Applied Data Science and Analysis
Vol.2023, **pp**. 66–77
DOI: https://doi.org/10.58496/ADSA/2023/005 ; ISSN: xxx-xxxx
https://mesopotamian.press/journals/index.php/ADSA

Research Article

A Quality Assessment Methodology for Sign Language Mobile Apps Using Fusion Of Enhanced Weighted Mobile App Rating Scale (MARS) and Content Expert Standardized Criteria

Dianese David[1] Abdullah Hussein[1]

[1] *Faculty of Computing and Meta-Technology (FKMT), Universiti Pendidikan Sultan Idris (UPSI), Perak, Malaysia*

ARTICLE INFO

Article History
Received 09 Apr 2023
Accepted 30 Jun 2023

Keywords
Sign Language
Criteria
App Ranking
Mobile Apps

ABSTRACT

Mobile sign language apps have drawn a lot of interest recently as a way to minimize communication barriers between hearing people and people with hearing impairments. However, there are issues with the criteria and standards that should be taken into account when developing these apps. This study proposes a set of development criteria for sign language mobile apps and standardizes these criteria using the Fuzzy Delphi approach. Furthermore, the Fuzzy-Weighted Zero Inconsistency (FWZIC) approach is utilized to assign weights to the criteria and establish a ranking order. An initial set of requirements is developed based on the literature review. The Fuzzy Delphi technique is used, involving a panel of experts made up of developers, sign language experts, and users of sign language mobile apps, to assess the validity and reliability of the criteria. The FWZIC technique is used to give the criterion weights and determine their ranking order in order to further improve the decision-making process. The relative relevance of each criterion is determined by the FWZIC technique, which involves expert input and makes use of their knowledge and expertise. A thorough ranking is generated by taking into account the effects of each criterion on several zones, assisting in efficient decision-making during the creation of sign language mobile apps. Six Malaysian Sign Language apps that have been shortlisted are being utilized as a proof of concept to test the idea. The result of 6 apps is obtained based on the final standard criteria, their weights, and rankings.

*Corresponding author. Email: alamoodi.abdullah91@gmail.com

1. INTRODUCTION

Sign language is like spoken languages where it evolved naturally when different groups of people interacted with one another, resulting in a wide range of variants and there are between 138 and 300 different varieties of sign language being used now all over the world based on World Federation of the Deaf [1]. Examples of Sign Language that are available are Bangla Sign Language [2, 3], Bahasa Isyarat Malaysia [4-6], American Sign Language (A.S.L.) [7-17], Thai Sign Language [18-21], and many more. Among these sign languages, American Sign Language (A.S.L.) is the primary language widely used worldwide, and some other sign languages are adapted from A.S.L. According to [22]. Learning sign language provides many advantages where people will build a strong appreciation for deaf culture as good sign language users and even will be able to encourage language acceptance and understanding among others [23]. This will boost community and societal awareness, thus breaking down the barriers between hearing-disabled and hearing individuals. However, several different approaches or traditional methods of learning sign language are practised, but due to the development of mobile applications fueled by technological breakthroughs that have dramatically boosted the use of mobile phones and applications for learning sign language. Although a few existing sign language mobile learning applications [24] are often too limited and inadequate for efficient sign language learning. Therefore, an imperative need exists for automated and efficient assessment of mobile applications for sign language where users still heavily trust either application store ratings or the content rated by the application developer. Systematic assessment is necessary to evaluate app content [2]. Such a form of assessment necessitates a significance that is more than just information and expertise, and they are frequently time-consuming and practically challenging [25]. According to the author, the slow-paced scientific procedure targeted at app evaluation, the lack of characteristics that could assist in identifying precious apps, and the paucity of available apps that have been tested are all challenges to app quality assessments. Exploring this domain of evaluation was urged for a range of factors, such as the level of the knowledge in an app or content for an app [2], which would aid in the creation of a reliability app [26] and the consideration of strategies to ensure app quality [27]. One apparent method in this domain was using the Mobile Apps Rating Scale (MARS), a standardized tool the Queensland University of Technology created. MARS is regarded as a straightforward and trustworthy tool for categorizing and rating the standard of mobile health apps. It can also create a checklist for generating new, greater apps. MARS uses 23 distinct MARS items to rate apps based on engagement, functionality, aesthetics, and information quality [25]. Its utilization, therefore, is important in the assessment of mobile applications. While looking at academic literature, it turned out that most sign language mobile apps have been mainly assessed by either conducting in-house testing and evaluation or running these apps in different experimentations set by their developers and authors. None of these apps had undergone a quality assessment in terms of a standardized approach except for one study by [27], where authors performed a quality assessment on mobile apps for sign language using (MARS). However, despite the potential of MARS as an assessment tool that can be used for mobile apps, especially those related to sign language, it still has its shortcoming, which does not make it by itself at least the best tool for Sign language mobile apps. MARS was previously utilized by [27] for quality assessment of A.S.L. sign language mobile apps. However, this approach exhibits its shortcomings, including its reliance on subjective criteria of assessment items which only rely on user interaction with the apps producing a mean subjective value, and ignoring additional important criteria especially the ones associated with apps content. Accordingly, more quality assessment criteria are warranted for further exploration. For instance, a sign language mobile app X will be considered good if the raters feel the app is good based on their own subjective views, they might overlook the features and functions of the app which clearly also are important, because at the end of the day, user perspective is only one part of the assessment and the functions and internal features of the app also holds significant impact. Another example, if an app was considered good in term of MARS criteria, but the actual features of the app were bad and the information presented misguided users, this will reflect on its long-term usage and though it was initially deemed good, because of its content limitations, it will not be considered for official communications between normal and deaf users. Another significant issue is that MARS criteria are only determined by specific number of users (experts) who will judge the app, and since human involvement is presented, this will introduce bias, and insufficient evidence to be considered as ultimate guide to assess the app, and for that content of the app along with its design are also worthy of considering while assessing the app. Another shortcoming of MARS is that this assessment tool heavily relies on the mean score of the used assessment items, and it overlooked the individual importance and significance of each of these items when performing the assessment (weight) which cannot be considered as ultimate solution for assessing the apps. Therefore, there's a need to address gap of MARS assessment by infusing its criteria with ones from the academic literature linked to mobile apps ctotent, and thereby creating a unified assessment methodology where not only subjective criteria reflecting user perspective are only considered, but also introducing a content, functions and design criteria which will be as important for consideration while assessing sign language mobile apps, and measuring their importance weights and its impact on the assessment. This research also addresses other sign language mobile apps challenges aside from MARS based on the literature. Therefore, for several reasons mentioned above, this research attempts to perform an enhanced quality assessment methodology for Sign Language mobile apps using Mobile App Rating Scale (*MARS*) criteria along with other criteria standardized and weighted from the literature reflecting two main aspects;

