ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 132

2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2017)

Evaluating the Postgraduate Examination Database Based on Analytic Hierarchy Process

Yehan Sun[1] and Ping Zhang[1,*]

[1]College of Geo-Exploration Science & Technology, Jilin University, Changchun, China, 130061

[*]Corresponding author

***Abstract*—Analyzing and evaluating the examination database quality is very important for test management, and the results of its analysis are also essential to estimate the work of test and teaching. This research has built up an evaluation model through the method of Analytic Hierarchy Process (AHP), which included three hierarchies. The first hierarchy was the target hierarchy (i.e. evaluation of the examination database). The second hierarchy was the rule hierarchy including three indicators (i.e. benefits of students from the examination database, quality of test questions, and operability of the examination database). The third hierarchy was the sub rule hierarchy including eleven indicators (i.e. acceptance of the database, testing professional knowledge level, prompting to study, quantity of test questions, question characteristics, coverage, types, difficulty, stability, system response time, easy usability). Based on the survey data of 68 postgraduates who have used the examination database, we applied the evaluation model based on AHP, the results have showed that the examination database was user-friendly and it could meet the requirements of postgraduate tests although the quality of test questions just passed the line. As a whole, the evaluation on the quality of the examination database based on AHP was good.**

Keywords-examination database; evaluation indicators; analytic hierarchy process models; survey data

I. INTRODUCTION

As the scientific technology developing, computerized automatic test paper formation through the established examination database is an important measurement to realize a scientific and standardized examination and separation of teaching and testing [1]. The examination database cannot only alleviate the workload of set questions for teachers, but also enhance the work efficiency and eliminate the subjective factors of the examiners. However, few evaluation systems can reflect the quality of examination database in a reasonable and scientific way. The main reason is lacking of quantitative evaluation standard and rational content assignment [1-2]. In hence, it is important to set up a reasonable evaluation model for the development of the test theory and requirement.

Many scholars evaluated the question database using various evaluation indicators, for example, depth, knowledge coverage, quantity of questions, reliability, properness, discrimination, comprehensiveness, grouping of questions, customization of tests and questions, combining question databases, importing and exporting questions, analysis and archiving of results and benefits of students [1-9]. Some evaluation systems shed light on the feedback of students on exams, which make sense for the research of the evaluation system of examination database [9]. Most of the examination database make use of modular design, including the parts of students answering and teacher or offstage managers managing the examination database, whose major functional part is auto-generating paper [10-14]. Most scholars evaluated the examination database systems only by using the quality of test questions. But it is not comprehensive because the examination database is a system via computer, and the operability of the examination database should be evaluated as well. So here an evaluation system based on AHP has been set up to assess the quality of the examination database comprehensively.

Many scholars using various models for evaluating the examination database (e.g. the rough set theory [8]; AHP model [2-4]; fuzzy evaluation model with fuzzy mathematics [15]). And some researchers set up statistical analysis model based on educational measurement theory, such as the integration of classical measurement theory and item response theory [11], validation model [16], regression algorithm model [17], and improved Rasch model [18-20]. This statistical analysis model mainly analyses quantum statistical index and contrast with the experience [21]. For instance, Rasch model is a simple mathematical expression made up by the ability of the tester and the degree of difficulty, and the quality of test paper is estimated by the fitting degree of the predicted results with real numerical values [18-20], which have a shortcoming of test dependency that using responding rate and scores represent the difficulty of exam item and the ability of testers [21].

Validity model tests validity which is providing the rational evidence for writing tests [21]. Regression algorithm model estimates whether the test paper measure the ability of the testers or not, but it focuses on the validation and rationality of individual questions, which is not evaluating well [21]. Fuzzy comprehensive evaluation makes an assessment in multi-indicator for test paper, which is simple and avoid the subjective factors in evaluating, but the weight of indices is confirmed by experts, which is less systematic and rational comparing to AHP methods [7]. Combining quantitative and qualitative analysis, AHP is a systematic and hierarchical analysis method, which breaks down the complex problems into multiple factors, forms a hierarchy structure according to the relationship of these factors, compares factors by per hierarchy, gets the total sorts of relative importance degree of decision schemes, and makes decision analysis [22-23]. In spite of the limitation of roughness and subjectivity, AHP methods are superior to other models for its systematic, practicality and simplicity [22]. Thus, the evaluation model through AHP method has been constructed in this study.

