Research on Evaluation System of User Experience With Online Live Course Platform

SHURUI WANG, YAN LIU, FANGHAO SONG, XIAOJIAO XIE, AND DONG YU
School of Mechanical Engineering, Shandong University, Jinan 250061, China
Corresponding authors: Yan Liu (liuyan2008@sdu.edu.cn) and Fanghao Song (songfanghao@sdu.edu.cn)

ABSTRACT While the COVID-19 has triggered the rapid and large-scale promotion of online courses, it has also exposed the problem of poor user experience on online course platforms. Therefore, this research hopes to propose a complete user experience evaluation system to clarify the deficiencies of the platform and give directions for improvement. This research combines qualitative and quantitative, from the perspective of user experience, integrates data obtained through literature research, interview and network research to determine evaluation indices. The Delphi method and analytic hierarchy process are used to determine the weights of various indices and set the scoring level. It is obtained that the “ease of operation” has the most important impact on the user experience of the online live course platform. Innovatively adopting the t-test method to establish a comparative analysis method for the model application, a standardized process of model application is designed, and a complete evaluation system is constructed. Two pieces of software are selected as experimental samples to verify the model from both statistics and practical applications. The test results show that although the two experimental sample software have significant differences in the scores of multiple second-level indices, the overall scores are still similar, which is consistent with the direct evaluation results. Therefore, the user experience evaluation system proposed in this research is reliable, which can provide a reference for the construction of online live course platform and the optimization of user experience.

INDEX TERMS User experience, evaluation index, analytic hierarchy process, online live course platform.

I. INTRODUCTION Since 2020, the COVID-19 has led various schools around the world to shift from traditional classrooms to online live classrooms. The changes in classroom models and learning scenarios correspond to a series of changes in teaching methods, learning methods, and teacher-student communication methods, which is a challenge for both teachers, students and online course platforms [1]. However, the mismatch between design and demand has resulted in a poor user experience (UE) for live courses. At present, some remote office and conference software have taken on the role of online live course platforms, but due to their original design purpose and the nature of the software, these software cannot fully meet the needs of the school’s online live course. In this case, it is particularly important for the user experience optimization construction of an online live course platform.

To standardize the construction of online courses and MOOCs and enable learners to achieve the desired learning effects, many scholars and institutions have researched on online courses and MOOC evaluation indices. In the early years, there were Online Instructional Design Evaluation Framework (OMVU-OID) [2], Specification for Service Quality Management System of e-Learning (CELTS-24) [3], EduTools [4], distance undergraduate course teaching strategy evaluation method [5], etc. These evaluation systems mainly discuss the evaluation indices of online courses from the teaching and technical dimensions. Recently, the research has been continuously improved, some new classification angles have been proposed, and the level of indices has been more refined. Aurora Institute (formerly iNACOL) divided National Standards for Quality Online Courses into 5 sections: online course content, instructional design, technology, student assessment and course management [6]. Based on the E-xcellence [7], the “Open Education Quality Label” was proposed by the European Federation of Distance Education Universities (EADTU). It divides the evaluation indices into two first-level indicates, one that applies at the institutional level, another that applies to individual courses, and there are 32 third-level indices [8]. German scholar Yousef et al. proposed The Criteria to Assure Design Quality of MOOCs, including the Pedagogical Criteria and the Technical Criteria.
two first-level indices, several second-level and third-level indices [9]. The GB/T 36642-2018 “Information technology — Learning, education and training — Online courses” released by the Education and Information Technology Standard Committee of the Ministry of Education of the People’s Republic of China, divides the evaluation criteria into two aspects, which are Online Course Evaluation Program Information Model and Basic Requirements of the Online Course Operation Platform, as well as several specific indices and requirements [10]. These systems and norms are very large, covering many aspects of online course design. The evaluation criteria and hierarchical structure are also relatively clear and have certain reference value. However, most of these specifications based on the design and development of online courses, but rarely measure user experience from the perspective of users.

The online live course platform is the carrier of live courses, and user experience with it has been proven to have an important impact on the learning effect of students [11]–[13]. Some scholars have adopted the method of behavior analysis and obtained the twenty most frequent failures from the perspective of platform technology through platform log data [14], but the research based on behavior log has ignored the emotional and psychological experience of users. Other scholars focused on qualitative research. For example, Funieru et al. conducted a qualitative comparative study on the two MOOC platforms of edX and Coursera from the perspectives of users and technology, as well as 8 indices based on their own experience [15]. Liu et al. discussed the role of culture on the user experience of the MOOC platform from the perspective of localization and interface design [16]. The quality of these qualitative studies can be assessed and has some generic guidelines [17], [18], but there are still strong subjective problems in the use of indices and the evaluation methods. There are also some scholars exploring the user experience of online course platforms based on usability testing experiments, such as, Zaharias et al. introduced the affective learning dimension into traditional usability testing questionnaires, and proposed learning motivation as a new e-learning user experience measurement method [19]; Tsironis et al. conducted a comparative study on the usability of the three MOOC platforms, edX, Coursera and Udacity, using task performance indices and self-report indices [20]. These studies put forward some practical user experience measurement methods but did not establish a clear evaluation system.