subjectivity and objectivity and with different level of significance (*weights*) which will play a major role in the assessment procedure. This methodology is applicable to any sign language mobile apps assessment, but it was applied on case study for Malaysian Sign language mobile apps as proof of concept using different criteria and experimentations till a consensus is achieved and quality of apps are assessed and presented. The remaining of the paper is organized as follow;

2. LITERATURE REVIEW

Phases 1, 2, and 3 of the approach are identified: Identification, Weighting, and Ranking. Each process has been covered in detail below and is related to the others to attain the final result.

2.1. IDENTIFICATION PHASE

This phase begins by identifying all the collected criteria through literature and standardising them using the Fuzzy Delphi method (FDM). FDM is known as a step-by-step process and has been discussed below.

Step 1: Expert Identification: The author [28] mentions that the Delphi method's expert panel size varies, but with a homogeneous group of experts, good results can be obtained even with small panels of 10–15 people. (N=17) Experts from various countries and professions were involved in this research voluntarily.

Step 2: Developing Expert Form: The questionnaire was developed using Google Forms and divided into three sections: Part A, Part B and Part C. Part A requires the personal information such as name, position and working experience of experts. This part uses long text and a checklist menu where experts can write their position and choose the working experiences from given choices. Part B is for rating the main criteria consisting of six criteria: design, content, recognition, translation, enabling feature and Cost, and lastly, Part C is for the six sub-criteria. Part B and Part C are created based on the multiple-choice grid with a Likert scale of Very Important, Important, Neutral, Low Important and Not Important. After adapting the questionnaire from the literature, it was given to the supervisor for feedback. Then, modifications were made to the questionnaire based on their feedback.

Step 3: Dissemination and Data Collection: Seventeen (17) experts submitted a completed response, and the response data was exported from the survey as input to analyze in the fuzzy Delphi method and displayed in Data collection.

Table 1: Likert scale

Very Important	5
Important	4
Neutral	3
Low Important	2
Not at all Important	1

Table 1 above shows the Likert scale to evaluate their agreement level for each criterion and sub-criterion.

Step 4: Likert Scale Conversion into Fuzzy Set:

All experts' collected results are converted into triangular fuzzy numbering from the linguistic variables, as seen in Table 2.

Table 2: Fuzzy Linguistic variable

			FUZZY Likert	
Not at all Important	1	0	0	0.2
Low Important	2	0	0.2	0.4
Neutral	3	0.2	0.4	0.6
Important	4	0.4	0.6	0.8
Very Important	5	0.6	0.8	1

Step 5: Data Analysis and Threshold Value

The last step is testing the acceptance conditions for each item (i.e., criteria and sub-criteria). All experts have agreed on the item if the value d is less than 0.2. Otherwise, a second round will be required to determine whether the item is required.

Step 6: Data Analysis Expert Consensus: This section involves calculating experts' consensus, which must be greater than 75% following the second condition to be accepted. The fuzzy Delphi method also entails assessing whether the experts' consensus is more than or equal to 75 per cent for the complete dimensions of each item. If the proportion of consensus for each item is the same or more than 75%, it is presumed that all items have attained expert consensus [29].

