SUS as $\{s_0^{sus}, s_1^{sus}, s_2^{sus}, s_3^{sus}, s_4^{sus}, s_5^{sus}, s_6^{sus}\}$ be the set of seven unbalanced linguistic terms that we note as $S_{SUS} = \{$*None*, *Worst Imaginable*, *Poor*, *Ok*, *Good*, *Excellent*, *Best Imaginable*$\} = \{N, WI, P, O, G, E, BI\}$.

  To manage this term set we consider jointly the 2-tuple representation and the hierarchical linguistic structures. To begin with, we need to choose a suitable hierarchical linguistic structure and assign the associated semantics to each term using the different levels of the hierarchy. Thus S_{SUS} scale is constructed through two steps:

  1. Define a linguistic hierarchy $LH = \cup_t l(t, n(t))$. Each level of the hierarchy represents $S^{n(t)}$ and is denoted as $l(t, n(t))$, where t denotes the level of the hierarchy and $n(t)$ express the granularity of the set of terms in that level, *i.e.*, the number of elements available to it. We set level 1 as $l(1, 3)$ to partition the scale from the center and generate the next level as $l(t + 1, 2n(t) - 1)$. Therefore, the second level of the hierarchy is $l(2, 5)$. We set a third level $l(3, 9)$ in order to adapt all terms of S_{SUS}. In other words, at level $t = 1$ of the hierarchy we have $n(1) = 3$, at level

$t = 2$ we have $n(2) = 5$, and at level $t = 3$ we have $n(3) = 9$. This leads to establish the hierarchy $LH = S^3 \cup S^5 \cup S^9$.

2. Represent the unbalanced terms of S_{SUS} in LH. Applying the procedure raised in [46], we obtain that S_{SUS} is represented in LH using the linguistic labels of levels 2 and 3 of the hierarchy as shown in Figure 4. The linguistic terms of S_{SUS} belong to different levels of the LH hierarchy. The semantic representation of those terms is shown in Figure 5 and corresponds to $N \leftarrow s^5_0$; $WI \leftarrow \overline{s^5_1} \cup \underline{s^9_2}$; $P \leftarrow s^9_3$; $OK \leftarrow \overline{s^9_4} \cup \underline{s^5_2}$; $G \leftarrow \overline{s^5_3} \cup \underline{s^9_6}$; $E \leftarrow s^9_7$; $BI \leftarrow s^9_8$.

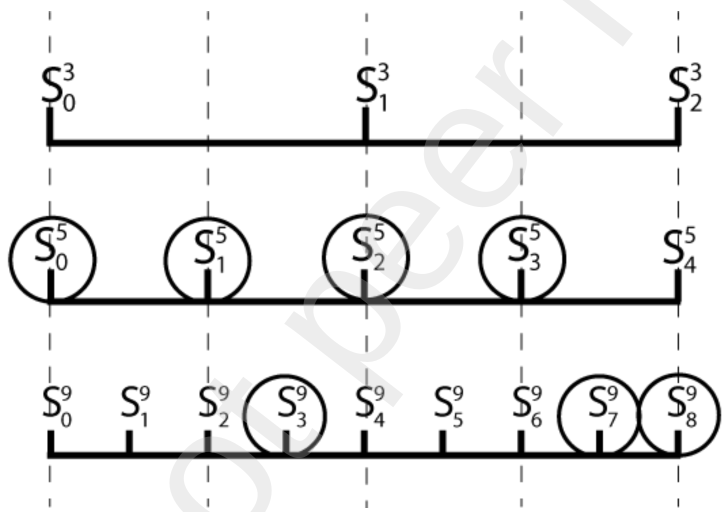


Figure 4: S_{SUS} term set according the hierarchy $LH = S^3 \cup S^5 \cup S^9$ are marked with circles.

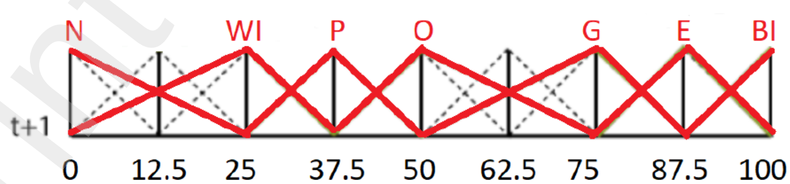


Figure 5: Semantic representation of S_{SUS} scale.

In order to work comfortably with this information, we define the set of Term Components (TC) associated with each linguistic term of S_{SUS} as the set of terms belonging to some level of the hierarchy that make up the semantics of the term in question as follows:

- $TC(s_0^{sus}) = \{s_0^5\}$
- $TC(s_1^{sus}) = \{s_1^5, s_2^9\}$
- $TC(s_2^{sus}) = \{s_3^9\}$
- $TC(s_3^{sus}) = \{s_4^9, s_2^5\}$
- $TC(s_4^{sus}) = \{s_3^5, s_6^9\}$
- $TC(s_5^{sus}) = \{s_7^9\}$
- $TC(s_6^{sus}) = \{s_8^9\}$

Thus far we have established the hierarchical structure associated with S_{SUS} as well as the semantics associated with each of its terms. It is time to transform the $SUS\_score_i^{k,l}, i = 1, \ldots, n$ scores of the users to the S_{SUS} scale. Next, Definition 1 presents the function that allows transforming SUS score values belonging to the interval $[0, 100]$ to the S_{SUS} scale:

Definition 1. *Let Score* $\in [0, 100]$ *be the score obtained after taking a SUS test (typically online calculator are available*[8]*). Let be* $S_{SUS} = \{s_i^{sus}; i = 0, \ldots, 6\}$ *the unbalanced SUS linguistic scale. Let* $LH = S^3 \cup S^5 \cup S^9$ *be the linguistic hierarchy associated with* S_{SUS} *such that* $S'^{n(t)} = \{s_0^{n(t)}, \ldots, s_{n(t)-1}^{n(t)}\}$, $t = 1, 2, 3$. *We obtain the 2-tuple belonging to the t level of LH associated with the Score as:*

$$(s', \alpha') = \Delta\left(\frac{(n(t)-1)Score}{100}\right) \tag{22}$$

where $t = 2$ *if Score* $\in [0, 25] \cup [50, 75]$ *and* $t = 3$ *if Score* $\in (25, 50) \cup (75, 100]$.

We define a transformation function for SUS (TF_{SUS}) *that associates the SUS score with its respective unbalanced linguistic 2-tuple as:*

[8]SUS calculator https://uiuxtrend.com/sus-calculator/

$$\begin{aligned} TF_{SUS} :&[0,100] \rightarrow (S_{SUS} \times [-0.5, 0.5)) \\ TF_{SUS}(Score) &= (s_i^{sus}, \alpha') \ \mid \ s' \in TC(s_i^{sus}) \end{aligned} \tag{23}$$

Example 1. *Let be the calculated value for C_1, $Score = 53$. This gives $t = 2$ (since $Score \in [50, 75]$) and, therefore, the 2-tuple associated with the 2 level of the hierarchy is $\Delta\left(\frac{4\times 53}{100}\right) = (s_2^5, 0.12)$. After that the associated unbalanced linguistic 2-tuple is $TF_{SUS}(53) = (s_3^{sus}, 0.12)$, that we understand as usability=$(OK, 0.12)$.*

We transform the values $SUS\_score_i^{k,l}, i = 1, \ldots, n$ given by user U_k playing the role R_l, into linguistic 2-tuples associated with the scale S_{SUS} by applying the Definition 1 and compile them into the $ID^{k,l}$ matrices. Specifically, for each user U_k playing the role R_l we fill in the first column ($j = 1$) of its associated matrix $ID^{k,l}$ as follows:

$$ID_{i1}^{k,l} = TF_{SUS}(SUS\_score_i^{k,l}) \ \ \forall i = 1, \ldots, n \tag{24}$$

- $C2 \cong NPS$. NPS values $NPS\_LTR\_score_i^{k,l}$ of user U_k playing the role R_l are available. These scores are values from 0 to 10. We consider $NPS\_LTR\_score_i^{k,l} = \{\emptyset\}$ when a user with that role does not evaluate alternative A_i with this test.

  In order to collect these values in the $ID^{k,l}$ matrix, we transform the non-empty scores, i.e., the LTR values obtained through the NPS test, to the S_{SUS} scale. To do so, we perform two steps:

  1. Transform NPS values $NPS\_LTR\_score_i^{k,l}$ to $NPS\_SUS\_score_i^{k,l}$ values. Sauro *et al.* [33] establishes a relationship between the values provided by users to the question posed in the NPS test (LTR values) and SUS values. The first proposed approach provides the equation $LTR = SUS/10$ that allows predicting the LTR value through a SUS value in a very simple way. Subsequently, through a study with more than 2000 users, the authors were able to obtain a better approximation of the regression equation that relates both values as $LTR = 1.33 + 0.08(SUS)$. We use the latter approximation since it is more accurate. Therefore, a correspondence is established between the scores obtained after answering the LTR question of the NPS test and a SUS test by:

$$NPS\_SUS\_score_i^{k,l} = \frac{NPS\_LTR\_score_i^{k,l} - 1.33}{0.08} \quad (25)$$

If the value $NPS\_SUS\_score_i^{k,l}$ is negative, then it is reset to 0. If that value is greater than 100, we set it to 100.

2. Transform the NPS values $NPS\_SUS\_score_i^{k,l}$ to linguistic representations S_{SUS} and collect them in the $ID^{k,l}$ matrices. We establish a correspondence between the values obtained in the NPS test represented as scores of a SUS test to the linguistic scale S_{SUS} by means of the Definition 1. We transform the scores $NPS\_SUS\_score_i^{k,l}, i = 1, \ldots, n$ of user U_k playing the role R_l into 2-tuples linguistic associated with the S_{SUS} scale and compile them into $ID^{k,l}$ matrices. Specifically, for each user U_k playing the role R_l we complete the second column ($j = 2$) of its associated $ID^{k,l}$ matrix as follows:

$$ID_{i2}^{k,l} = TF_{SUS}(NPS\_SUS\_score_i^{k,l}) \ \ \forall i = 1, \ldots, n \quad (26)$$

- $C3 \cong UT$. While the user U_k is playing the role R_l and performing the set of tasks, we collect some useful information: $Success\_score_i^{k,l}(q_v)$ (score on the success of the task), $Efficiency\_score_i^{k,l}(q_v)$ (score on the efficiency in time in the task) and an assessment regarding the user experience performing the UT, the $Satisfaction\_score_i^{k,l} \in S^5$. The first two are score assigned to a given q_v task and included as percentages in the final report, while satisfaction is considered as a linguistic variable for the test C_3. Specifically, for each user U_k playing the role R_l we complete the third column ($j = 3$) of its associated matrix $ID^{k,l}$ as follows:

$$ID_{i,3}^{k,l} = (Satisfaction\_score_i^{k,l}, 0) \ \ \forall i = 1, \ldots, n \quad (27)$$

- $C4 \cong ACC$. This tests scores alternative A_i with A, AA or AAA so it is directly in S^3 and is stored as a 2-tuple. Specifically, for each expert user U_k playing the role R_l we complete the fourth column ($j = 4$) of its associated $ID^{k,l}$ matrix as follows:

$$ID_{i4}^{k,l} = (Acc\_score_i^{k,l}, 0) \ \ \forall i = 1, \ldots, n \quad (28)$$

In summary, we highlight the linguistic membership scales of the elements of the $ID^{k,l}$ matrices, represented as 2-tuples:

- $ID^{k,l}_{i1} \rightsquigarrow S_{SUS}$
- $ID^{k,l}_{i2} \rightsquigarrow S_{SUS}$
- $ID^{k,l}_{i3} \rightsquigarrow S^5$
- $ID^{k,l}_{i4} \rightsquigarrow S^3$

3.4. Phase 3. Collective aggregation

In this phase, the individual ratings of the users playing different roles collected in the $ID^{k,l}$ matrices are aggregated into $r+1$ collective matrices: a matrix for each role $R_l, (l = 1, \ldots, r)$ and a global matrix that aggregates the matrices of each role.