In addition, looking at the above-mentioned studies on online courses, most of the research objects are video courses that have been recorded, while few studies on the evaluation indices of the live course form. There are certain differences between the two types of online courses in the form, so the evaluation indices will be different too. More exploration is needed for the evaluation research of live courses.

One way to establish an evaluation model is the analytic hierarchy process (AHP), a hierarchical and systematic analysis method [21]–[23] of multi-criteria decision-making (MCDM) [24]–[26] combined with qualitative and quantitative. At present, AHP has been combined a variety of methods and technologies, widely used in model construction, risk assessment and countermeasures proposed aspect [26] for industry and production [27], [28], geography and environment [29], [30], energy and sustainable development [31], [32], economy [33], [34], education [35] and other fields. The feasibility and openness provide the basis for the construction of the evaluation model based on UE in this research.

This research aims to combine the AHP with the UE perspective, and narrow the entry point to the construction of the evaluation model of the online live course platform. Based on literature research, in-depth interviews, and network research, combined with two rounds of Delphi method and AHP, build a more specific evaluation index system in order to provide more specific reference indices for platform improvement. This research also hopes to evaluate the ability of the model in practical applications through experiments. Taking into account the researcher’s country and cultural environment, Ding Talk and Tencent Meeting, the two most widely used software in China, are selected as experimental materials to verify the feasibility of the evaluation model in a comparative study. At the same time, the standardized process of the model was designed through experiments and a complete evaluation system was established.

In the rest of the paper, Section II introduces the main methods and detailed steps used in this study, Section III describes the establishment of indices, the determination of index weights, and the verification of the reliability and validity of the evaluation model. The evaluation model was verified by practical application and two application standardization processes were constructed in Section IV. Finally, Section V summarizes the paper.

II. METHODS

This research adopts a combination of qualitative and quantitative methods, using literature research, interview method, network research method and Delphi method to obtain qualitative data about evaluation indices, and using AHP to obtain quantitative index weights to construct evaluation models. Finally, the difference test of the questionnaire data obtained in the model application experiment is carried out to further verify the model, and the experimental process is constructed as a standardized evaluation process. The overall research process of this paper is shown in Fig 1.

A. INTERVIEW AND INTERNET RESEARCH

Five online live course platform users with a background in interactive design were selected as interviewees. Taking into account the convenience of participants, video interviews were used to conduct one-to-one semi-structured interviews. Interview questions include software experience (e.g. which platforms have you used to learn in the form of live courses?), preferences and levels (e.g. score these platforms on a scale of 1 to 9 and explain the reasons) and the following open core problem of the style: the advantages and disadvantages
of the platform you have used, and the functional points that you think can be improved on the existing live course platform. Based on the core question, the researcher was given the freedom to change the exact phrasing of questions and prompts [36]. The interview was controlled for about 20 minutes, and the whole interview process was recorded with the consent of the interviewee.

With the development of the Internet, people’s social interaction on the Internet has formed a large amount of information. Although this information is complicated and messy, the structured arrangement of it can still form valuable research materials [37], [38]. As far as this research is concerned, the evaluative information of the platform publicly released by users on the internet represents the user’s real experience of use, and these data play an important role in the construction of the evaluation system [39]. Collect online user evaluation data through the comment section of the IOS APP Store, and browse 50 high-scoring comments and 50 low-scoring comments for Ding Talk and Tencent Meeting. After these above, summarize interview content and comment content, and extract content keywords based on semantic analysis.

### B. DELPHI METHODS

The Delphi method uses a consensus-based iterative method to transform various opinions into a consensus [40], [41], and its anonymous response form makes the collected opinions more independent and authentic [42], which is conducive to the evaluation of indices. This study selected 12 experts from the perspectives of academic research and corporate practice, including 6 user experience research scholars, 4 UI or interaction designers, and 2 staff related to online live course platform. The survey is sent in the form of an electronic questionnaire, which was contingent on the accessibility of the experts [43], all experts were required to reply within 3 days. After the end of the first round, the expert suggestions are summarized and sorted, the feedback results are generated and distributed to the experts again. The content and sentences are modified according to the feedback results of the second round. Over all, the two rounds of Delphi will take one week.

### C. ANALYTIC HIERARCHY PROCESS

Analytic Hierarchy Process (AHP), proposed by American operations researcher Saaty [44]. The principle is to treat complex problems as a system and decompose the elements related to decision-making into goal layer, criterion layer, scheme layer, etc., finally, obtain the optimal scheme by solving the weight of the lower index to the upper layer [45]. The weight determination process quantifies the subjective assessment of experts and can check the consistency of the decision-makers’ evaluations [46]. The application of AHP is generally divided into three steps.