Step 7: Data Analysis Defuzzification Process and Fuzzy Score Value: This part includes the defuzzification process to acquire the fuzzy score. The value for 'A' must be more than or equal to the medial alpha level of 0.5 to fulfil the approval for the last condition [30]. 'A' can be further utilised in identifying the priority element based on opinions.

Step 8: Data Interpretation: This part is Fuzzy Delphi Method data interpretation, where all data interpretation for Fuzzy Delphi Method is described based on the acceptance and rejection rules. Three requirements are needed for an item to be discarded or kept based on the consensus of experts.

1. A threshold of less than or equal to 0.2 must be accomplished for an item to be accepted [31].
2. For an item to be considered, most of the threshold must be greater than or equal to 75% [32].
3. The level of importance for a factor within an item can be defined using pseudo-partition, which is established by consulting an expert by setting a threshold to prevent insufficient partitioning [30].

2.2. FINAL STANDARDIZED CRITERIA

The table below shows the results of each criterion for the respective condition. Criteria accepted is the criteria that have met all three conditions; meanwhile, the rejected column shows the criteria that have not met the stated conditions. It can be seen clearly that only four main criteria (design, content, translation, and Cost) have passed all three conditions, while the rest (recognition and enabling feature) have been rejected. See Table 3

Table 3: Standardized criteria

Condition	Criteria Accepted	Criteria Rejected
1	Design, Content, Translation, Enabling features and Cost	Recognition
2	Design, Content, Recognition, Translation and Cost	Enabling Feature
3	Design, Content, Recognition, Translation, Enabling Feature and Cost	
Criteria that passed all three conditions (*Design, Content, Translation and Cost*)		

This entire Fuzzy Delphi method is used to standardise and reach a consensus among seventeen experts on collected criteria through a literature review. Some alternative criteria from the MARS have been included, thus making it in total of (n=9) criteria (Design, Content, Translation, Cost, Engagement, Functionality, Aesthetics, Information quality and Content-specific criteria).

2.3. PROCESS OF SEARCHING RELEVANT APP

This section explains searching relevant mobile apps on Malaysian Sign Language. The search for Malaysian sign language mobile apps began by constructing appropriate keywords. These search terms were used to identify the app, 'Malaysia Sign Language', 'M.S.L.', 'Bahasa Isyarat Malaysia', 'BIM', 'Sign Language Malaysia' and 'Kod Tangan Bahasa Melayu'. Finally, out of 96 apps, only six were linked to Malaysian Sign Language, including Malaysian Sign Language, KoTBaM, Eddy: Digital Learning of Sign Language, Bimonar (Early Access), Easy Sign Language and Learn Sign Language. (See Figure 1)

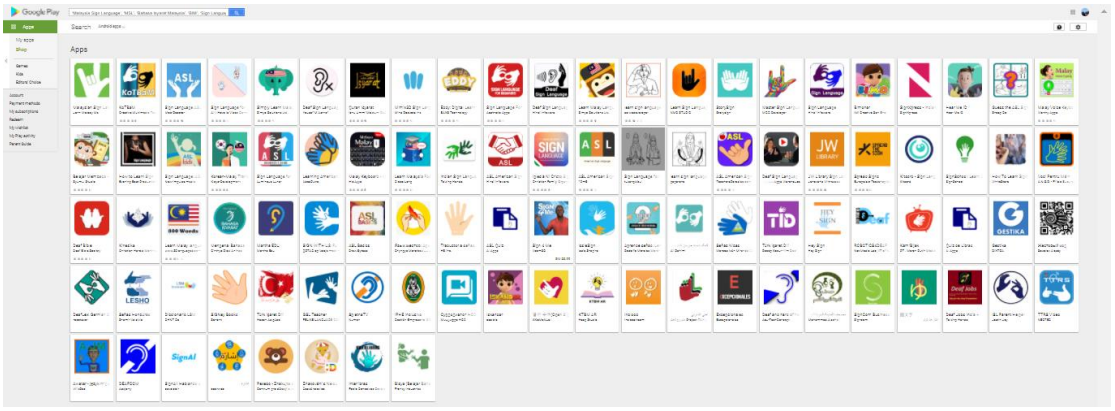

Figure 1: Apps Download

These apps were then downloaded using the Honor 50 Lite, android version 11, processor Qualcomm Snapdragon 662, RAM 8.0GB with a resolution of 2376 x 1080.

2.4. WEIGHTING PHASE

The fuzzy-weighted zero inconsistency (FWZIC) approach is proposed to compute the criteria' relevance with zero inconsistency. This approach was available in four separate versions, each with a different type of fuzzy. FWZIC is divided into five phases that must be fulfilled to be completed. The first three steps are the same regardless of whether the fuzzy environment is used, but the latter two need various mathematical operations based on the fuzzy environment.