Step 7. Unification of the information to S^9.

We must unify the values of the individual matrices in order to aggregate them. We transform each matrix $ID^{k,l}\{\emptyset\}$ containing the original user evaluations U_k playing the role R_l into a matrix with the unified evaluations in S^9, the deepest level of the hierarchy LH. We call these matrices *Unified Individual Decisions (UID)*. Depending on the linguistic scale used in the evaluation of each criterion, one transformation or another must be applied. Let us see how to proceed with each one:

- C_1 and C_2. Rates are on the scale S_{SUS}. We use the transformation function $\mathscr{LH}$ (see Section V.A of [47]) which allows to transform 2-tuples of an unbalanced linguistic scale, such as S_{SUS}, into a linguistic hierarchy, such as $LH = S^3 \cup S^5 \cup S^9$. Particularly, we set $\mathscr{LH}$:$(S_{SUS} \times [-0.5, 0.5)) \to (LH \times [-0.5, 0.5))$. After this conversion, the linguistic assessments are expressed in different linguistic domains which means that they cannot be processed directly. We require the transformation function $\mathscr{TF}$(see Section V.B of [47]) to convert the 2-tuples from different domains into a particular granularity label set of LH. We set $\mathscr{TF}^t_3$: $(LH \times [-0.5, 0.5)) \to (S^9 \times [-0.5, 0.5))$ as a special function that integrates a set of transformation functions TF [15] between levels of LH to the highest level of LH. In our case, by the definition of SUS, the obtained transformed 2-tuples can belong to either S^5 or S^9 getting:

$$UID_{ij}^{k,l} = \mathscr{T}\mathscr{F}_3^t(\mathscr{L}\mathscr{H}(ID_{ij}^{k,l})) \;\; \forall j = 1, 2; i = 1, \ldots, n \tag{29}$$

with t being level 2 or 3 of the hierarchy. If $ID_{ij}^{k,l} = \{\emptyset\}$, then $UID_{ij}^{k,l} = \{\emptyset\}$.

- C_3 and C_4. Rates are on the scale S^5 and S^3. We use the transformation function TF [15], which allows us to transform 2-tuples between any level of a linguistic hierarchy, in LH. Specifically, the third and fourth columns of the $UID^{k,l}$ matrices are obtained by transforming the 2-tuples of the $ID^{k,l}$ matrices to S^9 as follows:

$$UID_{ij}^{k,l} = TF_3^t(ID_{ij}^{k,l}) = \Delta\left(\frac{\Delta^{-1}(ID_{it}^{k,l}) \times 8}{n(t) - 1}\right) \;\; \forall j = 3, 4; i = 1, \ldots, n \tag{30}$$

with t being level 1 or 2 of the hierarchy. If $ID_{ij}^{k,l} = \{\emptyset\}$, then $UID_{ij}^{k,l} = \{\emptyset\}$.

Step 8. Aggregation for each role.

We define the *Unified Collective Decision matrix for role* R_l (UCD^l) containing the unified collective decisions in S^9 including the unified individual decisions of all users with role R_l. For this purpose, we consider the operator *2-tuple weighted average (2TWA)*. We obtain (UCD^l) as the aggregation of non-empty $UID^{k,l}$, ($k = 1, \ldots, u$) matrices by means of the 2TWA operator. Each element UCD_{ij}^l, ($i = 1, \ldots, n; j = 1, \ldots, m$) is defined as:

$$\begin{aligned} UCD_{ij}^l =& 2TWA_{W^l}(UID_{ij}^{1,l}, \ldots, UID_{ij}^{u,l}) \\ =& \Delta\left(\frac{\sum_{k=1}^{u} \Delta^{-1}(UID_{ij}^{k,l}) \times W_k^l}{\sum_{k=1}^{u} W_k^l}\right) = (s_{ij}^l, \alpha_{ij}^l) \end{aligned} \tag{31}$$

where the elements of the vector of weights $W^l = (W_1^l, \ldots, W_u^l)$ for role R_l are defined by $W_k^l = \frac{W_k'^l}{\sum_{k=1}^{u} W_k'^l}$, $k = 1, \ldots, u$, such as:

$$W_k'^l = \begin{cases} WU_k & if \;\; UID^{k,l} \neq \{\emptyset\} \\ 0 & if \;\; UID^{k,l} = \{\emptyset\} \end{cases} \tag{32}$$

Next we aggregate based on the weights of the criteria. This process generates a *unified collective decision* ($\boldsymbol{ucd}^l$) vector for each role R_l, which is used in the usability report.

$$\mathbf{ucd}_i^l = \Delta\left(\Delta^{-1}(s_{ij}^l, \alpha_{ij}^l) \times WC_j\right). \tag{33}$$

Step 9. Global aggregation

A *Unified Collective Decision* $UCD^{global} n \times m$ matrix containing the unified collective decisions in S^9 is defined by aggregating the unified collective decisions of each role R_l. For this purpose, the 2TWA operator is applied as the aggregation of the non-empty UCD_{ij}^l, $(l = 1, \ldots, r)$ matrices by means of the 2TWA operator. Each element of UCD_{ij}^{global}, $(i = 1, \ldots, n;\ j = 1, \ldots, m)$ is defined by Equation 34:

$$\begin{aligned} UCD_{ij}^{global} =& 2TWA_{WR}(UCD_{ij}^1, \ldots, UCD_{ij}^r) \\ =& \Delta\left(\sum_{l=1}^{r} \Delta^{-1}(UCD_{ij}^l) \times WR_l\right) = (s_{ij}^{global}, \alpha_{ij}^{global}) \end{aligned} \tag{34}$$

where WR is the normalized vector of weights of the roles verifying $\sum_{l=1}^{r} WR_l = 1$. Next, aggregation is performed based on the weights of the criteria. This process results in the generation of a global unified collective decision ($\mathbf{ucd}^{global}$) vector, which is then utilized in the usability report.

$$\mathbf{ucd}_i^{global} = \Delta\left(\Delta^{-1}(s_{ij}^{global}, \alpha_{ij}^{global}) \times WC_j\right). \tag{35}$$

3.5. Phase 4. Data exploitation

We apply the TOPSIS method (Section 2.5) on the Unified Collective Decisions matrices to generate several rankings of the alternatives, allowing to derive a ranking for specific roles and a general ranking. Thus, the model builds $r + 1$ *rankings*: one for each role R_l based on matrices $UCD^l, (l = 1, \ldots, r)$ and a global ranking based on matrix UCD^{global}. The ranking of the alternatives is established according to the relative closeness coefficient to the ideal alternative. The higher the coefficient value, the better is the alternative A_i. Then, all alternatives $A_i (i = 1, 2, \ldots, m)$ can be ranked according to a descending order of the relative closeness values.

Step 10. Generation of rankings for each role

The TOPSIS procedure is applied on the matrices UCD^l, $(l = 1, \ldots, r)$ whose values are 2-tuples $(s^l_{ij}, \alpha^l_{ij})$. Therefore, we set $(s_{ij}, \alpha_{ij}) = (s^l_{ij}, \alpha^l_{ij})$ to obtain r rankings, one for each role, which we denote by $Ranking^l, (l = 1, \ldots, r)$.

Step 11. Global ranking generation

The TOPSIS procedure is applied on the UCD^{global} matrix whose values are 2-tuples $(s^{global}_{ij}, \alpha^{global}_{ij})$. Therefore, we set $(s_{ij}, \alpha_{ij}) = (s^{global}_{ij}, \alpha^{global}_{ij})$ to obtain the global ranking that we denote by $Ranking^{global}$.

Step 12. Retranslation.
The elements of the vectors $\mathbf{ucd}^l$, $(l = 1, \ldots, r)$, and $\mathbf{ucd}^{global}$ are 2-tuples in S^9. This step conducts a retranslation of these values to S_{SUS} in order to provide more comprehensible linguistic terms for the users. Thus, it is not mandatory to perform this step to apply the LDM4USE model, but is convenient to complete the usability report.

We build an *adjective usability report* ($\mathbf{aur}^l$) vector for each role R_l and a global *adjective usability report* ($\mathbf{aur}^{global}$) vector containing the information from the $\mathbf{ucd}^l$ and $\mathbf{ucd}^{global}$ vectors, respectively, as linguistic terms of S_{SUS}. First, we define an identity function id to transform 2-tuple linguistic terms from S^9 to the linguistic hierarchy $LH = S^3 \cup S^5 \cup S^9$ by $id(s, \alpha) = (s, \alpha)$. Particularly, we set id:$(S^9 \times [-0.5, 0.5)) \rightarrow (LH \times [-0.5, 0.5))$. Subsequently, we use the inverse transformation function $\mathscr{LH}^{-1}$ (see Section V.A of [47]) which allows to transform linguistic 2-tuple expressed in a linguistic hierarchy, such as LH, into an unbalanced linguistic sale, such as S_{SUS}. Particularly, we set $\mathscr{LH}^{-1}$:$(LH \times [-0.5, 0.5)) \rightarrow (S_{SUS} \times [-0.5, 0.5))$. We compute the vectors $\mathbf{aur}^l$, $(l = 1, \ldots, r)$ and $\mathbf{aur}^{global}$ $(i = 1, \ldots, n; j = 1, \ldots, m)$ by Equations 36 and 37, respectively.

$$\mathbf{aur}^l_i = \mathscr{LH}^{-1}(id(\mathbf{ucd}^l_i)) \;\; \forall l = 1, \ldots, r \tag{36}$$

$$\mathbf{aur}^{global}_i = \mathscr{LH}^{-1}(id(\mathbf{ucd}^{global}_i)) \tag{37}$$

4. Case of study: virtual learning environments usability evaluation

In the decision-making field, real case studies and practical tools are essential for promoting and facilitating its application in various contexts [18, 48, 49, 40]. With the increase in the number of people who work from home or take online classes, new challenges over the use of technology are emerging. One of the major threats regarding technologies are loss of interest in *design for all* and adaptability to student's needs, since blended and hybrid education scenarios are increasingly present in High Education Institution (HEI) contexts. These new scenarios, particularly in the context of teaching and earning [24] need to be assessed. We propose to test one of the most popular Learning Management System (LMS), the one known as Moodle[9].