1. **Step 1: Construct judgment matrices**
   Construct reciprocal judgment matrices by comparing the importance of each element of a certain layer to the element of the upper layer. Construct a pairwise judgment matrix for m indices of the same layer:
   \[
   A = \begin{bmatrix}
   1 & a_{12} & a_{13} & \cdots & a_{1m} \\
   a_{21} & 1 & a_{23} & \cdots & a_{2m} \\
   a_{31} & a_{32} & \cdots & a_{ij} = 1/aij & \cdots \\
   \cdots & \cdots & \cdots & 1 & a_{im} \\
   a_{m1} & a_{m2} & \cdots & a_{mi} & 1
   \end{bmatrix}
   \]
   \[
   a_{ij} = \frac{1}{a_{ji}}
   \]

2. **Step 2: The consistency test of the judgment matrices**
   Single hierarchical arrangement and consistency test are performed in the top-down order, and the consistency index of the judgment matrices is CI:
   \[
   CI = \frac{\lambda_{\text{max}} - n}{n - 1}
   \]
\[ \lambda_{\text{max}} \] is the largest eigenvalue of matrix \( A \), \( n \) is the order of the matrix. The smaller the difference of \( \lambda_{\text{max}} - n \), the better the consistency of the matrix. It can be seen that the order \( n \) of the matrix has an impact on the consistency of the matrix, so the average random consistency index \( R_I \) is introduced. Using the consistency ratio \( CR \) as the test standard for the consistency of the judgment matrix, namely

\[ CR = \frac{CI}{RI} \quad (4) \]

If \( CR < 0.1 \), it is considered that the judgment matrix has satisfactory consistency, otherwise, the elements in the matrix need to be revised [47].

Step 3: Calculate the weight of each layer’s element

For the judgment matrix \( A \), satisfy the eigenvector corresponding to its largest eigenvalue:

\[ AW = \lambda_{\text{max}} W \quad (5) \]

After normalizing the feature vector to get the feature vector \( W_A = (w_1, w_2, \ldots, w_n) \). Take the ranking vectors (i.e. weights) \( W_1, W_2, W_3 \), and \( W_4 \) of the bottom layer to the upper layer to form a matrix \( D \), and multiply it with the weight vector matrix \( W_A \) of the upper layer to obtain the total ranking weight matrix of the bottom layer to the highest layer, namely

\[ W = DWA \quad (6) \]

The final vector \( W \) obtained is the ranking weight of the bottom index to the highest target.

### III. MATHCONSTRUCTION OF USER EXPERIENCE EVALUATION MODEL OF ONLINE LIVE COURSE PLATFORM

#### A. DETERMINATION OF EVALUATION INDICES

In OMVU-OID, each evaluation standard is divided into six components: standards and sub-standards, measurement criteria (also known as Description), benchmarks, ratings, weight and modification tips [2]. Because the two components of benchmarks and modification tips are of little significance to the establishment of this evaluation model, they are deleted. The standards at all level, the description items of indices, the weight of indices and the measurement criteria are defined as components of this evaluation model, and they are modified and established successively.

Integrating multiple user experience models, the user experience of this research is divided into four aspects: usability, ease of use, function and aesthetics. Summarize the content of interviews and online comments, extract content keywords based on semantic analysis, and combine with the previous literature research data to obtain a preliminary index system. The establishment of indices at all levels adopts a top-down thinking method. It is preliminarily determined to take X1 ease of operation, X2 functional completeness, X3 interface rationality, and X4 technical reliability as the first-level indices. Under this framework, 19 second-level indices are proposed, each second-level index proposes corresponding descriptions to clarify the meaning and comment content. After two rounds of the Delphi method, the expert opinions tended to be consistent. The merger and standardization of the second-level indices and the description items of the second-level indices were completed, the final evaluation indices were determined. In the final evaluation indices, “Y2 low threshold”, “Y6 repeatability” and “Y9 data visibility” are clarified and standardized as “versatility”, “reusability” and “course data visibility”; “Y7 high interaction” and “Y10 evaluaibility” are used as elements in the description items of the second-level indices, and to be classified into “teaching presence” and “freedom respectively”; the second-level index of “Y12 simplicity”, which’s concept overlaps with “Y11 clarity”, is deleted, and increased the “consistency” as a new second-level index; “Y16 delay” and “Y17 stress resistance” are combined to express as “stability”. Fig 2 shows the revision process of the two rounds of the Delphi method and final indices.

#### B. DETERMINATION OF WEIGHTS AND MEASUREMENT CRITERA

The AHP is used to determine the weight of each index. The user experience evaluation index of the online live course platform is taken as the target layer, and the indices of two level is taken as the criterion layer to establish a hierarchical structure model, as shown in Fig 3.

Invite 12 experts to judge and score the relative importance of 4 first-level indices and 16 second-level indices to construct judgment matrices. A total of 12 first-level indices judgment matrices and 48 second-level indices judgment matrices are established. Five final judgment matrices shown in Table 3 to Table 7 are obtained by averaging these judgment matrices.

For first-level indices judgment matrix \( A \), \( \lambda_{\text{max}} = 4.214 \), CI = 0.071, RI = 0.89, CR = 0.0776 < 0.1, which means the matrix has satisfactory consistency. For second-level indices, the second-level indices judgment matrix B1 for ease of operation, \( \lambda_{\text{max}} = 4.171 \), CI = 0.057, CR = 0.0639 < 0.1; the second-level indices judgment matrix B2 for functional completeness, \( \lambda_{\text{max}} = 4.157 \), CI = 0.052, CR = 0.0577 < 0.1; the second-level indices judgment matrix B3 for interface rationality, \( \lambda_{\text{max}} = 4.154 \), CI = 0.051, CR = 0.0695 < 0.1; the second-level indices judgment matrix B4 for technical reliability, \( \lambda_{\text{max}} = 4.046 \), CI = 0.015, CR = 0.0172 < 0.1. All of the five final judgment matrices all passed the consistency check.

