Research on indicators selection model based on IRT

Mei Luo[1, a], Jianhong Jiang[2,b]

[1] College of management, Guilin University of Technology, Guilin Guangxi 541004, China

[2] College of Business, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China

[a]aubrey.luo@gmail.com, [b]aero723@gmail.com

Keywords: Indicators selection, Item Response Theory, Rasch model, three-parameter logic model.

Abstract. This paper proposed an indicators selection model that can be used when evaluator wants to ensure objective measurement and utilizes multivariate analysis invalidate but does not have enough data to support knowledge mining. The selection model is constructed based on the IRT theory. The Rasch model has been used to estimated ability characteristics of evaluators, combined with The three-parameter logic model to select indicators, unsuitable evaluators and indicators with ineffective evaluation factors were screened, the rest of indicators are closely related to evaluation object.

Introduction

The basic step of constructing an evaluation indicators system is to construct the pre-selection indicators set that based on definition of evaluated object and qualitative analysis. Indicators should be selected because of the number of the indicators and the correlation between indicators. The selection methods are different according to the evaluation object. There are no unified standards and methods to select indicators, methods are different according to the evaluation object, and can be concluded to three types, subjective methods, knowledge mining methods and mathematical statistics methods. Each method has their merits and demerits. The subjective methods, such as Delphi and Analytic Hierarchy Process have strong subjectivity. The knowledge mining methods, such as neural network need trade multi-group evaluation data, and rough set delete the significant indicator easily. Compared with the former two methods, the mathematical statistics methods are most frequently used because of the objectivity and simple operation. Though statistics methods can ensure selected indicators have low relativity between each other, it cannot reflect the complete connotation of the evaluation object and sometimes it will be invalidated. There should be a indicators selection method when these problems occurred.

Item Response Theory

Item Response Theory is a mathematical model that mainly used to describe the item properties (difficulty, discrimination, pseudo-guessing) and how the subject's ability (latent trait) affects their answer reaction. It is also known as latent trait theory or latent trait model, which is a modern psychometric theory. IRT can guide item selection and test preparation, understand the characteristics of the questionnaire and estimate the subject's ability.

IRT assume the subjects have "latent trait", latent trait is a statistical analysis idea based on observed test reactions. In tests, latent trait generally refers to the ability of potential, and it is often used test scores to estimate this potential [1]. IRT considers the subject's response in the test and the results have a special relationship with their latent trait. IRT is based on the latent trait. It established a certain item response model to describe this relationship. Item parameters established by the IRT has permanent characteristics, which means different measurement scale scores can be unified. Difficulty, discrimination and other information of the items can be intuitively viewed through a comprehensive analysis of item response curve. So IRT can guide item selection and make test to compare scores. It can be used to determine indicator system [2]. The three-parameter logic model is one of the more commonly used models in IRT, which formula is as follows:

$$P(\theta) = c + \frac{(1-c)}{1+e^{-Da(\theta-b)}} \quad (1)$$

In Eq. 1, given that D is 1.7, where p(θ) is the probability of any subjects whose ability is θ that has correct responses to answer the questions, a is discrimination, b is difficulty, and c is seudo-guessing.

By analyzing the three-parameter logic model, pseudo-guessing, difficulty and discrimination of each test item can be estimated. Pseudo-guessing represents the subjects were inclined to agree with the importance of the index, the greater value indicated subject are easy to identify important of the item without the subject's ability level. The difficulty shows that subject were less likely to agree with the importance of the item. Greater value of discrimination shows the higher degree of distinction between the subjects, lower value indicates the item can not provide enough information to the subjects.

Other common IRT model is latent trait model proposed by Rasch[3]. This model is a simplified model of IRT, the model assumes that each question has the same identification, subject's ability, item's difficulty, and subject's possibility of given the correct answer can be described by the formula. Rasch model has two requirements for the measurement: for any items, higher ability of individual should have a greater possibility to make the correct answer than the lower ability of individuals; any individual's performance on easy topics should always be better than difficult topic.

Indicators selection model based on IRT

In order to effectively select indicators, and get rid of relative or unreasonable indicators, this paper constructs the indicators selection model based on the IRT. Fig. 1 shows the selection procedure. Firstly, by analyzing questionnaire data select the evaluators who have the better response. Then identify indicators of poor response by estimating the critical value and item information, while using the three parameter logistic model to estimate the questionnaire information and screen inappropriate indicators. Finally, select experts from the better response evaluators, confirm the selected indicator and establish the amend indicators system.