HEIs interested in providing inclusive virtual environments should pay adequate attention to three aspects:

1. The usability. To focus on the platform, which should be as useful as possible. This is relevant when the group of users is wide as when a Massive Open Online Course (MOOC) is offered.
2. The educational methodologies. To focus on the contents and materials that teachers share with their students, which should be designed based on the paradigm of Universal Design for Learning (UDL) [50].
3. The inclusive aspects. To enhance the system with Assistive Technologies (AT) which are designed to help groups of people with special needs.

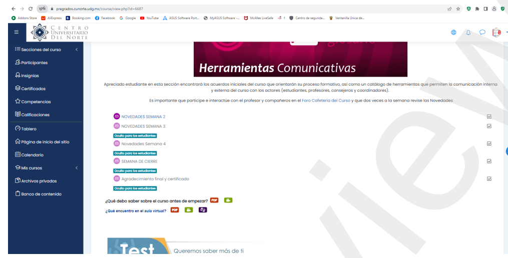
These three aspects have in common the use of technology. In our opinion, it is essential to rely additionally the UDL perspective in order to meet diverse needs through the good use of technology.

In order to showcase the practicality and utility of this proposal, the identical MOOC course has been deployed across diverse Moodle platforms within distinct university centers at the University of Guadalajara, Mexico. We got access to the **MOOC Course on Inclusive Educational Contexts: Design for All** [25] (a.k.a. DUA-MOOC), whose contents are the teaching practices to be taken to comply with the 9 UDL guidelines [50]. We have chosen to disable any AT in the three LMS , but users can enable

[9] About Moodle https://docs.moodle.org/402/en/About_Moodle

the AT that their operating system provides (e.g. Microsoft Narrator). In this way, an ideal scenario is achieved for the A/B testing comparison of the alternatives website, since the results will shed light only on the degree of usability of each MOODLE platform.

(a) A_1 = DUA-MOOC at CUTonalá.



(b) A_2 = DUA-MOOC at CUNorte.

(c) A_3 = DUA-MOOC at CULagos.

Figure 6: Using the same course on each website set the focus of the evaluation into the MOODLE theme.

4.1. Case use problem structure

We present a case use for the LDM4USE methodology given at Section 3. We aim to assess the usability of three Moodle learning platforms. These sites correspond to three separate Universities centers, and they differ in their released Moodle version and the theme or skin. In common, they have the same course content to be used in the Usability Test.

Step 1. Definition of the alternative set.

Three websites are established as the set of alternatives: University of Guadalajara Tonalá Center Moodle platform[10] (in sort, CUTonala), University of Guadalajara Northern University Center Virtual Campus[11] (in sort, CUNorte), and University of Guadalajara Center of Los Lagos Moodle platform[12] (noted, CULagos). Thus our set of alternatives are $A = \{\mathrm{CULagos}, \mathrm{CUNorte}, \mathrm{CUTonala}\}$. Figure 6 shows the DUA-MOOC main entry at each alternative website.

Step 2. Definition of the criteria set and derivation of the weight vector.

Let $C = \{\mathrm{SUS}, \mathrm{NPS}, \mathrm{UT}, \mathrm{ACC}\}$ be the set of criteria through which the usability of each alternative A_i is evaluated. Particularly for C_3, moderator has to define the content of the usability test and the estimated of maximum time per task. Public available *Usability Test for Moodle under Universal Design for Learning*[13] defines $UT = \{q_1, \ldots, q_{28}\}$. The requested tasks to be performed are listed in Table 1. Subsequently, we determine the criteria weights.

Step 2.1. Obtain pairwise judgments regarding the importance of criteria and filling in the criteria comparison matrix CP.

The moderator figure is responsible for determining the weight of each criterion C_j. It begins by establishing the importance of one criterion over

[10] CUTonala https://moodle2.cutonala.udg.mx/course/view.php?id=1605

[11] CUNorte https://pregrados.cunorte.udg.mx/course/view.php?id=6687

[12] CULagos https://plataforma.lagos.udg.mx/course/view.php?id=2278

[13] OD-Moodle-Usability-Assessment at GitHub https://github.com/ari-dasci/OD-Moodle-Usability-Assessment

Task category	Task List
Log in to the platform	1. Login to Moodle
	2. Find a course
	3. Access into the course
Technical support access	4. Find technical support documentation (manual, FAQ)
	5. Fill the technical support contact form
	6. Switch site language
User account management	7. Edit your profile
	8. Upload/Update profile photo
Access to information and resources/content	9. Read news items in what's new
	10. Download a file
	11. Download a file from resource Directory
	12. Track a URL link external to the platform
	13. Display an embedded video
	14. View a Page resource
	15. On the Page: read the text
	16. On the Page: visualize an image
Communication	17. Send a message to a co-worker/teacher
	18. Participate within a Chat
Accomplishment of course activities	19. Upload a file in the Task resource
	20. Answer a Questionnaire resource
	21. Add an entry to Glossary resource
	22. Select a sub-group
	23. Participate in the Forum resource
	24. In the editor: do text formatting
	25. In the editor: insert a new link
	26. In the editor: insert an image
	27. In the editor: do zoom (resize it or make full screen)
	28. Track your grades

Table 1: Task list for Moodle Usability Test under the UDL paradigm. Source [25], pp 159.

another, using a linguistic assessment $s \in S^5_{CP}$ to express the relative importance of each criteria test. Table 2 details the moderator's assessment of each pairs of criteria. We transform the linguistic labels into TFNs and complete the information of the CP matrix by using Equation 11 (see Table 3).

Step 2.2. Computing fuzzy synthetic extension.
The fuzzy synthetic extension is calculated by means of the Equations 12, 13 and 14. See Table 4 for the expression that is derived in this step.

Step 2.3. Degree of possibility.
The degree of possibility is calculated by means of the Equation 15, resulting

	C_1	C_2	C_3	C_4
C_1	Just import.	Very strongly import.	Equally import.	Weak import.
C_2		Just import.	Equally import.	Just import.
C_3			Just import.	Weak import.
C_4				Just important

Table 2: Assessment of the CP matrix by the moderator.

	C_1			C_2			C_3			C_4		
	$\boldsymbol{l}$	$\boldsymbol{m}$	$\boldsymbol{u}$	$\boldsymbol{l}$	$\boldsymbol{m}$	$\boldsymbol{u}$	$\boldsymbol{l}$	$\boldsymbol{m}$	$\boldsymbol{u}$	$\boldsymbol{l}$	$\boldsymbol{m}$	$\boldsymbol{u}$
C_1	1	1	1	5	7	9	1	1	3	1	3	5
C_2	$^1/_9$	$^1/_7$	$^1/_5$	1	1	1	1	1	3	1	1	1
C_3	$^1/_3$	1	1	$^1/_3$	1	1	1	1	1	1	3	5
C_4	$^1/_5$	$^1/_3$	1	1	1	1	$^1/_5$	$^1/_3$	1	1	1	1

Table 3: The CP matrix is set with moderator's rates for each criterion.

in the values given at Table 5.

Step 2.4. Obtaining the vectors of weights for the set C.

The vector of weights $WC' = \{1, 0.222, 0.515, 0.048\}$ is obtained through the Equation 16. Finally, using Equation 17, we derive the normalized weight vector of the criteria

$$WC = \{0.567, 0.114, 0.292, 0.027\}.$$

We know the moderator's assessment regarding relative importance of criteria is a correct evaluation by means of the consistency index (CI). In this scenario, $CI = -.08 \leq .09$ is valid [51]. Otherwise, moderator is prompted to change CP matrix.

Step 3. Definition of the user set.

Let $E = \{E_1, \ldots, E_4\}$ be the set of experts and $D = \{D_1, \ldots, D_{11}\}$ the set of end-users, so let $U = \{U_1, \ldots, U_{15}\}$ be the union of experts and end-users, such that $U = E \cup D$. Each user U_k evaluated the usability of each alternative A_i through each criterion C_j.

Furthermore, it is assumed that the expert set E have more knowledge in the UX discipline of Human Computer Interaction (HCI), so that their opinion can be of more importance for the general usability assessment that the given by D. In any case, let it be the moderator's decision to change

	Eq. 13		
Criteria	**l**	**m**	**u**
C_1	8.00	12.00	18.00
C_2	3.11	3.14	5.20
C_3	2.67	6.00	8.00
C_4	2.40	2.67	4.00
	Eq. 14		
	0.028	0.042	0.062
	Eq. 12		
	l	**m**	**u**
C_1	0.23	0.50	1.11
C_2	0.09	0.13	0.32
C_3	0.08	0.25	0.49
C_4	0.07	0.11	0.25

Table 4: Fuzzy synthetic extension calculation.

$C_1 > C_j$	**PD**	$C_2 > C_j$	**PD**	$C_3 > C_j$	**PD**	$C_4 > C_j$	**PD**
SC1>SC2	1.00	**SC2>SC1**	0.20	**SC3>SC1**	0.51	**SC4>SC1**	0.05
SC1>SC3	1.00	**SC2>SC3**	0.67	**SC3>SC2**	1	**SC4>SC2**	0.89
SC1>SC4	1.00	**SC2>SC4**	1.00	**SC3>SC4**	1	**SC4>SC3**	0.55

Table 5: Calculus of the possibility degree.

this policy for another of interest, for example, when participants are users with real disabilities and they want to increase the weight of this collective. We let the moderator to express this particular importance for each group. For this case, we have $WE = 100\%$ and $WD = 90\%$, and considering the membership of each user to one of these two groups, the following vector is stored.

$$WU = \{1, 1, 1, 1, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9, .9\}$$

Later, considering the role selected by the users, we calculate the corresponding normalized vector of user weights W^l following Equation 32.

Step 4. Definition of the set of roles.

Moderator selects the following roles $R = \{$*Blind*, *Ear infection*, *Arm injury*$\}$ as the set of possible choices, or equivalently, as the maximum number of times that a given user can answer the A/B test by playing roles without ever repeating. Then, moderator assigns importance to the roles by setting the vector of role weights $WR' = \{100, 75, 75\}$, which normalized to $WR = \{0.40, 0.30, 0.30\}$.

It is crucial to emphasize that users commence with the usability test (C_3), and subsequently, they are free to choose any other test, provided they respond to all of them. This consideration is associated with the time spent on the website during the usability test, as this criterion can be the most exploratory test for alternatives. Therefore, it holds significant value in experimenting with the system's usability.

Knowing which role each user plays, allows us to calculate the normalization of the user weights for each role. Our report confirms that users U_4, U_6 and U_{12} selected R_1 as they felt they want to play that role. Users U_2, U_5, U_{10}, U_{13} and U_{15} played R_2. Finally, users U_1, U_3, U_7, U_8, U_9, U_{11} and U_{14} selected role R_3 to play. When a user select a role, this is maintained for all the tests and for all alternatives. People is free to select any other role and start anew the A/B testing, answering from another point of view/need all tests for all alternatives.