For first-level indices judgment matrix \( A \), the feature vector is normalized to get \( W_A = (0.3935, 0.0941, 0.1558, 0.3565) \), the values correspond to the weights of “X1 ease of operation”, “X2 functional completeness”, “X3 interface rationality”, and “X4 technical reliability”. For second-level indices, the feature vector of the judgment matrix \( B_1 \) of ease of operation \( W_1 = (0.3726, 0.2954, 0.0863, 0.2457) \), the

### Table 2. Random consistency and corresponding values of matrix order.

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>RI</td>
<td>0</td>
<td>0</td>
<td>0.52</td>
<td>0.89</td>
<td>1.12</td>
<td>1.26</td>
<td>1.36</td>
<td>1.41</td>
<td>1.46</td>
</tr>
</tbody>
</table>

The matrix. It can be seen that the order n of the matrix has an impact on the consistency of the matrix, so the average random consistency index \( R_I \) is introduced. Using the consistency ratio \( CR \) as the test standard for the consistency of the judgment matrix, namely

\[ CR = \frac{CI}{RI} \quad (4) \]
feature vector of the judgment matrix $B_2$ of functional completeness $W_2 = (0.3943, 0.3273, 0.1969, 0.0815)$, the feature vector of the judgment matrix $B_3$ of interface rationality $W_3 = (0.0956, 0.4850, 0.2667, 0.1527)$, the feature vector of the judgment matrix $B_4$ of technical reliability $W_4 = (0.2254, 0.3774, 0.1005, 0.2968)$. These weights corresponding to these values are the ranking weights of various second-level indices to the corresponding first-level indices. Take the ranking vector (i.e. weight) $W_1, W_2, W_3, W_4$ of the second-level indices layer to the first-level indices layer as

\[ W = \begin{pmatrix}
W_1 \\
W_2 \\
W_3 \\
W_4 
\end{pmatrix} \]

and compute the vector $W$ using the hierarchical model of the evaluation system. The final evaluation result of the user experience of the online live course platform is obtained by superimposing the weights $I$ and the feature vector $F$ of the user experience. The final evaluation result $E$ is calculated as

\[ E = I \cdot W \cdot F \]
TABLE 4. Final judgment matrix $B_1$ of first-level indices.

<table>
<thead>
<tr>
<th>$B_1$</th>
<th>Y1</th>
<th>Y2</th>
<th>Y3</th>
<th>Y4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Y2</td>
<td>1/2</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Y3</td>
<td>1/4</td>
<td>1/3</td>
<td>1</td>
<td>1/3</td>
</tr>
<tr>
<td>Y4</td>
<td>1</td>
<td>1/2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

TABLE 5. Final judgment matrix $B_2$ of first-level indices.

<table>
<thead>
<tr>
<th>$B_2$</th>
<th>Y5</th>
<th>Y6</th>
<th>Y7</th>
<th>Y8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y5</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Y6</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Y7</td>
<td>1/3</td>
<td>1/2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Y8</td>
<td>1/4</td>
<td>1/3</td>
<td>1/4</td>
<td>1</td>
</tr>
</tbody>
</table>

Thus, the column vector to form the matrix D, and multiply it with the weight vector matrix $W_A$ of the first-level indices layer to obtain the total ranking weight of the bottom layer to the highest layer matrix, which is

$$W = DWA$$

$$= \begin{bmatrix}
0.3726 & 0 & 0 & 0 \\
0.2954 & 0 & 0 & 0 \\
0.0863 & 0 & 0 & 0 \\
0.2457 & 0 & 0 & 0 \\
0 & 0.3943 & 0 & 0 \\
0 & 0.3273 & 0 & 0 \\
0 & 0.1969 & 0 & 0 \\
0 & 0.0815 & 0 & 0 \\
0 & 0 & 0.0956 & 0 \\
0 & 0 & 0.4850 & 0 \\
0 & 0 & 0.2667 & 0 \\
0 & 0 & 0.1527 & 0 \\
0 & 0 & 0 & 0.2254 \\
0 & 0 & 0 & 0.3774 \\
0 & 0 & 0 & 0.1005 \\
0 & 0 & 0 & 0.2968 \\
\end{bmatrix} \begin{bmatrix}
0.1466 \\
0.1162 \\
0.0339 \\
0.0967 \\
0.0371 \\
0.0308 \\
0.0185 \\
0.0077 \\
0.0149 \\
0.0756 \\
0.0416 \\
0.0238 \\
0.0803 \\
0.1345 \\
0.0358 \\
0.1058 \\
\end{bmatrix} = \begin{bmatrix}
0.3935 \\
0.0941 \\
0.1558 \\
0.3565 \\
\end{bmatrix} \begin{bmatrix}
0.05 \\
0.051 \\
0.15 \\
0.25 \\
0.35 \\
0.45 \\
\end{bmatrix} = \begin{bmatrix}
0.3935 \\
0.3565 \\
\end{bmatrix}$$

Thus, obtained the weight vector $W$ is the final weight of the second-level indices Y1 to Y16 in turn. It can be clearly seen from Fig 4 that among the first-level indices, “X1 ease of use” occupies the most important position in the user experience evaluation model, followed by “X4 technical reliability”; and in the second-level indices, “Y1 fastness” is the most important index, followed by “Y14 stability”, “Y2 versatility” and “Y16 safety”.