Phase 1: Criteria definition: Different sets of criteria are provided to FWZIC when FDM is completed. These criteria are examined and specified in the first step. The next stage is to segregate and group all of the criteria and their sub-criteria.

Phase 2: Structured expert judgment: Structured expert judgement (S.E.J.) entails assessing the relative value of the previously stated criteria. For the criteria, an expert panel with experience in the research study's area does the examination. Following the compilation of the expert list, the nomination process is carried out in the following steps.

Step 1: Expert identification; In the context of the FWZIC, an expert is someone who has been or is still involved in the research study's areas and is considered competent by others. In the literature, experts are classified as 'domain' or' substantive' experts.
Step 2: Expert selection: Following the completion of expert identification, a group of experts for the case study is chosen. This step requires at least four specialists [33]. All former experts are contacted through email to determine their availability and willingness to participate on the panel as potential experts.
Step 3: Evaluation form development: The evaluation form was created since it is vital for gathering expert consensus. It is evaluated for reliability and authenticity before being finalised, and it is reviewed by all of the experts from the previous step.
Step 4: Defining the importance level scale: Using a 1–5 Likert scale, all of the experts chosen in the previous phase determine the importance level for each criterion.
Step 5: Converting linguistic scale into numerical scale: For the sake of the analysis, all preference values are converted from subjective to numerical form. As a result, each expert's importance level for each criterion in the utilised Likert scale is translated into a numerical scale, as shown in **Error! Reference source not found.**

Phase 3: Building the Expert Decision Matrix (E.D.M.): The primary sections of the E.D.M., which contain criteria and alternatives, are constructed, as indicated in the table below. The list of selected experts and each expert's choice within a particular criterion are defined in the preceding phase. The E.D.M. is created at this stage. The decision criteria and alternatives are the two primary components of the E.D.M., as indicated in Table 4. The table shows a crossover between the criteria and the experts, with each expert determining the priority level of each criterion.

Table 4: *Criteria Weighting*

Expert	Standardized criteria				MARS criteria				
	Design	Content	Translation	Cost	Engagement	Functionality	Aesthetics	Information quality	Content-specific criteria
1	4	5	5	5	4	5	3	3	2
2	5	5	5	5	5	5	4	5	5
3	4	5	5	4	3	3	3	5	4
4	5	5	5	5	5	5	5	5	5
5	2	3	2	4	3	2	1	3	3

According to the table, a crossover is made between the evaluation criterion model and the S.E.J. panel. Each criterion (Cj) in the attribute intersects with each selective expert (Ei), where the expert has scored the suitable level of importance per each criterion. The E.D.M. is the base for further analysis steps in the proposed method, which will be illustrated in the following sections.

Phase 4: Application of fuzzy memberships: The fuzzy membership function and associated defuzzification procedure are implemented to the E.D.M. data in this step, transforming the data to improve precision and ease of use in further analysis. Nevertheless, assigning a specific preference rate to any criterion is complex; the problem is ambiguous and imprecise. The fuzzy technique's purpose is to use ambiguous numbers instead of crisp numbers to calculate the relative value of qualities (criteria) to handle the issue of imprecise and uncertain problems. The most prevalent sort of fuzzy numbers used in FWZIC are triangular fuzzy numbers (TFNs). TFNs are written as A= (a.b.c). Because of their conceptual and computational ease, they are often employed in actual applications [34], as seen in the triangular membership in Figure 2.

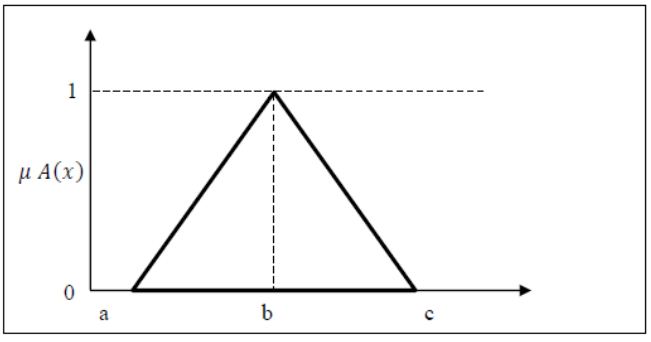

Figure 2. Membership of TFNs

TFN A's membership function (x) is provided by

$$\mu A(x) = \begin{cases} 0 \ if \ x < a \\ \dfrac{x-a}{b-a} \ if \ a \le x \le b \\ \dfrac{c-x}{c-b} \ if \ b \le x \le x \\ 0 \ if \ x > c \end{cases}, \qquad where \ a \le b \le c.$$

Formula 1 The membership function (x)

Let x = (a1, b1, c1) and y = (a2, b2, c2) be two non-negative TFNs, respectively, and R+ The arithmetic operations are defined according to the extension principle.:

1. $\tilde{x} + \tilde{y}$ = (a1 + a2, b1 + b2, c1 + c2),
2. $\tilde{x} - \tilde{y}$ = (a1 − c2, b1 − b2, c1 − a2),

3. $\alpha\tilde{x}$ = (αa1, αb1, αc1),
4. $\tilde{x}-1 \cong$ (1/c1, 1/b1, 1/a1),
5. $\tilde{x} \times \tilde{y} \cong$ (a1a2, b1b2, c1c2),
6. $\tilde{x}/\tilde{y} \cong$ (a1/c2, b1/b2, c1/a2).