By using Eq. 32 on WU for each role R_l ($l = 1, 2, 3$), we derive the normalized user weights:

$$W^1 = \{0, 0, 0, 0.357, 0, 0.321, 0, 0, 0, 0, 0, 0.321, 0, 0, 0\}$$

$$W^2 = \{0, 0.217, 0, 0, 0.196, 0, 0, 0, 0, 0.196, 0, 0, 0.196, 0, 0.196\}$$

$$W^3 = \{0.154, 0, 0.154, 0, 0, 0, 0.138, 0.138, 0.138, 0, 0.138, 0, 0, 0.138, 0\}$$

4.2. Case use data elicitation

We have conducted this A/B testing with real people at the *Graphics, Interfaces and Usability* (GIU) course. GIU is a fourth year course of the Computer Science Degree at the University of Guadalajara, Mexico. The professors of the course act as experts and the students as end-users. The test C are proposed to the students as practical exercises to perform accessibility and usability tests. The full A/B testing (assessments for all the alternatives and all the tests) took less than two hours, including the time spent giving instructions and two short break times.

Full set of answers (450 responses for SUS, 45 responses for NPS, 1260 responses for the UT and 12 responses for ACC) can be downloaded from the project repository at GitHub[14]. Particularly after performing the UT, we can derive from the tool useful information about tasks q_v (i.e. difficulty of the task) that are not part of the linguistic assessment matrix, but would be useful to design and develop better usability tests. To understand how the model use these information in a meaningful way, next we move into Phase 2 steps description.

Step 5. Gathering user evaluations.

Tests are executed sequentially. Either by using several data collection instruments, or assisted by software, we have to run each C_j and gather users' responses, and adapt them to our linguistic decision making model. Therefore, responses collect for the SUS Questionnaire (C_1) presented in Table 6. The NPS ratings (test C_2) are in Table 7. Full input data for C_3 test is not given here due to space restriction. Instead, we provide Table 8 with the information gathered from user U_4^1. It shows efficiency and success metrics per task of this particular user when running the UT for each of the alternatives. A quick look shows that U_4^1 performed best with A_3. Finally, and only for the group of experts D, the automatic accessibility assessment reports are linguistically interpreted in terms of number of errors and warnings. This information is provided in Table 9.

	U_1^3	U_2^2	U_3^3	U_4^1	U_5^2	U_6^1	U_7^3	U_8^3	U_9^3	U_{10}^2	U_{11}^3	U_{12}^1	U_{13}^2	U_{14}^3	U_{15}^2
A_1	42.5	30	52.5	60	50	30	62.5	60	65	32.5	30	37.5	47.5	40	42.5
A_2	50	55	62.5	42.5	80	57.5	55	47.5	62.5	50	60	45	57.5	62.5	47.5
A_3	60	27.5	75	30	40	27.5	50	55	72.5	50	57.5	37.5	57.5	60	70

Table 6: Input view of $C_1 \cong SUS$ responses for each A_i.

	U_1^3	U_2^2	U_3^3	U_4^1	U_5^2	U_6^1	U_7^3	U_8^3	U_9^3	U_{10}^2	U_{11}^3	U_{12}^1	U_{13}^2	U_{14}^3	U_{15}^2
A_1	4	4	6	3	6	6	5	5	7	6	7	6	5	1	6
A_2	7	7	1	2	8	7	5	8	6	8	8	7	5	2	8
A_3	3	3	5	1	5	7	1	8	5	8	7	7	1	7	6

Table 7: Input view of $C_2 \cong NPS$ with NPS_LTR_{score} data for each A_i.

[14]S-USE-AB_Tool https://github.com/ari-dasci/S-USE-AB-Tool/blob/main/use-cases/moodle.md

UT		$A_1 = \{\text{CULagos}\}$				$A_2 = \{\text{CUNorte}\}$				$A_3 = \{\text{CUTonalá}\}$			
q_v	$MaxTime$	$Time_1^{4,1}$	$Effic._1^{4,1}$	$Success_1^{4,1}$	$Satisf._1^{4,1}$	$Time_2^{4,1}$	$Effic._2^{4,1}$	$Success_2^{4,1}$	$Satisf._2^{4,1}$	$Time_3^{4,1}$	$Effic._3^{4,1}$	$Success_3^{4,1}$	$Satisf._3^{4,1}$
q_1	5	5	1	1	s_3^5	4	1	1	s_4^5	6	0	1	s_1^5
q_2	20	16	1	1	s_4^5	24	0	1	s_4^5	29	0	1	s_0^5
q_3	10	11	0	1	s_4^5	11	0	1	s_4^5	12	0	1	s_1^5
q_4	30	38	0	0	s_0^5	31	0	0	s_0^5	21	1	1	s_2^5
q_5	30	32	0	0	s_0^5	32	0	0	s_0^5	31	0	0	s_0^5
q_6	30	25	1	1	s_3^5	21	1	1	s_4^5	37	0	0	s_0^5
q_7	120	112	1	1	s_3^5	92	1	1	s_4^5	155	0	1	s_1^5
q_8	30	36	0	1	s_3^5	41	0	1	s_3^5	42	0	0	s_0^5
q_9	120	131	0	1	s_3^5	169	0	1	s_3^5	135	0	0	s_0^5
q_10	30	34	0	0	s_0^5	33	0	1	s_2^5	31	0	0	s_0^5
q_11	45	58	0	0	s_0^5	41	1	1	s_3^5	61	0	0	s_0^5
q_12	30	30	0	1	s_2^5	21	1	1	s_4^5	41	0	1	s_2^5
q_13	120	113	1	1	s_2^5	149	0	0	s_1^5	124	0	0	s_0^5
q_14	45	36	1	1	s_4^5	55	0	1	s_3^5	55	0	0	s_0^5
q_15	120	134	0	1	s_1^5	86	1	1	s_2^5	135	0	0	s_0^5
q_16	20	26	0	1	s_1^5	20	1	1	s_3^5	25	0	0	s_0^5
q_17	45	60	0	1	s_3^5	60	0	1	s_4^5	50	0	0	s_0^5
q_18	90	84	1	1	s_1^5	101	0	1	s_2^5	123	0	0	s_0^5
q_19	90	85	1	1	s_2^5	74	1	1	s_3^5	122	0	0	s_0^5
q_20	600	769	0	0	s_2^5	499	1	1	s_4^5	621	0	0	s_0^5
q_21	180	182	0	1	s_3^5	259	0	1	s_3^5	249	0	1	s_1^5
q_22	60	77	0	0	s_1^5	83	0	1	s_3^5	85	0	0	s_0^5
q_23	60	54	1	1	s_2^5	69	0	0	s_0^5	69	0	1	s_2^5
q_24	120	110	1	1	s_3^5	168	0	1	s_3^5	138	0	1	s_0^5
q_25	30	40	0	0	s_0^5	26	1	1	s_3^5	32	0	0	s_0^5
q_26	30	23	1	1	s_3^5	27	1	1	s_3^5	33	0	1	s_0^5
q_27	45	57	0	0	s_0^5	49	0	0	s_1^5	51	0	1	s_1^5
q_28	60	52	1	1	s_2^5	69	0	1	s_3^5	62	0	1	s_0^5
Average			**42.86%**	**71.43%**	$(s_2^5, -0.04)$		**39.29%**	**82.14%**	$(s_3^5, -0.29)$		**3.57%**	**42.86%**	$(s_0^5, 0.39)$

Table 8: A given user U_4 form E set is running C_3 under role R_1.

Step 6. Construction of the individual decision matrices.

This steps aims to construct the individual decision matrices with the answers of U users. In order to carry out linguistic collective aggregation, data must undergo a homogenization process. For clarity purposes, we adhere to the procedure of user U_4 evaluating alternative A_1 while assuming the role R_1:

- **Responses to** $C_1 \cong SUS$. The answers to the 10 items of SUS questionnaire are $\{2, 3, 4, 2, 3, 2, 2, 2, 3, 1\}$. Therefore, using Equation 18, we have $SUS\_score_1^{4,1} = 60$. This numerical value is transformed by means of Equation 24 getting $ID_{1,1}^{4,1} = TF_{SUS}(60) = (s_2^{sus}, 0.4) \in S_{SUS}$.
- **Responses to** $C_2 \cong NPS$. Direct response to LTR question (how likely is it that you would recommend the LMS to an acquaintance or friend), is 4. This numerical value is transformed to a linguistic 2-tuple by means of Equation 25, and thus $NPS\_SUS_{score} = 33.375$. By applying Equation 26 we have $ID_{1,2}^{4,1} = TF_{SUS}(33.375) = (s_3^{sus}, -0.335) \in S_{SUS}$.
- **Responses to** $C_3 \cong UT$. We apply an UT designed to fully use a

UX experts					End-Users										
U_1^3	U_2^2	U_3^3	U_4^1	U_5^2	U_6^1	U_7^3	U_8^3	U_9^3	U_{10}^2	U_{11}^3	U_{12}^1	U_{13}^2	U_{14}^3	U_{15}^2	
A_1	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-
A_2	A	AA	A	AA	-	-	-	-	-	-	-	-	-	-	-
A_3	A	A	A	A	-	-	-	-	-	-	-	-	-	-	-

Table 9: Input view of $C_4 \cong ACC$ with labels for accessibility.

LMS environment (in [25], page 150). It consists in 28 task (listed in Table 1). User U_4^1 gave the following results:

1. Efficiency rate: out of 28 activities, 12 were completed with an efficiency rate of $Efficiency\_score_1^{4,1} = 42.86$ (by eq. 19).
2. Success rate: out of the 28 activities, 20 were completed correctly obtaining $Success\_score_1^{4,1} = 71.43$ (by eq. 20).
3. Satisfaction level: satisfaction varied according to the task, but in average (applying eq. 21 and 27) we can derive for this user the $Satisfaction\_score_1^{4,1} = (s_2^5, -0.036)$. Therefore, $ID_{1,3}^{4,1}$=$(s_2^5, -0.036) \in S^5$.

- **Responses to $C_4 \cong$ Accessibility**. Acting as expert, U_4 used WAVE tool to browse the LMS and to have the reporting for accessibility evaluation. The interpretation of the same, is summarized in the valuation given A label. Therefore, $ID_{1,4}^{4,1} = (A, 0) = (s_0^3, 0) \in S^3$.

Thus, the elements of U_4^1 individual decision matrix for the evaluation of the alternative A_1 are:

$$ID_1^{4,1} = \{(s_2^{sus}, 0.4), (s_3^{sus}, -0.335), (s_2^5, -0.036), (s_0^3, 0)\}$$

Table 10 summarizes the matrices obtained in this step with regard to $U_k, (k = 1, \ldots, 15)$. Note that each $ID^{k,l}$ matrix is displayed in a single row of values, and that users have been grouped by roles.