After completing the weight calculation of the evaluation index, the evaluation standard is determined. There are three main types of expressions for evaluation criteria, namely adjective expressions, numeric expressions, and expressions that combine adjectives and numbers [48]. This study adopts the combination one, and uses a scale of 1-9 as the scoring standard for each index. 1 point means that a certain platform performs very poorly on this index item, 9 points means very good. At this point, the construction of the evaluation index system is completed as shown in Table 8:

C. RELIABILITY AND VALIDITY TEST OF THE EVALUATION MODEL

Correlation analysis is used to measure whether two variables are related or independent [49]. Pearson correlation analysis is used to verify the independence of the indices. It is
generally believed that the Pearson correlation coefficient greater than 0.8 means that the two variables have a higher correlation (the two indices in the evaluation model are highly correlated) and should be modified. The “Tencent Meeting User Experience Evaluation Questionnaire” was distributed to learners who used the Tencent Meeting to conduct online live course learning through the form of online questionnaires. A total of 60 questionnaires were distributed, 60 valid questionnaires were recovered, and SPSS 25.0 was used to analyze the questionnaire data. The analysis shows that the Pearson correlation coefficients of 16 second-level indices are all less than 0.8, which proves that the indices have good independence.

The test indices of the evaluation model mainly refer to reliability (consistency of measurement) and validity (measurement of target attributes) [50].

Reliability analysis is to check whether the sample data is reliable, usually internal consistency test is performed to verify whether the measurement of the item is the same kind of problem [51] [52]. The Cronbach’s alpha is used as the test value of the scale reliability, if the coefficient $\alpha$ is greater than 0.8, it is considered that the scale has high consistency, which means high reliability, and it can be used. If the coefficient $\alpha$ is less than 0.6, means that the reliability is not good and the question item needs to be modified. Using SPSS 25.0 to analyze the relative weight data of the user experience evaluation index of the online live course platform, the coefficient $\alpha$ is 0.958, which shows that the index weight has high reliability; to analyze the data of “Tencent Conference User Experience Evaluation Questionnaire” and the coefficient $\alpha$ is 0.892, indicating that the evaluation model has high reliability.

Validity analysis is to test whether the item design of the scale is reasonable. Factor analysis is adopted to verify the structural validity of the evaluation system (that is, whether the corresponding relationship between the research factors and the scale items is in line with expectations). Using SPSS 25.0 to perform exploratory factor analysis on the 16 indices’ data obtained from the “Tencent Meeting User Experience Evaluation Questionnaire”, the Kaiser-Meyer-Olkin (KMO) value was 0.803 greater than 0.6, the significance level of Bartlett test of sphericity was less than 0.05. The cumulative variance contribution rate of 4 factors divided into 16 items is 70.732%, there is no serious deviation in the corresponding relationship between the question item and the factor, and the factor loading on the corresponding factor of the 16 question items is greater than 0.5. Therefore, it is proved that the evaluation model has passed the validity test and has good validity.

### IV. APPLICATION OF EVALUATION MODEL AND STANDARDIZED PROCESS DESIGN

#### A. APPLICATION ANALYSIS OF EVALUATION MODEL

The fundamental purpose of constructing the evaluation model is to form a systematic and reasonable quantitative evaluation of the user experience of the online live course platform, to show the direction of improvement and optimization of the platform with more intuitive data, so as to promote the improvement of user experience and guarantee the learner’s learning effect. Although this evaluation model has undergone statistical tests to verify its rationality, the effect of the actual application is more valuable for research.
1) PROCESS FOR MODEL APPLICATION

Based on the idea of comparative analysis, two widely used software in china, ding talk (represented by d in the figure and table below) and tencent meeting (represented by t in the figure and table below) were selected as the research objects. The online environment for the evaluation is based on multiple devices and systems such as computers, mobile phones, and tablets. The undergraduate and postgraduate courses in higher education are used as the course content, and the course teachers are the teachers who should complete the courses offline in the university. Invite 40 learners who have experience in the use of the two pieces of software. these learners come from several universities in china. send them the “ding talk user experience evaluation questionnaire” or “tencent meeting user experience evaluation questionnaire”, asked them to rate the two pieces of software according to the evaluation model constructed in the previous article, and report the overall experience of the software. That’s for exploring whether the scores obtained by the two pieces of software according to this evaluation model are consistent with the preference degree in their intuitive impression.