This paper has established an evaluation system including three hierarchies: target hierarchy (i.e. evaluation of the examination database), rule hierarchy (i.e. benefits of students, quality of test questions and operability of the database) and sub rule hierarchy (i.e. acceptance of the database, testing professional knowledge level, prompting to study, quantity of test questions, question characteristics, coverage, types, and difficulty, stability, system response time, easy usability). Using AHP methods, the evaluation model has been established. Using the survey data of 68 postgraduates having using the examination database and running the evaluation model, the examination database has been evaluated.

II. EVALUATION MODEL BASED ON THE AHP

A. Data Source and Data Processing

On June 30, 2016, 68 postgraduate students were answered the questionnaires, who had not only finished the test on the examination database, but also had objective appraisal about the examination database system. The survey data was the data source for this study. All the questions in questionnaires were answered, so the data were valid in total. The data processing included giving the codes for each survey question answers and then putting the answers codes into the computer. There were 25 questions in the questionnaire, and each question had four answer options. Each question answers were not only divided into four levels: excellent, good, pass and fail, but also assigned the numerical values of 100, 80, 60 and 40 correspondingly.

B. Establishing the Hierarchical Structure

AHP applications are found useful when problems require considerations of both quantitative and qualitative factors. AHP decomposes the problem into small parts in order to facilitate the decision-making in the appraisal task [22]. In this study, according to the relationship between factors, we set up a hierarchical structure including three hierarchies (i.e. target hierarchy, rule hierarchy and sub rule hierarchy). The target hierarchy was on the top of the hierarchies, which was the target of the final quality of the examination database system. Below the target hierarchy, the rule hierarchy had factors (i.e. benefits of students, quality of test questions and operability of the database) to measure the target, and each factor was named a rule. When rules were too much (such as more than nine), it could be decomposed the rule hierarchy into sub rule hierarchy (i.e. acceptance of the database, testing professional knowledge level, prompting to study, quantity of test questions, question characteristics, coverage, types, and difficulty, stability, system response time, easy usability) (FIGURE I).

Previous scholars evaluated the quality of the examination database only through the quality of test questions, but our evaluation system not only consider the quality of examination evaluation, but also pay attention to the evaluation of examination database system operability, and the feedback from the students. So our evaluation system was hierarchical and scientific.

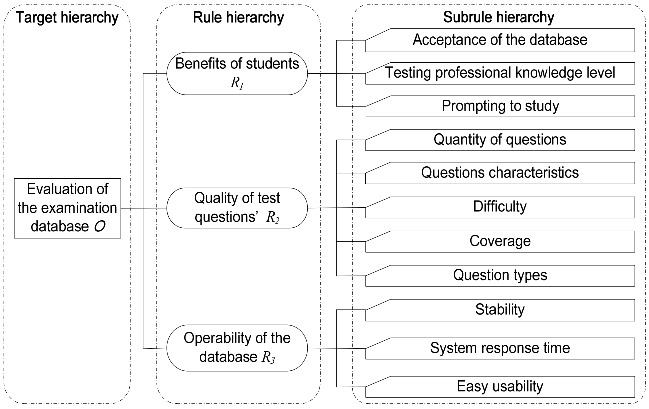
Target hierarchy
Rule hierarchy
Subrule hierarchy
Evaluation of the examination database O
Benefits of students R1
Quality of test questions' R2
Operability of the database R3
Acceptance of the database
Testing professional knowledge level
Prompting to study
Quantity of questions
Questions characteristics
Difficulty
Coverage
Question types
Stability
System response time
Easy usability

FIGURE I. THE EVALUATION MODEL

C. Construction of Judgment Matrix

The appraisal was constructed top-down using pair-wise comparisons. We made pair-wise comparisons of the same hierarchy, rather than all the factors. To make comparisons, we used a scale of numbers that indicates how many times more important one element was over another element with respect to the rule or target [23]. Table 1 exhibited the scale. Among them, the judgment matrix can be expressed as follows:

$$U = \left(u_{ij}\right)_{n\times n} \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & u_{ij} & \cdots \\ u_{n1} & u_{n1} & \cdots & u_{nn} \end{pmatrix} \tag{1}$$

Where U was a judgment matrix, u_{ij} represented the relative strength of the factor i and factor j with respect to the upper hierarchy, j is the matrix dimension, which was the number of indicators in the hierarchy. Among them, $u_{ij} > 0, u_{ij} = \frac{1}{u_{ji}}(i, j = 1,2,...,n)$. If the judgment matrix had transitivity relation as:

$$u_{ij} \cdot u_{jk} = u_{ik}(i, j, k = 1,2,...,n) \tag{2}$$

Where U was the consistency matrix. In general, the judgments u_{ij} were rarely perfect, and the transitivity relation was therefore frequently violated. In this case, the judgment matrix was said to be inconsistent [23].