4.3. Case use collective aggregation

In order to aggregate the information obtained previously from the set of users U, and considering the heterogeneity of the information, an unification process was necessarily previous to the aggregation of the information. The

	$A_1 = \{CULagos\}$				$A_2 = \{CUNorte\}$				$A_3 = \{CUTonala\}$			
U_r^l	$C_1 \cong SUS$	$C_2 \cong NPS$	$C_3 \cong UT$	$C_4 \cong ACC$	$C_1 \cong SUS$	$C_2 \cong NPS$	$C_3 \cong UT$	$C_4 \cong ACC$	$C_1 \cong SUS$	$C_2 \cong NPS$	$C_3 \cong UT$	$C_4 \cong ACC$
U_4^1	$(s_2^{sus}, 0.4)$	$(s_3^{sus}, -0.33)$	$(s_2^5, -0.04)$	$(s_0^3, 0)$	$(s_3^{sus}, 0.4)$	$(s_3^{sus}, -0.16)$	$(s_3^5, -0.29)$	$(s_1^3, 0)$	$(s_2^{sus}, 0.4)$	$(s_1^{sus}, -0.16)$	$(s_0^5, 0.39)$	$(s_0^3, 0)$
U_6^1	$(s_2^{sus}, 0.4)$	$(s_3^{sus}, -0.33)$	$(s_2^5, 0.07)$		$(s_2^{sus}, 0.3)$	$(s_3^{sus}, -0.16)$	$(s_2^5, 0.21)$		$(s_2^{sus}, 0.2)$	$(s_1^{sus}, -0.16)$	$(s_2^5, 0.36)$	
U_{12}^1	$(s_3^{sus}, 0)$	$(s_2^{sus}, 0.33)$	$(s_3^5, -0.5)$		$(s_4^{sus}, -0.4)$	$(S_0^{sus}, 0)$	$(S_2^5, 0.43)$		$(s_3^{sus}, 0)$	$(s_4^{sus}, -0.33)$	$(s_3^5, -0.11)$	
U_2^2	$(s_2^{sus}, 0.4)$	$(s_1^{sus}, -0.16)$	$(s_2^5, -0.29)$	$(s_0^3, 0)$	$(s_2^{sus}, 0.2)$	$(s_0^{sus}, 0.33)$	$(s_2^5, 0.46)$	$(s_1^3, 0)$	$(s_2^{sus}, 0.2)$	$(s_0^{sus}, 0)$	$(s_2^5, -0.14)$	$(s_0^3, 0)$
U_5^2	$(s_2^{sus}, 0)$	$(s_2^{sus}, 0.33)$	$(s_3^5, 0.14)$		$(s_6^{sus}, 0.4)$	$(s_7^{sus}, -0.33)$	$(s_2^5, 0.39)$		$(s_3^{sus}, 0.2)$	$(s_4^{sus}, -0.33)$	$(s_3^5, -0.46)$	
U_{10}^2	$(s_3^{sus}, -0.4)$	$(s_2^{sus}, 0.33)$	$(s_3^5, -0.11)$		$(s_2^{sus}, 0)$	$(s_3^{sus}, -0.16)$	$(s_2^5, 0.32)$		$(s_2^{sus}, 0)$	$(s_3^{sus}, -0.16)$	$(s_2^5, 0.43)$	
U_{13}^2	$(s_4^{sus}, -0.2)$	$(s_4^{sus}, -0.33)$	$(s_3^5, -0.25)$		$(s_2^{sus}, 0.3)$	$(s_4^{sus}, -0.33)$	$(s_2^5, 0.11)$		$(s_2^{sus}, 0.3)$	$(s_0^{sus}, 0)$	$(s_2^5, 0.46)$	
U_{15}^2	$(s_3^{sus}, 0.4)$	$(s_4^{sus}, -0.33)$	$(s_3^5, -0.11)$		$(s_4^{sus}, -0.2)$	$(s_7^{sus}, -0.33)$	$(s_2^5, -0.11)$		$(s_3^{sus}, -0.2)$	$(s_7^{sus}, -0.33)$	$(s_2^5, -0.29)$	
U_1^3	$(s_3^{sus}, 0.4)$	$(s_3^{sus}, -0.16)$	$(s_2^5, 0.25)$	$(s_0^3, 0)$	$(s_2^{sus}, 0)$	$(s_2^{sus}, 0.33)$	$(s_2^5, 0.18)$	$(s_0^3, 0)$	$(s_2^{sus}, 0.4)$	$(s_4^{sus}, -0.33)$	$(s_3^5, -0.29)$	$(s_0^3, 0)$
U_3^3	$(s_2^{sus}, 0.1)$	$(s_2^{sus}, 0.33)$	$(s_2^5, 0.25)$	$(s_0^3, 0)$	$(s_3^{sus}, -0.5)$	$(s_7^{sus}, -0.33)$	$(s_3^5, 0.18)$	$(s_0^3, 0)$	$(s_3^{sus}, 0)$	$(s_7^{sus}, -0.33)$	$(s_3^5, -0.25)$	$(s_0^3, 0)$
U_7^3	$(s_3^{sus}, -0.5)$	$(s_3^{sus}, -0.16)$	$(s_3^5, -0.46)$		$(s_2^{sus}, 0.2)$	$(s_7^{sus}, -0.33)$	$(s_3^5, -0.43)$		$(s_2^{sus}, 0)$	$(s_3^{sus}, -0.16)$	$(s_2^5, 0.32)$	
U_8^3	$(s_2^{sus}, 0.4)$	$(s_2^{sus}, 0.33)$	$(s_3^5, 0.21)$		$(s_4^{sus}, -0.2)$	$(s_3^{sus}, -0.16)$	$(s_3^5, 0.04)$		$(s_2^{sus}, 0.2)$	$(s_3^{sus}, -0.16)$	$(s_3^5, -0.11)$	
U_9^3	$(s_3^{sus}, -0.4)$	$(s_4^{sus}, -0.33)$	$(s_3^5, -0.11)$		$(s_3^{sus}, -0.5)$	$(s_4^{sus}, -0.33)$	$(s_2^5, 0.32)$		$(s_3^{sus}, -0.1)$	$(s_0^{sus}, 0)$	$(s_2^5, 0.43)$	
U_{11}^3	$(s_2^{sus}, 0.4)$	$(s_0^{sus}, 0)$	$(s_1^5, 0.46)$		$(s_2^{sus}, 0.4)$	$(s_0^{sus}, 0.33)$	$(s_2^5, 0.14)$		$(s_2^{sus}, 0.3)$	$(s_3^{sus}, -0.16)$	$(s_3^5, 0.14)$	
U_{14}^3	$(s_3^{sus}, 0.2)$	$(s_2^{sus}, 0.33)$	$(s_3^5, 0.21)$		$(s_3^{sus}, -0.5)$	$(s_7^{sus}, -0.33)$	$(s_3^5, -0.11)$		$(s_2^{sus}, 0.4)$	$(s_2^{sus}, 0.33)$	$(s_3^5, 0.25)$	

Table 10: Elements of $ID^{k,l}$ matrices, represented as 2-tuples.

necessary steps used to compute the UDC^{global} from the $ID^{k,l}$ matrices are detailed below.

Step 7. Unification of the information to S^9.

Values in Table 10 are unified to S^9 with the application of Equations 29 and 30. The result of the aforementioned procedures for each alternative are shown in Table 11.

	$A_1 = \{CULagos\}$				$A_2 = \{CUNorte\}$				$A_3 = \{CUTonala\}$			
U_r^l	$C_1 \cong SUS$	$C_2 \cong NPS$	$C_3 \cong UT$	$C_4 \cong ACC$	$C_1 \cong SUS$	$C_2 \cong NPS$	$C_3 \cong UT$	$C_4 \cong ACC$	$C_1 \cong SUS$	$C_2 \cong NPS$	$C_3 \cong UT$	$C_4 \cong ACC$
U_4^1	$(s_5^9, -0.2)$	$(s_3^9, -0.33)$	$(s_4^9, -0.07)$	$(s_0^9, 0)$	$(s_3^9, 0.4)$	$(s_6^9, -0.32)$	$(s_5^9, 0.43)$	$(s_4^9, 0)$	$(s_2^9, 0.4)$	$(s_2^9, -0.32)$	$(s_1^9, -0.21)$	$(s_0^9, 0)$
U_6^1	$(s_2^9, 0.4)$	$(s_3^9, -0.33)$	$(s_4^9, 0.14)$		$(s_5^9, -0.4)$	$(s_6^9, -0.32)$	$(s_4^9, 0.43)$		$(s_2^9, 0.2)$	$(s_2^9, -0.32)$	$(s_5^9, -0.29)$	
U_{12}^1	$(s_3^9, 0)$	$(s_5^9, -0.34)$	$(s_5^9, 0)$		$(s_4^9, -0.4)$	$(s_0^9, 0)$	$(s_5^9, -0.14)$		$(s_3^9, 0)$	$(s_4^9, -0.33)$	$(s_6^9, -0.21)$	
U_2^2	$(s_2^9, 0.4)$	$(s_2^9, -0.32)$	$(s_3^9, 0.43)$	$(s_0^9, 0)$	$(s_4^9, 0.4)$	$(s_1^9, -0.34)$	$(s_5^9, -0.07)$	$(s_4^9, 0)$	$(s_2^9, 0.2)$	$(s_0^9, 0)$	$(s_4^9, -0.29)$	$(s_0^9, 0)$
U_5^2	$(s_4^9, 0)$	$(s_5^9, -0.34)$	$(s_6^9, 0.29)$		$(s_6^9, 0.4)$	$(s_7^9, -0.33)$	$(s_5^9, -0.21)$		$(s_3^9, 0.2)$	$(s_4^9, -0.33)$	$(s_5^9, 0.07)$	
U_{10}^2	$(s_3^9, -0.4)$	$(s_5^9, -0.34)$	$(s_6^9, -0.21)$		$(s_4^9, 0)$	$(s_6^9, -0.32)$	$(s_5^9, -0.36)$		$(s_4^9, 0)$	$(s_6^9, -0.32)$	$(s_5^9, -0.14)$	
U_{13}^2	$(s_4^9, -0.2)$	$(s_4^9, -0.33)$	$(s_0^9, 0)$		$(s_5^9, -0.4)$	$(s_4^9, -0.33)$	$(s_0^9, 0)$		$(s_5^9, -0.4)$	$(s_0^9, 0)$	$(s_5^9, -0.07)$	
U_{15}^2	$(s_3^9, 0.4)$	$(s_4^9, -0.33)$	$(s_6^9, -0.21)$		$(s_4^9, -0.2)$	$(s_7^9, -0.33)$	$(s_4^9, -0.21)$		$(s_6^9, -0.4)$	$(s_7^9, -0.33)$	$(s_3^9, 0.43)$	
U_1^3	$(s_3^9, 0.4)$	$(s_6^9, -0.32)$	$(s_5^9, -0.5)$	$(s_0^9, 0)$	$(s_4^9, 0)$	$(s_5^9, -0.34)$	$(s_4^9, 0.36)$	$(s_0^9, 0)$	$(s_5^9, -0.2)$	$(s_4^9, -0.33)$	$(s_5^9, 0.43)$	$(s_0^9, 0)$
U_3^3	$(s_4^9, 0.2)$	$(s_5^9, -0.34)$	$(s_5^9, -0.5)$	$(s_0^9, 0)$	$(s_5^9, 0)$	$(s_7^9, -0.33)$	$(s_6^9, 0.36)$	$(s_0^9, 0)$	$(s_6^9, 0)$	$(s_7^9, -0.33)$	$(s_6^9, -0.5)$	$(s_0^9, 0)$
U_7^3	$(s_5^9, 0)$	$(s_6^9, -0.32)$	$(s_5^9, 0.07)$		$(s_4^9, 0.4)$	$(s_7^9, -0.33)$	$(s_5^9, 0.14)$		$(s_4^9, 0)$	$(s_6^9, -0.32)$	$(s_5^9, -0.36)$	
U_8^3	$(s_5^9, -0.2)$	$(s_5^9, -0.34)$	$(s_6^9, 0.43)$		$(s_4^9, -0.2)$	$(s_6^9, -0.32)$	$(s_6^9, 0.07)$		$(s_4^9, 0.4)$	$(s_6^9, -0.32)$	$(s_6^9, -0.21)$	
U_9^3	$(s_5^9, 0.2)$	$(s_4^9, -0.33)$	$(s_6^9, -0.21)$		$(s_5^9, 0)$	$(s_4^9, -0.33)$	$(s_5^9, -0.36)$		$(s_6^9, -0.2)$	$(s_0^9, 0)$	$(s_5^9, -0.14)$	
U_{11}^3	$(s_2^9, 0.4)$	$(s_0^9, 0)$	$(s_3^9, -0.07)$		$(s_5^9, -0.2)$	$(s_1^9, -0.34)$	$(s_4^9, 0.29)$		$(s_5^9, -0.4)$	$(s_6^9, -0.32)$	$(s_6^9, 0.29)$	
U_{14}^3	$(s_3^9, 0.2)$	$(s_5^9, -0.34)$	$(s_6^9, 0.43)$		$(s_5^9, 0)$	$(s_7^9, -0.33)$	$(s_6^9, -0.21)$		$(s_5^9, -0.2)$	$(s_5^9, -0.34)$	$(s_7^9, -0.5)$	