2) ANALYSIS OF EVALUATION DATA

A total of 40 questionnaires were distributed and 40 valid questionnaires were collected. Among them, the evaluations of ding talk and tencent meeting are ordinary.

One thinking is the evaluation of a single product, that is, the analysis starts from the scores of each index item of the single product itself. For the two pieces of software, the unweighted data of the second-level indices is easier to visually see the advantages and disadvantages, so the original data is used for preliminary analysis. It can be seen from table 9 that the average scores of 20 users for the two pieces of software are mostly around 6 points (that is, little good), indicating that the performance of the two pieces of software’s second-level indices is acceptable. In the application of the evaluation model, the lower scores appear in the index items Y4, Y5, Y6, and Y7 of the tencent meeting, which is basically consistent with the shortcomings and function point data obtained in the previous interview. While higher scores appear on ding talk’s index Y8, tencent meeting’s index Y9 and two pieces of software’s index Y13.

The other thinking is a comparison study of products, which means to select multiple products to analyze the relative advantages and disadvantages of each product in each index item. independent-samples t-test or independent-samples nonparametric (npar) test is used to analyze the difference between the scores of the two pieces of software on each index and the total score. the independent-samples t-test is used to test whether there is a significant difference between two groups of normally distributed data, while the independent-samples npar test is used to test whether there is a significant difference between two or more non-normally distributed data.

Using spss 25.0 to test the normality of the data. Due to the small sample size (less than 50), the shapiro-wilk test (s-w test) is used. Normality test results show that at least one set of software scoring data has a significant p-value of less than 0.05 in most indices, which proves that most scoring data are non-normally distributed, so independent sample NPar tests are performed on them. the significance p-values of the three second-level indices Y5, Y7, and Y16, the four first-level indices, and the total score are all greater than 0.05, which conforms to the normal distribution. Therefore, the data of these eight indices are subjected to independent sample T test.

The independent-sample npar test uses the mann-whitney test. It can be seen from fig 5 that for the second-level indices, the significance p values of Y1, Y4, Y6, and Y15 are all less than 0.05, indicating that the scores of the two pieces of software have significant differences in these five indices. According to the mean rank data, tencent meeting is significantly better than ding talk in terms of “y1 fastness”, while ding talk is significantly better than tencent meeting in terms of “Y4 fault tolerance”, “Y6 reusability”, “Y8 course data visibility”, and “Y15 show diversity”.

Independent-samples t-test was performed on the index data of all levels that conform to the normal distribution. It can be seen from Fig 6 that the significance p-value of the second-level indices Y5, Y7 and the first-level index X2 is less than 0.05. For the second-level indices, the “Y5 teaching resource accessibility” and “Y7 teaching presence” of Ding Talk is significantly better than Tencent Meeting; for the first-level indices, Ding Talk is significantly better than Tencent Meeting. The overall experience of the software is reported by these learners from several universities in china.

### Table 9. Mean table of user ratings for second-level indices.

<table>
<thead>
<tr>
<th>Index</th>
<th>Software</th>
<th>Mean</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>D</td>
<td>5.50</td>
<td>0.94591</td>
<td>0.21151</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>6.45</td>
<td>1.23438</td>
<td>0.27601</td>
</tr>
<tr>
<td>Y2</td>
<td>D</td>
<td>6.70</td>
<td>1.17429</td>
<td>0.26258</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>6.45</td>
<td>0.75915</td>
<td>0.16975</td>
</tr>
<tr>
<td>Y3</td>
<td>D</td>
<td>6.25</td>
<td>0.71635</td>
<td>0.16018</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>5.85</td>
<td>0.74561</td>
<td>0.16662</td>
</tr>
<tr>
<td>Y4</td>
<td>D</td>
<td>5.75</td>
<td>0.91047</td>
<td>0.20359</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>4.40</td>
<td>1.35536</td>
<td>0.30262</td>
</tr>
<tr>
<td>Y5</td>
<td>D</td>
<td>6.70</td>
<td>1.03110</td>
<td>0.23056</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>4.55</td>
<td>1.46808</td>
<td>0.32827</td>
</tr>
<tr>
<td>Y6</td>
<td>D</td>
<td>6.40</td>
<td>0.94032</td>
<td>0.21026</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>4.05</td>
<td>1.84890</td>
<td>0.41343</td>
</tr>
<tr>
<td>Y7</td>
<td>D</td>
<td>6.75</td>
<td>1.08942</td>
<td>0.24360</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>4.20</td>
<td>1.70488</td>
<td>0.38113</td>
</tr>
<tr>
<td>Y8</td>
<td>D</td>
<td>7.30</td>
<td>1.12858</td>
<td>0.25236</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>5.75</td>
<td>1.06992</td>
<td>0.23924</td>
</tr>
<tr>
<td>Y9</td>
<td>D</td>
<td>6.60</td>
<td>1.18766</td>
<td>0.26557</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>7.00</td>
<td>1.12390</td>
<td>0.25131</td>
</tr>
<tr>
<td>Y10</td>
<td>D</td>
<td>6.15</td>
<td>1.34849</td>
<td>0.30153</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>6.55</td>
<td>0.75915</td>
<td>0.16975</td>
</tr>
<tr>
<td>Y11</td>
<td>D</td>
<td>6.25</td>
<td>1.40955</td>
<td>0.31519</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>6.15</td>
<td>0.87509</td>
<td>0.19658</td>
</tr>
<tr>
<td>Y12</td>
<td>D</td>
<td>6.40</td>
<td>1.14248</td>
<td>0.25547</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>6.90</td>
<td>1.20916</td>
<td>0.27048</td>
</tr>
<tr>
<td>Y13</td>
<td>D</td>
<td>7.05</td>
<td>0.94451</td>
<td>0.21210</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>7.20</td>
<td>1.10501</td>
<td>0.24709</td>
</tr>
<tr>
<td>Y14</td>
<td>D</td>
<td>6.60</td>
<td>1.14248</td>
<td>0.25547</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>6.45</td>
<td>1.09904</td>
<td>0.24575</td>
</tr>
<tr>
<td>Y15</td>
<td>D</td>
<td>6.85</td>
<td>0.93330</td>
<td>0.20869</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>5.80</td>
<td>1.19649</td>
<td>0.26754</td>
</tr>
<tr>
<td>Y16</td>
<td>D</td>
<td>6.60</td>
<td>1.27321</td>
<td>0.28470</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>6.00</td>
<td>1.77705</td>
<td>0.39736</td>
</tr>
</tbody>
</table>
in addition to the problems mentioned above, the Tencent Meeting also exposed the problem of less basic curriculum-related data display and limited resource format supporting display in comparison with Ding Talk, while Ding Talk exposed the cumbersome steps to register and enter the course.