The value of each linguistic term with TFN is shown in Table 5.

Table 5: Linguistic terms and their equivalent TFNs

Linguistic terms	TFNs
Not important	(0.00,0.10,0.30)
Slight important	(0.10,0.30,0.50)
Moderately important	(0.30,0.50,0.75)
Important	(0.50,0.75,0.90)
Very important	(0.75,0.90,1.00)

According to Table 5, all linguistic variables are translated into triangular fuzzy numbers, with the fuzzy number serving as the variable for each criterion for each expert. Diversely, the expert was asked to determine the relative importance of evaluation criteria within factors that are measured using linguistic variables.

COMPUTING THE FINAL WEIGHT

The final weight is calculated in three steps, with the following formulae used for associated calculations:

Step 1: The previous equations are utilised with TFNs [34]. Calculate the ratio of fuzzification data using the equation Formula 2. Table 6 shows the results.

$$\frac{\boldsymbol{Imp}(\widetilde{\boldsymbol{E1}}/\boldsymbol{C1})}{\sum_{j=1}^{n} \boldsymbol{Imp}(\widetilde{\boldsymbol{E1}}/\boldsymbol{C}_{1j})}$$

Formula 2 Ratio of fuzzification

Where $Imp(E1/C1\tilde{)}$ represents the fuzzy number of Imp (E1/C1) and $\Sigma\boldsymbol{Imp}(\boldsymbol{E1}/\boldsymbol{C}1\tilde{j})nj$=1 is the summation of all fuzzy number values of the important that assigned by expert per criterion.

Step 2: The final fuzzy values of the weight coefficients of the evaluation criterion $(\boldsymbol{w\tilde{1}},\boldsymbol{w\tilde{2}},...,\boldsymbol{w\tilde{n}})\boldsymbol{T}$ are computed using the mean values. The Fuzzy E.D.M. (EDM̃) is used to compute the final weight value of each criterion via Formula 3.

$$\widetilde{w_j} = \left(\sum_{i=1}^{m} \frac{Imp(\widetilde{E_{ij}}/C_{ij})}{\sum_{j=1}^{n} Imp(\widetilde{E_{ij}}/C_{ij})}\right)/m\ ), \text{ for } i = 1,2,3,..m \text{ and } j = 1,2,3,..n\ .$$

Formula 3 Fuzzy E.D.M.

The $w\tilde{j}$ is the fuzzy weight of each criterion, computed by dividing the ratio value of the criterion over the number of experts accordingly.

Table 6: Fuzzy weight

Expert	Standardized criteria												MARS criteria														
	Design			Content			Translation			Cost			Engagement			Functionality			Aesthetics			Information quality			Content-specific criteria		
	m1	m2	m3	m1	m2	m3	m1	m2	m3	m1	m2	m3	m1	m2	m3	m1	m2	m3	m1	m2	m3	m1	m2	m3	m1	m2	m3

1	0.50	0.75	0.90	0.75	0.90	1.00	0.50	0.75	0.90	0.30	0.50	0.75	0.50	0.75	0.90	0.75	0.90	1.00	0.30	0.50	0.75	0.30	0.50	0.75	0.10	0.30	0.50
2	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.50	0.75	0.90	0.75	0.90	1.00	0.75	0.90	1.00
3	0.50	0.75	0.90	0.75	0.90	1.00	0.30	0.50	0.75	0.75	0.90	1.00	0.30	0.50	0.75	0.30	0.50	0.75	0.30	0.50	0.75	0.75	0.90	1.00	0.50	0.75	0.90
4	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00	0.75	0.90	1.00
5	0.10	0.30	0.50	0.30	0.50	0.75	0.30	0.50	0.75	0.30	0.50	0.75	0.30	0.50	0.75	0.10	0.30	0.50	0.00	0.10	0.30	0.30	0.50	0.75	0.30	0.50	0.75
$\tilde{w}$	0.06	0.11	0.20	0.08	0.13	0.23	0.06	0.11	0.21	0.07	0.12	0.22	1.29	0.11	0.21	1.11	0.11	0.19	0.04	0.08	0.16	0.07	0.12	0.22	1.79	0.11	0.20
Defuzzification	0.12			0.15			0.13			0.14			0.54			0.47			0.09			0.14			0.70		
Final weight	0.048			0.058			0.052			0.061			0.217			0.189			0.037			0.053			0.285		
Rank	8			5			7			4			2			3			9			6			1		