Table 11: Unified Individual Decisions (UID) matrices expressed with S^9.

Step 8. Aggregation for each role.

Another useful information is the unified collective decision vector, which is computed for each role. First, judgments in S^9 are clustered into a UCD^l matrix for each role R_l. Let us follow the case of $l = 1$ and users $U_k, k = \{4, 6, 12\}$ assessments represented in the UCD^1 matrix. Then we apply Equation 31 by using the vector of user weights W^1 in order to derive UCD^l_{ij} elements. Later, by means of Equation 33 and the vector of criteria weights (see Step 2.4. at Section 4.1) we compute the unified collective decision $\mathbf{ucd}^l$ vector per role. Results of both procedures are shown in Table 12. This has to be subsequently repeated to cover UCD^l $(l = 2, \ldots, r)$.

A_i	C_1	C_2	C_3	C_4	$\mathbf{ucd}^1_i$
	$WC^1 = 0.567$	$WC^2 = 0.114$	$WC^3 = 0.292$	$WC^4 = 0.027$	
A_1	$(s^9_3, 0.45)$	$(s^9_3, 0.31)$	$(s^9_4, 0.34)$	$(s^9_0, 0)$	$(s^9_4, -0.4)$
A_2	$(s^9_4, -0.15)$	$(s^9_4, -0.15)$	$(s^9_5, -0.08)$	$(s^9_4, 0)$	$(s^9_4, 0.17)$
A_3	$(s^9_3, -0.47)$	$(s^9_2, 0.32)$	$(s^9_4, -0.34)$	$(s^9_0, 0)$	$(s^9_3, -0.23)$

Table 12: Role playing with R^1 derive in $\mathbf{ucd}^1_i$, the 2-tuple vector with the usability assessment for each alternative.

Step 9. Global aggregation.

To integrate all the information, we examine the complete matrix of UCD^l_{ij} elements. Subsequently, we aggregate this information by taking into account the weights assigned to the roles, as outlined in Section 4.1, denoted as WR. We apply Equation 34 to compute each global unified collective decision UCD^{global}_{ij} element, used to report a linguistic score for each usability test and for each alternative.

Furthermore, a linguistic score can be derived to each alternative by means of Equation 35. We noted as the unified collective decision $\mathbf{ucd}^{global}_i$ vector and represent the *usability* of the websites. Table 13 shows 2-tuples on S^9 that are the collective representation of the usability assessments given by all users role-playing on the alternative websites through a set of tests.

4.4. Case use data exploitation

TOPSIS was used to rank the alternatives evaluated. The procedure for the two types of ranking achieved in the proposed model is detailed below:

	$\mathbf{ucd}_{i1}^{global}$	$\mathbf{ucd}_{i2}^{global}$	$\mathbf{ucd}_{i3}^{global}$	$\mathbf{ucd}_{i4}^{global}$	$\mathbf{ucd}_{i}^{global}$
A_1	$(s_4^9, -0.45)$	$(s_4^9, -0.34)$	$(s_5^9, -0.47)$	$(s_0^9, 0)$	$(s_4^9, -0.24)$
A_2	$(s_4^9, 0.30)$	$(s_4^9, 0.41)$	$(s_5^9, -0.36)$	$(s_3^9, -0.2)$	$(s_4^9, 0.37)$
A_3	$(s_4^9, -0.35)$	$(s_3^9, 0.25)$	$(s_4^9, 0.45)$	$(s_0^9, 0)$	$(s_4^9, -0.26)$

Table 13: For each alternative, the LDM4USE model can calculate and report the combined usability scores based on the weighted roles in each test.

Step 10: Generation of rankings by role.

For the following steps, the calculations are based on the matrix UCD^l for each role R_l, given at Table 14. Note that UCD^1 is repeated at Table 12. In the next, the step-wise derivation is described, the positive ideal solution A^{+l} and the negative ideal solution A^{-l} for R_1 are determined by Equations 3 - 6 and are shown below.

$$A^{+l} = \{2.183, 0.439, 1.438, 0.108\}, A^{-l} = \{1.434, 0.264, 1.067, 0.000\}, R_1$$

Next, the separation measures, D_i^{+l} and D_i^{-l}, of each alternative A_i from the positive ideal solutions A^{+l} and the negative ideal solutions A^{-l} are calculated by Equations 7-8 as expressed in Table 15. Also, the relative proximity coefficients $RC_i^l(i = 1, 2, 3)$ are calculated by Equation 9. They are use to rank alternatives as $Ranking^l$ and the results are shown in the Table 14.

		$C_1 \cong SUS$	$C_2 \cong NPS$	$C_3 \cong UT$	$C_4 \cong ACC$	$\mathbf{ucd}_i^l$	$Ranking^l$
		$WC_1 = 0.567$	$WC_2 = 0.114$	$WC_3 = 0.292$	$WC_4 = 0.027$		
$R^1 \cong See$	A_1	$(s_3^9, 0.45)$	$(s_3^9, 0.31)$	$(s_4^9, 0.34)$	$(s_0^9, 0)$	$(s_4^9, -0.4)$	2
	A_2	$(s_4^9, -0.15)$	$(s_4^9, -0.15)$	$(s_5^9, -0.08)$	$(s_4^9, 0)$	$(s_4^9, 0.17)$	1
	A_3	$(s_3^9, -0.47)$	$(s_2^9, 0.32)$	$(s_4^9, -0.34)$	$(s_0^9, 0)$	$(s_3^9, -0.23)$	3
$R^2 \cong Hearing$	A_1	$(s_3^9, 0.22)$	$(s_4^9, -0.38)$	$(s_4^9, 0.24)$	$(s_0^9, 0)$	$(s_3^9, 0.48)$	3
	A_2	$(s_5^9, -0.37)$	$(s_5^9, -0.42)$	$(s_4^9, -0.34)$	$(s_4^9, 0)$	$(s_4^9, 0.33)$	1
	A_3	$(s_4^9, -0.12)$	$(s_3^9, 0.13)$	$(s_4^9, 0.39)$	$(s_0^9, 0)$	$(s_4^9, -0.16)$	2
$R^3 \cong Touch$	A_1	$(s_4^9, 0.02)$	$(s_4^9, 0.18)$	$(s_5^9, 0.07)$	$(s_0^9, 0)$	$(s_4^9, 0.24)$	3
	A_2	$(s_5^9, -0.43)$	$(s_5^9, -0.02)$	$(s_5^9, 0.24)$	$(s_0^9, 0)$	$(s_5^9, -0.31)$	2
	A_3	$(s_5^9, -0.07)$	$(s_5^9, -0.4)$	$(s_6^9, -0.43)$	$(s_0^9, 0)$	$(s_5^9, -0.06)$	1

Table 14: Ranking of alternatives by roles.

According to $Ranking^1 = A_2 \succ A_1 \succ A_3$ which is focus on role $R_1 = \{See\}$, the best usability is shown at website A_2 and alternative A_3 has the lowest degree of usability for users with this role. With respect to role

$R_2 = \{Hearing\}$, we get another perspective as $Ranking^2 = A_2 \succ A_3 \succ A_1$, confirming that A_2 has better usability level over the other two alternatives. Finally, $Ranking^3 = A_3 \succ A_2 \succ A_1$ vary to allow us to discover some feature in A_3 that helped users with $R_3 = \{Touch\}$ impairment more that the other two alternatives.

Step 11: Global ranking generation.

In order to achieve a global ranking, the UCD^l matrices are considered as the basis (see Table 14). Next, we calculate the positive ideal solution A^+ and the negative ideal solution A^- by means of Equations 3 - 6 and are presented below.

$$A^{+l} = \{2.439, 0.503, 1.354, 0.076\}, A^{-l} = \{2.015, 0.370, 1.299, 0.000\}.$$

Subsequently, the separation measures, D_i^+ and D_i^-, of each alternative A_i from the positive ideal solution A^+ and the negative ideal solution A^- are calculated by Equations 7-8. Similarly, the relative proximity coefficients $RC_i (i = 1, 2, 3)$ are computed by Equation 9. All these results are given in Table 15.

A_i	D_i^+	D_i^-	RC_i	$Ranking^{global}$
A_1	0.440	0.053	0.108	3
A_2	0.000	0.454	1.000	1
A_3	0.401	0.058	0.126	2

Table 15: Values used to sort the alternatives as in $Ranking^{global}$

According to $Ranking^{global} = A_2 \succ A_1 \succ A_3$, of the three alternative websites, A_2 shows that is more suitable for real users considering an academic environment.

Step 12. Retranslation

We compute the *adjective usability report* ($\mathbf{aur}^l$) vector for each role R_l and the global *adjective usability report* ($\mathbf{aur}^{global}$) vector by means of Equations 36 and 37, respectively. Results are shown in Table 16. According to the rankings and to the adjective usability reports, is evident that A_2 is the best choice considering the usability aspects of the interface. Linguistically the three alternatives have an *OK* score according to Adjective SUS labels, but thanks to the proposed model we can manage better understanding of the user experiences at these three sites.