Four judgment matrix U_0, U_1, U_2, U_3 were established in this study, where P was a judgment matrix of the target O (i.e. evaluation the examination database). U_1, U_2, U_3 were judgment matrices of rule R_1 (i.e. benefits of students), R_2 (i.e. quality of test questions), R_3 (i.e. operability of the database).

TABLE I. PRIORITY BASED SCALE USED IN THE AHP

Fundamental scale	Explanation
1	Equal importance
3	Moderate importance
5	Strong importance
7	Very strong
9	Extreme importance
2.4.6.8	Intermediate levels

D. Relative Weight Vetor and Consistency Test

Judgment matrix was the basic information and analysis foundation of AHP. The normalized eigenvector of the factors relative to the upper hierarchy is the relative weight vector [22]. In this study, the eigenvectors of the judgment matrix were calculated by using square root method, which were the index weight of factors in each hierarchy. The results are calculated as follows:

$$M_j = \prod_{k=1}^{n} u_{jk} \tag{3}$$

$$W_{U_i k} = \frac{\sqrt[n]{M_j}}{\sum_{k=1}^{n} \sqrt[n]{M_{j_k}}} \tag{4}$$

$$\lambda_{i\max} = \frac{1}{n}\sum_{j=1}^{n} \frac{(U_i W_{U_i})_k}{W_{U_{ij}}} \tag{5}$$

Where U_i was a judgment matrix $(i = 0,1,2,3)$. M_j was the product of the row j of U_i. W_{U_i} was normalized eigenvector of the matrix judgment U_i. $\lambda_{\max}$ was the maximum eigenvalue of U_i (TABLE II).

The consistency index (CI) was the deviation of the maximum eigenvalue ($\lambda_{\max}$) from the number of rules or sub rules (n) used in the comparison process:

$$CI = \frac{\lambda_{\max} - n}{n-1} \tag{6}$$

$$CR = \frac{CI}{RI} \tag{7}$$

Where CI was the consistency index, and CR was consistency ratio. CR was a measure called consistency ratio that gave a feedback to the decision maker on the consistency of the entered judgment matrices [22]. RI was the ratio index, which was the average of the consistency index of 500 randomly generated matrices [22].

If the CR was higher than 0.1, it was recommended to revise the comparisons in order to reduce the inconsistency. If the CR was lower than 0.1, it meant that the relative weight was rational. The CR of the judgment matrices U_i were lower than 0.1 (TABLE II), which meant that the relative weight of the rules and sub rules were rational.

TABLE II. JUDGMENT MATRIX AND CONSISTENCY TEST

U	w	$\lambda_{\max}$	CR
U_0	$(0.26,0.41,0.33)^T$	3.05	0.05
U_1	$(0.17,0.39,0.44)^T$	3.02	0.01
U_2	$(0.24,0.50,0.11,0.26,0.09)^T$	5.33	0.07
U_3	$(0.24,0.21,0.55)^T$	3.02	0.01

E. Combining Weight Vector and Consistency Test

Using the results of the relative weight vector, the combining weight vectors through top-down processing, was written as follows:

$$w_j = \sum_{j=1}^{n_i}\sum_{i=1}^{3} W_{i\,j} W_{0i} \tag{8}$$

The consistency test was written as follows:

$$CI_c = \sum_{j=1}^{n_i} w_j CI_{c\,j} \tag{9}$$

$$RI_c = \sum_{j=1}^{n_i} w_j RI_{c\,j} \tag{10}$$

$$CR_c = \frac{CI_c}{RI_c} \tag{11}$$

where, w_j was the combining weight of the j sub rule, $j = 1,2,...,11$. n_i was the dimension of the judgment matrix U_i, which was equal to the number of the factors in the matrix. W_i was the relative weight vector of U_i. CI_c was total consistency index. CR_c was total consistency ratio. RI_c was total ratio index. When CR_c was lower than 0.1, the combining weight vector was rational. In this study, $CR_c = 0.051$, which was lower than 0.1. The combining weight was rational (TABLE III).

F. The Evaluation Model

After finishing the above steps, the evaluation model was established. It can be written as follows:

$$S = \sum_{i=1}^{n} d_i w_i \tag{12}$$

where, S represented the score of the target O. d_i was the score of index in sub rule hierarchy. w was the combining weight vector. The score of each rule (i.e. R_1, R_2, R_3) can be calculated by (12) as well, and w_i should be W_i as mentioned above, d_i should be the sub rule under its rule hierarchy. In order to quantify the analysis, this study divided the evaluation scores into 4 level. If the score was between 100-85, it was excellent; if the score was between 84-70, it was good; if the score between 69-55, it was pass; if the score was less than 55, it was fail.