According to this evaluation model, relevant suggestions for Tencent Meeting are made: (1) adds anti-inadvertent prompts for important operations, (2) enrich the number and forms of uploading and downloading of various teaching resources, (3) increase the function of automatically uploading screen recording video after the course, (4) enrich the levels and functions of communication tools, (5) add reminders to start classes to provide reminders for learners, (6) more specific and intuitive display of specific data such as participants and the duration of each student’s participation is convenient for teachers to count the basic situation of the class, (7) update display compatible technology, support more different types of software to effectively display. The suggestion for Ding Talk is to simplify the registration process, reduce the level of operation required to enter the course, and increase the pop-up prompt after login to quickly join the course.

It can be seen from the difference test results that although the user experience of the two pieces of software is significantly different in several second-level indices and one first-level index, the total scores are basically the same (there is no significant difference). In the meantime, the total scores slightly greater than the scoring standard 6 points set in the previous article, it indicates that the final evaluation results of the two pieces of software are “slightly good”, which is consistent with the result of directly selecting the preferences of the two pieces of software without using this evaluation model. Therefore, it can be seen that this evaluation model has practical application value.

B. THE STANDARDIZED PROCESS OF THE EVALUATION SYSTEM

This evaluation system can be used for single product evaluation and product comparison evaluation. The two application methods are used to construct standardized procedures for two application methods. The flowchart is as follows.

For single product evaluation, first recruit users of this product to score 16 second-level indices. After pre-testing, this questionnaire takes about 100 to 300 seconds to answer the question carefully, so sample data with a response time of less than 100 seconds should be excluded. Secondly, the mean value of each second-level index is calculated to obtain the average user rating of 16 indices, and the score of the most mean values is compared with the measurement criteria to obtain the performance level of most index items of the product. If the score of the most mean values is greater than 5, the performance of the product is acceptable on various second-level indices; if it’s equal to or less than 5, the descriptions of the corresponding indices are required to find the shortcomings of the product, and make a wide range
of changes to the product to improve the user experience. Then, focus on the index items with scores lower than the score of the most mean values. Scores on these index items are lower than scores of most index items, which proves that it does not reach the general level of the product in these indices. These are the product’s disadvantages and urgently needs improvement. Index items that are higher than the mean mode and greater than or equal to 6.5 can be regarded as the product’s advantages and temporarily maintained. After the disadvantages and mode items are improved, they can be further refined to highlight product advantages. Finally, the 16 second-level indices are weighted to get the total score of the product under this evaluation system, and the measurement criteria is compared to obtain the overall evaluation score of the product.

For product comparison evaluation, first, select competing products of the same type as the comparative research object. Recruit users of the two products to score the two products separately, and to preprocess the questionnaire data. Secondly, test the difference between the user’s ratings of the two products on each second-level index — If the two sets of user rating data are normally distributed, use the independent sample t-test method, otherwise use the independent sample nonparametric test method. Use 0.05 as the significance test standard. The index item with significance less than 0.05 proves that there is a significant difference between the two products in the score of the index item. Which is better is judged by the mean value or mean rank value: if the product’s scores are significant higher than the competing product’s, these indices needs to be highly valued and should be regarded as the primary task of product modification, combined with description items of indices and research on competing products to make greater improvements for this product to make up for its disadvantages and retain users. The index items with significance greater than 0.05 indicate that there is no significant difference in the scores of the two products on these second-level indices. It can be regarded as an opportunity for the product to improve, as a second important task for improvement. Researchers can dig more details based on the description items of indices to strive to transform into the advantage of the product. Then, analyze the scores of the four first-level indices, focusing on the advantages and disadvantages of the products in the general direction. Finally, analyze the total scores of the two products based on this evaluation system, and compare them with the measurement criteria to obtain the product’s overall level.