A_i	R_l	$\mathbf{ucd}_i^l$	$\mathbf{aur}_i^l$	$\mathbf{ucd}_i$	$\mathbf{aur}_i$	*Usability*
A_1	$R^1 \cong See$	$(s_4^9, -0.4)$	$(s_3^{sus}, 0.20)$	$(s_4^9, -0.24)$	$(s_3^{sus}, -0.12)$	*Ok*
	$R^2 \cong Hearing$	$(s_3^9, 0.48)$	$(s_2^{sus}, 0.48)$			
	$R^3 \cong Touch$	$(s_4^9, 0.24)$	$(s_3^{sus}, 0.12)$			
A_2	$R^1 \cong See$	$(s_4^9, 0.17)$	$(s_3^{sus}, 0.085)$	$(s_4^9, 0.37)$	$(s_3^{sus}, 0.185)$	*Ok*
	$R^2 \cong Hearing$	$(s_4^9, 0.33)$	$(s_3^{sus}, 0.165)$			
	$R^3 \cong Touch$	$(s_5^9, -0.31)$	$(s_3^{sus}, 0.345)$			
A_3	$R^1 \cong See$	$(s_3^9, -0.23)$	$(s_2^{sus}, 0.23)$	$(s_4^9, -0.26)$	$(s_3^{sus}, -0.13)$	*Ok*
	$R^2 \cong Hearing$	$(s_4^9, -0.16)$	$(s_3^{sus}, 0.42)$			
	$R^3 \cong Touch$	$(s_5^9, -0.06)$	$(s_3^{sus}, 0.47)$			

Table 16: Final scores to reporting are given under S^{SUS} Adjective scale.

4.5. A web tool based DSS to usability assessment

Our solution stands out for its flexibility in combining the various tests that a designer has to face, when doing accessibility and usability assessment. We call moderator the person in charge of validating which is the best design among the alternatives established in the A/B test. Using our tool, they sets up the project, shares links with the two groups of users and, once the tests have been carried out, obtains a final report in which the usability is given in SUS Adjective scale.

The following is a summary of the steps to create a project in USE A/B Testing DSS[15].

1. **Alternatives Definition:** The moderator, after creating the project template, adds each alternative for evaluation using the "Add" icon. In this modal window, they enter the name, URL, and insert the site logo image for each alternative.
2. **Criteria and Weight Definition:** The moderator selects questionnaires from the Test set (C) in USE-AB-testing DSS to evaluate the

[15] https://lionware.dev/S-USE-AB-TOOL-DSS

alternatives (A). New tests can be added from the System Test section to assess different types of sites or platforms. Then, linguistic labels indicating the importance of criteria in the comparison matrix are chosen. A consistency index $< .10$ confirms the validity of criteria importance, allowing the evaluation project configuration to proceed.

3. **Users and Weight Definition:** The moderator specifies the importance of opinions from two user types in USE-AB-test DSS (recipients and experts) to determine their relative significance.
4. **Roles and Weight Definition:** The moderator selects disabilities from predefined system categories. Enabling at least one disability in each category allows users and experts to assume related roles. Using a slider, the moderator determines the importance of incorporating user opinions in this category compared to others.
5. **Invitation Links Generation:** The system generates two invitation links, one for end-users and one for experts. The moderator shares these links with participants and opinions are gathered directly to the project.

We set the Moodle project on USE-AB-test DSS for the present use case. After the completion of all participants, the moderator request the project report. All three alternatives are labeled equally with the term *Ok*. However, a detailed view of the report by test and by role, as shown in Figure 7, helps to draw interesting conclusions: A_1 and A_3 has serious accessibility issues, A_2 strong point are NPS promoters and the Usability Test in all role-playing scenarios, also plays better A_3 under *touch* impairment.

5. Conclusions

This proposal incorporates linguistic techniques that are trends in the evaluation of the usability of IT systems. The importance of user roles –as user particular conditions or needs to identify– are aspects of critical importance in field of UX and usability assessment. Particularly this categorization contributes to the identification of areas of opportunity to improve end-user satisfaction, as well as to reduce the frustration in the use of the IT systems.

Under the perspective of testing the usability under the user's satisfaction opinion as a collective and the use of standard tests –SUS, NPS, UT and Accessibility– we have derived a proposal to perform A/B testing that is not limited by number of alternatives, neither limited to the number of tests.

Following Design Thinking paradigm, we have enabled the use of *personas* and *role-playing*. Also, we believe that by using a flexible evaluation model that consider user knowledge regarding how to assess real websites (because our model weight users' opinions), may help to balance the point of view of the development team with regard to the rest of stakeholders and final users. Other aspects to fine tune the A/B test include: tests relative importance, roles and roles' weights.

The utilization of linguistic variables simplifies the qualification of user perception. Additionally, the 2-tuple computational linguistic model, in conjunction with techniques such as Fuzzy AHP and Fuzzy TOPSIS, enables the processing of linguistic information without loss. Through the incorporation of collective opinions from decision-makers, including both experts and end-users, and the application of various ranking methods, we illuminate areas for improvement.

Our model is simple to be developed online, and can help to define new tests and roles, an to run an A/B test like the one described at Section 4. Our DSS[16] is designed to be able to cover the four phases needed to solve the usability assessment LDM problem: 1) to define the A/B test, 2) gather user information, 3) unification and collective aggregation, and 4) exploitation through rankings and the generation of the usability feedback report. This report describes the usability scores per role and per test, and also globally for all alternative website using the Adjective SUS scale, which is an standard in the UX field.

We have applied the model to a real use case regarding the assessment of Moodle, a LMS very popular in HEIs. According to the evaluation made by users U_k, alternative A_2 is the best website, information that is consistent with the results obtained in R^2 and R^3 of the same A_2. On the contrary, A_1 is the worst alternative, as demonstrated by the overall ranking in agreement with the evaluations performed for R^1 and R^2. Alternative A_2 is the best evaluated for R_1 and R_2, due to the fact that in *CUNorte* there is a group of visually impaired users, who requested to the technology area the installation of accessible components. On the contrary, the websites for A_2 and A_3, have the basic installation of Moodle, reason why the evaluations are below those made for A_1. Our end-users were students of a course that have to cover accessibility and usability evaluation. Thus this experience helped them to

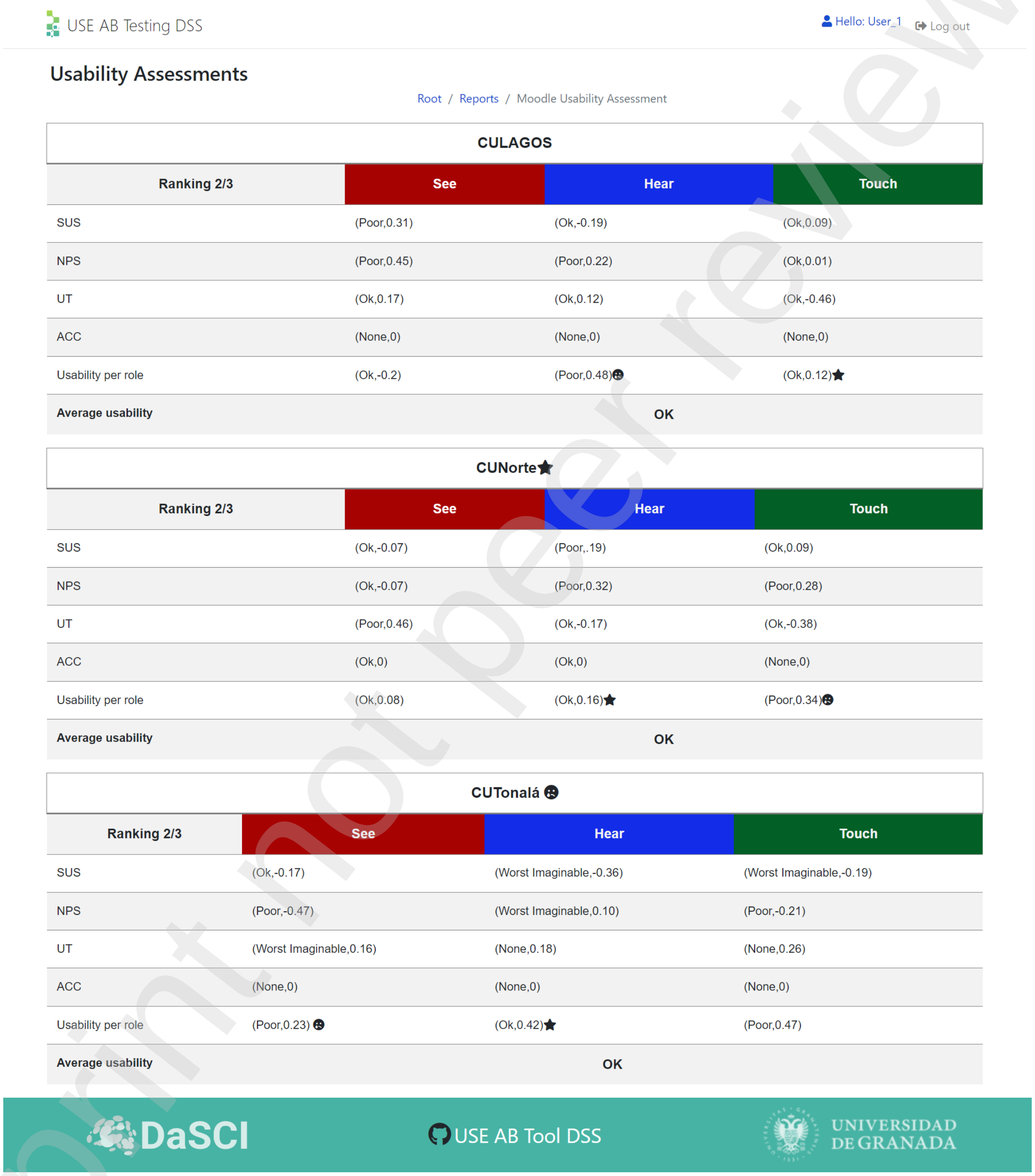
[16]USE-AB-test DSS https://github.com/ari-dasci/S-USE-AB-Tool/

improve these skills.

However, as future work, we are considering differentiating usability dimensions and use the model on more complex scenarios with real users requiring online assistive technology support. Such a case would require not only to apply our model, but also to see if these technologies are enabling or, on the contrary, are disabling because they are not properly integrated with the evaluated system.

Acknowledgments

This paper has been supported by Ministry of Science and Innovation, Spanish Government, grant number PID2020-119478GB-I00.