TABLE III. THE COMBINING WEIGHTS

Rule Hierarchy	Benefits of students	Quality of test Papers	Operability of the database	Combining Weight W
Sub rule Hierarchy	0.260	0.413	0.327	
Acceptance of the database	0.044	0.070	0.055	0.058
Testing professional knowledge level	0.101	0.160	0.127	0.134
Prompting to study	0.115	0.183	0.145	0.153
Difficulty	0.063	0.101	0.080	0.084
Coverage	0.077	0.122	0.097	0.102
Quantity of questions	0.028	0.045	0.036	0.038
Questions characteristics	0.067	0.106	0.084	0.089
Question types	0.024	0.038	0.030	0.032
Stability	0.056	0.087	0.069	0.073
System response time	0.062	0.099	0.079	0.083
Easy usability	0.143	0.227	0.180	0.190

III. RESULTS AND DISCUSSION

After running the evaluation model based on the AHP, the benefits of students are 78.72, the quality of test questions is 68.52, the operability of the database is 76.61. According to the model, the final quality of the examination database is 77.27. Therefore, the benefits of students and the operability of the database are good; the quality of test questions is pass the line. The quality of the examination database is good totally.

For the part of benefits of student, acceptance of the database was 65, testing professional knowledge level was 79.69, and prompting to study was 82.92. All in all, it shows that the students cannot adapt to the examination database test and cannot accept the examination form totally, but they can gain the positive impact from the examination form, that is prompting the students to learn and enhancing the professional skills.

For the quality of test questions part, the quantity of the tests was 70.10, the questions' characteristics was 72.10, the difficulty was 68.21, the coverage was 64.69, and the type of questions was 69.15. It shows that the quantity of the test paper was rational and the questions were practical and theoretical; the difficulty, coverage and question types are unreasonable. All the results were in line with the survey expectations.

For the operability of the examination database part, stability was 77.85, system response time was 69.65, easy usability was 86.15. The results illustrated that the interface was convenient for students taking an exam; the database ran smoothly and stably.

The quality of the examination database was good in total according to the evaluation model, which was in line with our expectations. Increasing the evaluation indicators of the AHP model properly, the evaluation can be more accurate and comprehensive.

IV. CONCLUSION

Based on the survey data from 68 postgraduate students about the examination database in June 2016, we set up the evaluation system based on the AHP, which included three hierarchies. The first hierarchy was the target hierarchy (i.e. evaluation of the examination database). The second hierarchy was the rule hierarchy including three indicators (i.e. benefits of students from the examination database, quality of test questions, and operability of the examination database).

TABLE IV. EVALUATION SCORES FOR THE EXAMINATION DATABASE

Indicator	Score	Indicator	Score	Indicator	Score
Evaluation of the examination database S	77.27	Benefits of students S_1	78.72	Acceptance of the database d_1	65.00
				Testing professional knowledge level d_2	79.69
				Prompting to study d_3	82.92
		Quality of Test questions S_2	68.52	Difficulty d_4	68.21
				Coverage d_5	64.69
				Quantity of questions d_6	70.10
				Question characteristics d_7	72.10
				Question types d_8	69.15
		Operability of the database S_3	76.61	Stability d_9	77.85
				System response time d_{10}	69.65
				Easy usability d_{11}	86.15

The third hierarchy was the sub rule hierarchy including eleven indicators (i.e. acceptance of the database, testing professional knowledge level, prompting to study, quantity of test questions, question characteristics, coverage, types, difficulty, stability, system response time, easy usability). The evaluation model can evaluate not only the quality of test questions, but also the benefits of students, and the operability of the database. Moreover, the results showed the benefits of students was 78.72, the quality of test questions was 68.52, the operability of the database was 76.61. And the general quality of the examination database was 77.27. This demonstrated that the evaluation system was reasonable and functional; the evaluation model through AHP performed as expected. All in all, the evaluation system and the evaluation model based on the AHP can be applied to other examination database evaluation.

Acknowledgement

This research was supported by Postgraduate Teaching Reform Funding from Jilin University of China. (No. 4500914A3103; 450060211402; 419100200157; 2015XYB235)

References

[1] X. Zhang, “Study of general testing storehouse of automatic generation of testing items and its evaluating system”. Journal of Shanxi Institute of Technology, Vol. 20, no. 3, pp. 88-91, 94, 2004.