Hints: 1- The Ratio value is computed by summing the same group's criteria per expert and then dividing each criterion value over their sum. 2- $\tilde{w}$ is the summation of all ratio values for each fuzzy member divided by several experts. 3- To defuzzied $\tilde{w}$ the summation of the triangular membership will compute and divided by 3 using the centroid method (refer to section 3.4.5), for example (0.06+0.11+0.20)/3 = 0.12. 4- The final weight will compute by dividing each of the defuzing $\tilde{w}$ on their summation, for example: in the main criteria level. S = 0.12+0.15+0.13+0.14+0.54+0.47+0.09+0.14+0.70= 3.20
Design final weight = 0.12/2.47 = 0.048

Step 3: The centroid approach is the most often used defuzzification method for determining the final weight (Local Weight (L.W.)). When employing TFNs, the mathematical equation is (a+b+c)3. Before computing the final weight coefficient values, each criterion's weight of importance should be assigned based on the sum of all criteria's weights for the rescaling purpose used in this stage.

Step 4: Compute the Global weight (G.W.): Finally, the global essential weights for each main criteria and sub-criteria computed with FWZIC are derived using the equation:

Formula 4 Compute the Global weight

$$GW = LW\ of\ main\ criterion * LW\ of\ subcriteria$$

Table 7: Results of Global Weight

Sub criteria	G.W.	Rank
Resizing	0.017184	1
Number of dictionaries	0.017168	2
Star rating	0.00342	3
Recommendation	0.003135	4
Willingness to pay	0.00285	5

Ease of use	0.002835	6
Interactivity	0.002821	7
Performance	0.002646	8
Interest	0.002604	9
Entertainment	0.001953	10
Offline and online usage	0.00087	11
Free apps	0.000854	12
Storing new word	0.000754	13
Image dictionary content	0.000754	14
Quality of information	0.000742	15
Visual information	0.000689	16
Word	0.000676	17
Help support	0.000672	18
Searching online for content	0.000638	19
Quantity of information	0.000636	20
Text to sign	0.000624	21
Customization	0.000576	22
Feedback for developer	0.000576	23
Sound	0.000572	24

2.5. DEVELOPMENT OF MOBILE ASSESSMENT

2.5.1. Development of Mobile Assessment

Upon the completion of criteria weighting, a mobile app rating scale (MARS) criteria is utilised to assess the quality of After determining the relative importance of each criterion, the quality of apps is evaluated using the Mobile App Rating Scale (MARS). When it comes to measuring the quality of mHealth apps, MARS is the first of its kind to provide a multidimensional measure of engagement, functionality, aesthetics, information quality, and subjective quality [35]. More than five experts use a 5-point Likert scale (1-Inadequate, 2-Poor, 3-Acceptable, 4-Good, 5-Excellent) to rate mobile apps on these indications and the weighted criteria. In the event that no application meets all criterion, the ranking will be determined by the application that meets the most of them. There will be a mix of Yes/No, objective, and subjective (5-point Likert Scale) scoring for some categories. As a result, the highest-scoring app according to these metrics will be the winner. Table 8.

Table 8: Final quality assessment criteria

	Main Criteria And Detailed Criteria	Main Criteria	Detailed Criteria	Average	Rank
App_1	0.093787	4.871167	1.136316	2.033756	3
App_2	0.093824	5.231833	1.089089	2.138249	2
App_3	0.147961	12.28717	2.0071	4.814076	1
App_4	0.0606	4.590667	0.988226	1.879831	5
App_5	0.087688	4.741833	0.984201	1.937908	4
App_6	0.004814	0.337167	0.080996	0.140992	6

A weighted average of the experts' ratings (AVGS) will be calculated for each app in the previous table. Each criterion will be assigned a score based on this formula. The final score for each AVGS is calculated by multiplying the AVGS by the criteria weight. The highest-scoring app will be given the highest ranking after all criteria scores are tallied up. The MARS criteria will be evaluated in their unmodified form and given weights to account for variations in app rank, with the results factored into the overall quality rating and ranking. App 3 (Eddy: Digital Learning of Sign Language) ranks highest with a value of 4.814076, according to the findings. The next highest value is 2.138249, achieved by App 2 (KoTBaM). App 1 (Malaysian Sign Language) has an average score of 2.033756, good enough for third place. App 5 (Easy Sign Language)

comes in at number 4 with a total score of 1.937908. App 4 (Bimonar) and App 6 (Learn Sign Language) come in at 5th and 6th place, with a score of 1.879831 and a score of 0.140992, respectively.