USE AB Testing DSS

Hello: User_1 Log out

Usability Assessments

Root / Reports / Moodle Usability Assessment

CULAGOS			
Ranking 2/3	See	Hear	Touch
SUS	(Poor,0.31)	(Ok,-0.19)	(Ok,0.09)
NPS	(Poor,0.45)	(Poor,0.22)	(Ok,0.01)
UT	(Ok,0.17)	(Ok,0.12)	(Ok,-0.46)
ACC	(None,0)	(None,0)	(None,0)
Usability per role	(Ok,-0.2)	(Poor,0.48)☹	(Ok,0.12)★
Average usability		OK	

CUNorte★			
Ranking 2/3	See	Hear	Touch
SUS	(Ok,-0.07)	(Poor,.19)	(Ok,0.09)
NPS	(Ok,-0.07)	(Poor,0.32)	(Poor,0.28)
UT	(Poor,0.46)	(Ok,-0.17)	(Ok,-0.38)
ACC	(Ok,0)	(Ok,0)	(None,0)
Usability per role	(Ok,0.08)	(Ok,0.16)★	(Poor,0.34)☹
Average usability		OK	

CUTonalá ☹			
Ranking 2/3	See	Hear	Touch
SUS	(Ok,-0.17)	(Worst Imaginable,-0.36)	(Worst Imaginable,-0.19)
NPS	(Poor,-0.47)	(Worst Imaginable,0.10)	(Poor,-0.21)
UT	(Worst Imaginable,0.16)	(None,0.18)	(None,0.26)
ACC	(None,0)	(None,0)	(None,0)
Usability per role	(Poor,0.23) ☹	(Ok,0.42)★	(Poor,0.47)
Average usability		OK	

DaSCI USE AB Tool DSS UNIVERSIDAD DE GRANADA

Figure 7: Usability evaluation report for three Moodle platforms obtained with USE A/B testing DSS.

References

[1] W. Chisholm, G. Vanderheiden, and I. Jacobs, "Web content accessibility guidelines 1.0," *Interactions*, vol. 8, no. 4, pp. 35–54, 2001.

[2] ISO/IEC International Standard, "IEC 9216," 1998.

[3] ISO 25000, "ISO/IEC 25010:2011: Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models," 2011.

[4] World Wide Web Consortium (W3C), "Web Content Accessibility Guidelines (WCAG) 2.2," http://www.w3.org/TR/WCAG22/, 2023.

[5] P. Jordan, B. Thomas, I. McClelland, and B. Weerdmeester, *Usability Evaluation In Industry*. Taylor & Francis, 1996. [Online]. Available: https://books.google.es/books?id=IfUsRmzAqvEC

[6] S. Krug, *Don't make me think!: a common sense approach to Web usability*. Pearson Education India, 2000.

[7] L. Martínez, D. Ruan, and F. Herrera, "Computing with Words in Decision Support Systems: An overview on Models and Applications," *International Journal of Computational Intelligence Systems*, vol. 3, no. 4, pp. 382–395, 2010.

[8] E. Triantaphyllou, *Multi-criteria Decision Making Methods: A Comparative Study*. Boston, MA: Springer US, 2000.

[9] S. A. Adepoju, I. O. Oyefolahan, M. B. Abdullahi, and A. A. Mohammed, "A survey of research trends on university websites' usability evaluation," *i-manager's Journal on Information Technology*, vol. 8, p. 11, 01 2019.

[10] B. K. Eldrandaly, A. A. Al, R. K. Chakrabortty, and M. Abdel-Basset, "An efficient framework for evaluating the usability of academic websites: Calibration, validation, analysis, and methods," *Neutrosophic Sets and Systems*, vol. 53, no. 1, pp. 1–24, 2023.

[11] G. Agrawal, A. Dumka, and M. Singh, "Usability and accessibility-based quality evaluation of Indian airline websites: An MCDM approach," *Universal Access in the Information Society*, 2022.

[12] O. Estrada-Molina, D. R. Fuentes-Cancell, and A. A. Morales, "The assessment of the usability of digital educational resources: An interdisciplinary analysis from two systematic reviews," *Education and Information Technologies*, vol. 27, no. 3, pp. 4037–4063, 2022.

[13] Y. Liu, C. M. Eckert, and C. Earl, "A review of Fuzzy AHP methods for Decision-Making with subjective judgements," *Expert Systems with Applications*, vol. 161, p. 113738, 2020.

[14] T. Lockwood, *Design Thinking: Integrating innovation, customer experience, and brand value*. Simon and Schuster, 2010.

[15] F. Herrera and L. Martinez, "2-Tuple Fuzzy Linguistic Representation Model for Computing with Words," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 6, pp. 746–752, 01 2001.

[16] C.-T. Chen, "Extensions of the TOPSIS for group decision-making under fuzzy environment," *Fuzzy Sets and Systems*, vol. 114, no. 1, pp. 1–9, 2000.

[17] A. Bangor, P. Kortum, and J. Miller, "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale," *Usability Studies*, vol. 4, no. 3, pp. 114–123, may 2009.

[18] F. J. Cabrerizo, J. A. Morente-Molinera, I. J. Pérez, J. López-Gijón, and E. Herrera-Viedma, "A decision support system to develop a quality management in academic digital libraries," *Information Sciences*, vol. 323, pp. 48–58, 2015.

[19] J. Nielsen, *Usability engineering*. Academic Press, 1993.

[20] M. Farzandipour, E. Nabovati, H. Tadayon, and M. S. Jabali, "Usability evaluation of a nursing information system by applying cognitive walkthrough method," *International Journal of Medical Informatics*, vol. 152, p. 104459, 2021.

[21] F. Shahini, D. Wozniak, and M. Zahabi, "Usability Evaluation of Police Mobile Computer Terminals: A Focus Group Study," *International Journal of Human–Computer Interaction*, vol. 37, no. 15, pp. 1478–1487, 2021.

[22] A. Hussain, E. O. Mkpojiogu, N. Ishak, N. Mokhtar, and Z. C. Ani, "An Interview Report on Users' Perception about the Usability Performance of a Mobile E-Government Application," *Int. J. Interact. Mob. Technol.*, vol. 13, no. 10, pp. 169–178, 2019.

[23] J. Y. Lee *et al.*, "Development and usability of a life-logging behavior monitoring application for obese patients," *Journal of Obesity & Metabolic Syndrome*, vol. 28, no. 3, p. 194, 2019.

[24] D. Buenaño-Fernandez, W. Villegas-CH, and S. Luján-Mora, "The use of tools of data mining to decision making in engineering education—a systematic mapping study," *Computer Applications in Engineering Education*, vol. 27, no. 3, pp. 744–758, 2019.

[25] L. B. Herrera Nieves, "Moodle Usability Evaluation. Inclusive Virtual Educational Environments based on Universal Learning Design," Ph.D. dissertation, Universidad de Granada, April 2020.

[26] B. Albert and T. Tullis, *Measuring the User Experience: Collecting, Analyzing, and Presenting UX Metrics.* Morgan Kaufmann, 2013.

[27] F. F Reichheld, "The One Number you Need to Grow," *Harvard business review*, vol. 81, pp. 46–54, 06 2004.

[28] J. Brooke, "SUS: A Retrospective," *Usability Studies*, vol. 8, no. 2, pp. 29–40, feb 2013.

[29] M. Schmettow, R. Schnittker, and J. M. Schraagen, "An extended protocol for usability validation of medical devices: Research design and reference model," *Journal of biomedical informatics*, vol. 69, pp. 99–114, 2017.

[30] C. C. Quinn, S. Staub, E. Barr, and A. Gruber-Baldini, "Mobile support for older adults and their caregivers: dyad usability study," *JMIR aging*, vol. 2, no. 1, p. e12276, 2019.

[31] P. Ambarwati and M. Mustikasari, "Usability Evaluation of the Restaurant Finder Application Using Inspection and Inquiry Methods," *Jurnal Sistem Informasi*, vol. 17, no. 2, pp. 1–17, 2021.

[32] R. Owen, *Net Promoter Score and Its Successful Application.* Singapore: Springer Singapore, 2019, pp. 17–29.

[33] J. Sauro and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research.* Morgan Kaufmann, 2016.

[34] R. Koning, S. Hasan, and A. Chatterji, "Experimentation and Start-up Performance: Evidence from A/B Testing," *Management Science*, vol. 68, no. 9, pp. 6434–6453, 2022.

[35] R. King, E. F. Churchill, and C. Tan, *Designing with data: Improving the user experience with A/B testing.* O'Reilly Media, Inc., 2017.

[36] S. Firmenich, A. Garrido, J. Grigera, J. M. Rivero, and G. Rossi, "Usability improvement through A/B testing and refactoring," *Software Quality Journal*, vol. 27, no. 1, pp. 203–240, 2019.

[37] R. F. Dam and T. Y. Siang, "Personas A Simple Introduction," https://www.interaction-design.org/literature/article/personas-why-and-how-you-should-use-them, 2023.

[38] Microsoft, "Inclusive 101 design toolkit," https://download.microsoft.com/download/b/0/d/b0d4bf87-09ce-4417-8f28-d60703d672ed/inclusive_toolkit_manual_final.pdf, 2016.

[39] T. Malhotra and A. Gupta, "A systematic review of developments in the 2-tuple linguistic model and its applications in decision analysis," *Soft Computing*, pp. 1–35, 2020.

[40] P. Liu, X. Dong, P. Wang, and R. Du, "Managing manipulation behavior in hydrogen refueling station planning by a large group decision making method with hesitant fuzzy linguistic information," *Information Sciences*, vol. 652, p. 119741, 2024.

[41] F. Herrera and L. Martínez, "A 2-tuple fuzzy linguistic representation model for Computing with Words," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 6, pp. 746–752, 2000.

[42] C.-L. Hwang, K. Yoon, C.-L. Hwang, and K. Yoon, "Methods for multiple attribute decision making," *Multiple attribute decision making: methods and applications a state-of-the-art survey*, pp. 58–191, 1981.

[43] H.-C. Liu, M.-L. Ren, J. Wu, and Q.-L. Lin, "An interval 2-tuple linguistic MCDM method for robot evaluation and selection," *International Journal of Production Research*, vol. 52, no. 10, pp. 2867–2880, 2014.

[44] F. Herrera and E. Herrera-Viedma, "Linguistic decision analysis: steps for solving decision problems under linguistic information," *Fuzzy Sets and Systems*, vol. 115, no. 1, pp. 67–82, 2000.

[45] D.-Y. Chang, "Applications of the extent analysis method on fuzzy ahp," *European journal of operational research*, vol. 95, no. 3, pp. 649–655, 1996.

[46] F. Herrera, S. Alonso, F. Chiclana, and E. Herrera-Viedma, "Computing with Words in Decision Making: Foundations, Trends and Prospects," *Fuzzy Optimization and Decision Making*, vol. 8, no. 4, pp. 337–364, dec 2009.

[47] F. Herrera, E. Herrera-Viedma, and L. Martínez, "A fuzzy linguistic methodology to deal with unbalanced linguistic term sets," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 2, pp. 354–370, 2008.

[48] H. T. Phan, N. T. Nguyen, V. C. Tran, and D. Hwang, "An approach for a decision-making support system based on measuring the user satisfaction level on twitter," *Information Sciences*, vol. 561, pp. 243–273, 2021.

[49] İlker Gölcük, "An interval type-2 fuzzy axiomatic design method: A case study for evaluating blockchain deployment projects in supply chain," *Information Sciences*, vol. 602, pp. 159–183, 2022.

[50] Center for Applied Special Technology (CAST), "Universal Design for Learning Guidelines v.2," 2011. [Online]. Available: https://www.cast.org/impact/universal-design-for-learning-udl

[51] F. R. Lima Junior, L. Osiro, and L. C. R. Carpinetti, "A comparison between fuzzy ahp and fuzzy topsis methods to supplier selection," *Applied Soft Computing*, vol. 21, pp. 194–209, 2014.