[2] J. W. Lu, J. Chen, “Research on Test Paper Auto-production of Quality Evaluation Modeling Based on Analytic Hierarchy Process (AHP)”. Computer Knowledge and Technology ,Vol. 19, pp. 4457-4459, 2014.

[3] X. H. Tian, Y. Yang, “Analysis and Research on Test Question Target Evaluation Based on Test Question Bank”. Computer Knowledgeand Technology. Vol. 4, no. 20, pp. 404,417, 2007.

[4] Z. P. Zhang, X. P. Yuan, “Research on intelligent test paper construction algorithm and test paper quality evaluating system”. Science Paper Online, Vol. 2, no.10, pp. 730-734, 2007.

[5] R. J. Wu, “Study on Intelligently Composing Test Paper based on Based on Ant Colony Optimization”. Computer Simulation, Vol. 28, no.8, pp. 380-384, 2011.

[6] W. Zhang, “The Enlightenment of Cambridge Assessment ESOL for Foreign Language Tests in China”. Examinations Research, Vol. 6, pp. 71-79, 2012.

[7] Y. B. Zhang, Y. Y. Li, "The evaluation model of the quality of test papers". Vocational Education Research. Vol. 5, pp. 51-52, 2008.

[8] Z. L. Guo, Q. Z. Zhang, "Analysis on test papers quality evaluation based on rough set theory". Journal of Southwest University for Nationalities (Natural Science Edition), Vol. 38, no.5, pp. 695-700, 2012.

[9] K. Barbara, R. Mgadalena, R.Anna, B. Andrzej, K. Wojciech, "Creating digital question database: Use of self-tests in teaching medical subjects". Studies in Logic, Grammar, and Rhetoric, Vol. 43,.no.56, pp..211-227, 2015.

[10] B. H. Harris, J. L. Walsh, S. Tayyaba, D. A. Harris, D.J. Wilson, P.E. Smith, "A novel student-led approach to multiple-choice question generation and online database creation, with targeted Clinician Input". Teaching and Learning in Medicine, Vol. 27,.no.2,.pp.182-188,.2015.

[11] A. M. Yang, J. P. Wu., L. X. Wang, "Research and design of test question database management System based on the three-tier structure". WSEAS Transaction on Systems.Vol.12, no.7, pp.1473-1483, 2008.

[12] S. W. Xu, X. M. Wang, "Network test system design and implementation". Energy Procedia. Vol.17, pp. 694-699, 2012.

[13] Y. Yu, J. H. Wang, "Research on network examination system model based on Multi-AGENT". Advanced Materials Research. Vol. 566, pp. 685-690, 2012.

[14] R. Li, "Database design on intelligent test paper composing system of college English online examination". Advanced Materials Research. Vol. 846-847, pp. 1772-1775, 2013.

[15] J. J. Wang, "Group Fuzzy Analytical Hierarchy Process Based on the 0-1 Programming". Fuzzy Systems and Mathematics, Vol. 29, no.1, pp.11-125, 2015.

[16] R.S.Janice, M.Nancy, "Book Review: Measurement and Assessment in Teaching. Journal of Psychoeducational Assessment". Vol. 24, no.3, pp. 292-298, 2006.

[17] X. Huang, X. G. Hu, G. Q. Chen, “A quality evaluation system of examination questions based on the regression algorithm”. Journal of Hefei University of Technology (Natural Science), Vol. 27, no.1, pp.101, 2004.

[18] R. Huang, X. Z. Zhang, S. Y. Zhao, “Estimating the test quality with Rasch model and factor analysis”. Journal of Guizhou Normal University (Natural Sciences), Vol. 33, no. 2, pp. 36-39.

[19] Y. Y. Sheng, Y. C. Zhao, “The evaluation of university teachers' classroom teaching ability based on multidimensional Rasch model”. Higher Education Exploration, Vol. 2, pp. 70-74, 2015.

[20] S. Y. Zhao, F. X. He and Y. Liu, "The application of Rasch model in achievement test quality analysis". Educational Research and Experiment, Vol. 1, pp. 87-91, 2013.

[21] S. Z. Mei, "Exploration and analysis on technical model of test paper quality evaluation". Journal of Huaibei Coal Industry Teachers College (Natural Science Edition). Vol.4, no.4, pp. 82-87, 2015.

[22] P. Sujin, Y. Hui-Chang, H. Gyunyoung, Z. Muhammad, K. U. Rahman, "Study on Nuclear Accident Precursors Using AHP and BBN". Science and Technology of Nuclear Installations, Vol. 2014, http://dx.doi.org/10.1155/2014/206258.

[23] M. Andrecut, "Decision Making via AHP". Statistics. 2014.