V. DISCUSSION
This research has four innovations: The first aspect, it starts from the current social hot issue of large-scale online learning, and cuts into it from a more macroscopic perspective, paying attention to the live courses in online courses. In addition to the content and quality of the course itself, the user experience of the platform is unearthed as another factor that affect the learning effect of learners. Based on this, a new perspective of online course evaluation “user experience of course platform” is proposed, which refines the dimension and expands the field of online course evaluation. The second aspect, through the division of user research dimensions and the combing of online course norms at home and abroad, put forward an evaluation model from four dimensions, which are “easy of operation”, “complete function” and “interface rationality” and “technical feasibility”, enriching and developing the theoretical system of online course evaluation. The third is the combination of statistical significance verification with practical application verification methods. After conducting an empirical study on the evaluation model, it is found that the evaluation model can be better applied to real case evaluations. It has reference value for the platform to discover its own deficiencies and improve user experience. The fourth is to build an analysis model for product comparison, compare the scores of the two products from bottom to top, then, the relative advantages and weaknesses can be obtained, and the optimization and improvement of the products can be made. At the same time, there are still some shortcomings in this research, such as the small sample size of the application analysis and the comparative analysis using only two pieces of software for verification. Its applicability in the application of larger samples and more software still needs further exploration and verification.

In the experimental part of this study, Ding Talk and Tencent Meeting were selected as the experimental objects. It’s because these two pieces of software have a wide range of
users in China, so it’s easy to select high-quality experimental samples. At the same time, both two pieces of software are free installation, which means it’s convenient for researchers to conduct an in-depth analysis of the differences of evaluation results from the perspective of software design, and propose reasonable improvement strategies. However, it is worth noting that cultural differences between countries and regions will affect learner behavior models, learning patterns, learning effects, etc. Teachers in different cultures will also have different focus on the design of live courses and the degree of standardization of operations. All of these will cause a difference in users experience. In addition, in countries and regions with better educational innovations, their online live course platform may have richer functions to meet more flexible and diversified teaching methods, and these may also bring more user experience evaluation dimensions and standards. Therefore, the research on the user experience of online live course platforms under different cultural backgrounds is a question worthy of discussion, which can be explored in future research.

VI. CONCLUSION

This research is based on the UE model and expands from the four dimensions: usability, ease of use, function and aesthetics. It sorts out a large number of existing research, summarizes the current research direction and context of online course platform evaluation; collects user evaluation data to obtain true user experience first-hand data. Putting the two under the framework of user experience, the user experience evaluation model of the online live course platform is constructed. At the same time, the Delphi method and AHP are used to comprehensively evaluate 16 second-level indices, obtaining the first-level index “ease of operation” and the second-level index “quickness” are the most important indices that affect the UE of the online live course platform. By scoring two representative software by users, it can be concluded that this evaluation model has good reliability and validity in statistical significance, as well as practical value in practical application. This research also constructed a model application method for product competition analysis. Based on the model application method for product competition analysis to make up for the limitations of single product evaluation and provide a more detailed quantitative method for product competition analysis. Based on two ideas of single product evaluation and product comparison evaluation, the standardized process of model application was designed to ensure the actual application effect of the model. The evaluation model and evaluation process together constitute the evaluation system of this study.

In the social situation of the COVID-19, online live courses quickly completed large-scale application and promotion. Compared with traditional classrooms, its advantages of flexibility in time and place are widely recognized. In the future, it is expected to form a blended learning model combined with traditional classrooms, which can be applied to the field of education. However, the research on ensuring the learning quality of online live courses is still in the exploratory stage. With the gradual normalization of the application of online live courses, more problems and opportunities will emerge, and the corresponding evaluation criteria will be more detailed. Under the interaction of practice and theory, learners will able to achieve better learning results.

REFERENCES

FANGHAO SONG received the D.A. degree in aesthetic cultures from the Aesthetics of Literature and Art Research Center, Shandong University, Jinan, Shandong, China, in 2017.

Since 1998, he has been a Professor and a Ph.D. Supervisor with the School of Mechanical Engineering, Shandong University, and the Director of the Institute of Modern Industrial Design. His research interests include design, interaction design, and industrial design. He was a member of the Interaction Design Committee of China Industrial Design Association in 2013 and a Senior Member of the Industrial Design Society of China Mechanical Engineering Society in 2014.

XIAOJIAO XIE received the bachelor’s degree in industrial design from the Wuhan University of Technology, Wuhan, Hubei, China, in 2018. She is currently pursuing the master’s degree with the School of Mechanical Engineering, Shandong University, Jinan, Shandong, China.

Her research interests include industrial design, human–computer interaction, and product design research.

DONG YU received the bachelor’s degree in industrial design from the University of Jinan, Jinan, Shandong, China, in 2018. He is currently pursuing the master’s degree with the School of Mechanical Engineering, Shandong University, Jinan.

His research interests include industrial design and interaction design research.

* * *