3. Discussion

We make countless choices every instant in the actual world, like the quote, "Life is full of options." When comparing options with simple qualities and a few comparisons, priorities can be easily assigned, clear reasons for the decision can be attained, and logical fallacies can be avoided. The Fuzzy Delphi approach has been widely used in several research domains to gather reliable expert input on a specific topic. The Fuzzy Delphi approach was used in this study to choose significant criteria from a wide range of choices and to improve the validation of mobile app assessment criteria. Among all the collected criteria, only four passed all three conditions, namely Design, Content, Translation and Cost, and were selected for the next assessment phase. Choosing among several attributes without knowing the priorities can significantly challenge decision-making. This entire Fuzzy Delphi method is used to standardise and reach a consensus among seventeen experts on collected criteria. To assist sign language mobile app assessment, there has been some research that involved the MARS criteria. Notably, these MARS criteria have already been adopted for mobile app assessment. In contrast, MARS criteria are insufficient for assisting sign language mobile app assessment and are instead used for general mobile app assessment due to domain problems. Although the MARS criteria are considered a valid and reliable scale for evaluating mobile apps, there is a need for more sturdy criteria for sign language mobile apps. Owing to this, criteria collected from the literature review were obtained and standardized using Fuzzy Delphi. The standardized criteria are then combined with MARS criteria for further assessment. FWZIC is performed in three layers (standardized criteria), (MARS criteria) and (standardized and MARS criteria) for both main and sub-criteria to identify the values and rank. The result reveals that MARS criteria have more influence for the main criteria, with 78.1%, compared to the standardized criteria, with only 21.9%. Meanwhile, standardized criteria are more significant for the sub-criteria with 82.5%, whereas MARS criteria are just 17.48% before FWZIC is performed. When FWZIC was executed separately on MARS and standardized criteria, it shows that standardized criteria are more influential than the MARS criteria. This resulted in standardized criteria having more influence than the MARS criteria when FWZIC was performed separately. This concludes that MARS and standardized criteria have their weight and are equally important. These criteria will eventually lead to a successful sign language mobile app development. This method used the weights assigned for each main and detailed criteria to identify the rank of six mobile apps relevant to Malaysian Sign Language. This computation was done in three layers (main criteria and detailed criteria), (main criteria) and (detailed criteria) only with the average values of each app. To determine the final rank, an average of nine layers of computation was performed to ensure the apps' rank. The computation was performed by (MARS main criteria) (MARS sub-criteria) (MARS all) (Standardized main criteria) (Standardized sub-criteria) (Standardized all criteria) (MARS and Standardized main criteria) (MARS and Standardized sub-criteria) (MARS and Standardized all). The study demonstrates a correlation between weights received by criteria and the ranking. The higher the weights received by criteria, the best is the app has been ranked. Criteria that assigned higher weight resulted in more influential criteria. More criteria may be included in an app. Nonetheless, it may not be the best app because it also depends on the expert's opinion of how far the requirements contained in the app are met. Throughout the computation for each stage, App 3 and App 6 have the same rank; meanwhile, App 1, App 2, App 4 and App 5 change their rank due to the weights and level of integration they obtained. The evaluation of mobile apps is a novel and crucial topic. This study has conducted a comprehensive review based on the Systematic Literature Review (S.L.R.) protocol on the smartphone applications of sign language. Mobile apps quality assessment protocol was proposed for assessing the quality of sign language mobile apps for Malaysian sign language with different criteria and experimentations until a consensus is achieved and the quality of apps is assessed. Criteria were collected from articles and standardized. These standardized criteria and MARS criteria were adopted for mobile app assessment. The results obtained were analysed and interpreted based on the criteria weights and level of these criteria integration into the apps. Both results are used for app ranking. While the criteria of assessment and its integration can influence the selection of the apps, to what extent can it influence the app? The evidence is clear that the criteria weights and level of integration influence Malaysian Sign Language mobile app selection. Moreover, a list of criteria influencing good mobile app development has been identified. This study provides future researchers with a comprehensive understanding of mobile app development and its criteria. From a long-term perspective, it would represent a significant step towards a more sustainable and reliable mobile app development, especially in sign language. This study reveals a set of criteria for reliable sign language mobile app development. However, these criteria might have some changes in terms of their influence over time. Timely research requires for successful sign language mobile app development.

4. Conclusion

The evaluation of mobile apps is a novel and crucial topic. for smartphone applications of sign language. Mobile apps quality assessment protocol was proposed for assessing the quality of sign language mobile apps for Malaysian sign language with different criteria and experimentations until a consensus is achieved and the quality of apps is assessed.

Criteria were collected from articles and standardized. These standardized criteria and MARS criteria were adopted for mobile app assessment. The results obtained were analysed and interpreted based on the criteria weights and level of these criteria integration into the apps. Both results are used for app ranking. While the criteria of assessment and its integration can influence the selection of the apps, but to which extent it can influence the app. The evidence is clear that the criteria weights and level of integration influence Malaysian Sign Language mobile app selection. Moreover, a list of criteria influencing good mobile app development has been identified. This study provides future researchers with a comprehensive understanding of mobile app development and its criteria. From a long-term perspective, it would represent a significant step towards a more sustainable and reliable mobile app development, especially in sign language. This study reveals a set of criteria for reliable sign language mobile app development. However, these criteria might have some changes in terms of their influence with time. Timely research requires for successful sign language mobile app development.

Conflicts of Interest

The paper explicitly states that there are no conflicts of interest to disclose.

Funding

We acknowledge the funding received from [Academic Institution] to carry out the research presented in this paper.

Data availability

N/A

REFERENCES

[1] D. Bragg *et al.*, *Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective*. 2019, pp. 16-31.

[2] J. K. a. B. Klímová, "Use of Smartphone Applications in English Language Learning—A Challenge for Foreign Language Education," *education sciences*, 2019, doi: 10.3390/educsci9030179.

[3] G. Abou Haidar, R. Achkar, D. Salhab, A. Sayah, and F. Jobran, "Sign language translator using the back propagation algorithm of an M.L.P.," in *2019 7th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, 2019: IEEE, pp. 31-35.

[4] H. Haron, H. Samad, F. Md Diah, and H. Yusof, "E-LEARNING APPROACH USING MOBILE APPS: MALAYSIAN SIGN LANGUAGE FOR DUMB AND DEAF," no. 1, pp. 1-7%V 1, 2019-06-23 2019. [Online]. Available: https://myjms.mohe.gov.my/index.php/ijarti/article/view/5991%J International Journal of Advanced Research in Technology and Innovation.

[5] R. L. Romero *et al.*, "Modifying the Mobile App Rating Scale With a Content Expert: Evaluation Study of Deaf and Hard-of-Hearing Apps," vol. 7, no. 10, 2019.

[6] M. G. Vintimilla, D. Alulema, D. Morocho, M. Proano, F. Encalada, and E. Granizo, "Development and implementation of an application that translates the alphabet and the numbers from 1 to 10 from sign language to text to help hearing impaired by Android mobile devices," in *2016 IEEE International Conference on Automatica (ICA-ACCA)*, 2016: IEEE, pp. 1-5.

[7] F. K. Ryan Lee Romero, Mark Hart, Amanda Ojeda, Itai Meirom, Stephen Hardy., "Modifying the Mobile App Rating Scale With a Content Expert: Evaluation Study of Deaf and Hard-of-Hearing Apps," *JMIR Mhealth Uhealth 2019*, vol. 7, 2019, doi: 10.2196/14198.

[8] K. A. Dawood, K. Y. Sharif, A. A. Ghani, H. Zulzalil, A. A. Zaidan, and B. B. Zaidan, "Towards a unified criteria model for usability evaluation in the context of open source software based on a fuzzy Delphi method," *Information and Software Technology*, vol. 130, p. 106453, 2021/02/01/ 2021, doi: https://doi.org/10.1016/j.infsof.2020.106453.

[9] A. A. Alsalem MA, Albahri OS, Dawood KA, Mohammed RT, Alnoor A, Zaidan AA, Albahri AS, Zaidan BB, Jumaah FM, Al-Obaidi JR., "Multi-criteria decision-making for coronavirus disease 2019 applications: a theoretical analysis review.," *Artificial Intelligence Review*, vol. 1, p. 84 2022, doi: 10.1007/s10462-021-10124-x

[10] H. I. Rahmat Dapari, Rosnah Ismail, Noor Hassim Ismail, "Application of Fuzzy Delphi in the Selection of COPD Risk Factors among Steel Industry Workers," *National Research Institute of Tuberculosis and Lung Disease*, vol. 16, pp. 46-52, August 25th 2016 2017.

[11] A. F. H. Mohamed Yusoff, Azmil & Muhamad, Norhisham & Wan Hamat, Wan., "Application of Fuzzy Delphi Technique Towards Designing and Developing the Elements for the e-PBM PI-Poli Module (Aplikasi Teknik Fuzzy Delphi Terhadap Elemen-Elemen Reka Bentuk Dan Pembangunan Modul e-PBM PI-Poli)," *Asian Journal of University Education,* vol. 17, pp. 292-304, 2021, doi: 10.24191/ajue.v17i1.12625. .

[12] F. Smarandache, J. E. Ricardo, E. G. Caballero, M. Y. L. Vázquez, and N. B. J. I. S. Hernández, "Delphi method for evaluating scientific research proposals in a neutrosophic environment," vol. 34, pp. 204-213, 2020.

[13] P.-L. Chang, C.-W. Hsu, and P.-C. Chang, "Fuzzy Delphi method for evaluating hydrogen production technologies," *International Journal of Hydrogen Energy,* vol. 36, no. 21, pp. 14172-14179, 2011.

[14] C.-H. Cheng and Y. Lin, "Evaluating the best main battle tank using fuzzy decision theory with linguistic criteria evaluation," *European journal of operational research,* vol. 142, no. 1, pp. 174-186, 2002.

[15] S. Bodjanova, "Median alpha-levels of a fuzzy number," *Fuzzy Sets and Systems,* vol. 157, no. 7, pp. 879-891, 2006.

[16] H.-C. Chu and G.-J. Hwang, "A Delphi-based approach to developing expert systems with the cooperation of multiple experts," *Expert systems with applications,* vol. 34, no. 4, pp. 2826-2840, 2008.