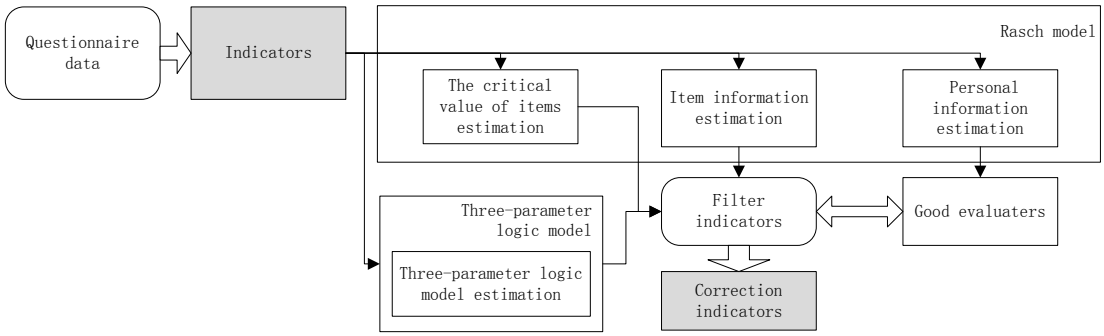

Fig. 1. Indicators selection model

Before applying IRT, these two assumptions need to verify to determine whether the material has good characteristics that suitable for IRT model analysis. Traditional unidimensionality testing methods are factor analysis and principal component analysis that test whether the data can only extract one factor or one principal component, or ratio of the first factor eigenvalues and the second factor eigenvalues greater than 2, while the variance of the first eigenvalues can be explained at least up to 20%.

For local independency test, the questionnaire of the object are the experts in related fields, the majority of questionnaire results with the following characteristics, questionnaire score depends on their industry experience and knowledge, the subject would not affected by the correctness of the answer, that means the answer state of the subject does not affect the following select results. Therefore the select probability among each question is no significant correlation, the questionnaire results meet the requirements of local independency.

The Rasch model utilizes the subject and the difficulty of the items for Parameter estimation, according to the subject's reaction on a series of items. This estimate exists fitting problem between

actual data and the desired model, hence model data fit test is necessary to examine the extent to which the data fit the theoretical models. The parameter estimation can be trusted when measurements with better fit. Rasch can fit through two forms of chi-square analysis method[4], deviation reaction (Outfit mean square, Outfit MNSQ) and information-weighted (Infit mean square, Infit MNSQ). Outfit MNSQ is the mean squared of residual, is sensitive to outliers, which is a standard to assess the quality of rating scale, the mean square value is greater than 2 indicate that the unmeasured variable information is more than measured[5], some options let the noise into the measurement process, the extreme values produce a great deal of residuals. Infit MNSQ is the weighted Outfit MNSQ(the weighted coefficient of variance), is more sensitive to the data when the difficulty of item approximate the subject's individual ability level. Acceptable ranges of Infit MNSQ and Outfit MNSQ depend largely on research purposes. Standardized forms of Outfit and Infit are Outfit ZSTD and Infit ZSTD, obey the t distribution that expectation is 0 and standard deviation is 1, which acceptable values is generally between -2 and + 2.

After the applicability analysis, use IRT to analyze the subject's information. Because subjects have different education level, business life and distribution range of industries, they also have different understanding and attitudes of the questionnaire. Therefore the subject's ability needs to be assessing to analyze the final affecting indicators.

By estimating infit and outfit values of the questionnaire data, get the subject's information, as shown in Table 1, the table lists the part of the evaluator's item response feature information. The form of outfit value is similar with infit value, outfit statistics are more sensitive to the singular values that occasionally given by the evaluator who have the better self-consistency. To improve measurement reliability, select the range of subject's MNSQ value from 0.7 to 1.2, and the range of ZSTD from -2- to +2. If the subject's MNSQ value greater than 1.2, it indicates that the variation of the subject in the assessment process is greater than the expected variability of the model, the subject has relatively poor self-consistency. If the subject's MNSQ value less than 0.7, it indicates that the variation in the assessment process, which is less than the model expected range, so the subject's score provide relatively less information regarding the ability to estimate the item value. By fit test, specific features of the subject's can be estimated. To improve the assessment quality of indicators, remove the subjects whose MNSQ value and ZSTD value out of the range, those subject's results are not satisfactory because of poor score consistency.

Table 1. Subjects screened by Rasch (partial)

Subjects	chisq	df	outfit	infit msq	outfit	infit t
P_1	11.86	48	0.243	0.245587	-4.56	-5.59
P_3	37.25	48	0.78	0.821287	-1.03	-0.89
P_{10}	46.61	48	0.978	1.007376	-0.06	0.10
P_{18}	32.571	48	0.68	0.713171	-1.61	-1.54

Table 2. Indicators screened by thresholds (partial)

Indicators	Threshold.1	Threshold.2	Threshold.3	Threshold.4
I_7	0.279721	1.076995	2.016157	2.790659
I_{23}	0.473859	1.271043	2.21107	2.784798
I_{29}	-0.05003	3.74724	7.98682	8.66091
I_{47}	1.974299	2.621574	3.711149	4.65238

After choosing apposite subjects, continue test the indicator fit. MNSQ values of most indicators are between 0.5 and 1.5, means those indicators have good item response. Abnormal performance of the indicators, which is outside this range, has been removed. Because meaning of those indicators or the questionnaire for the subjects, are hard to understand or have duplicate meaning.

In addition to the options category fitted outside classification, thresholds in category calibration can also be used to describe the characteristics of rating scales. Critical value (step calibration) is hard to estimate, because it is difficult to distinguish between two options. It is difficult to assess the real difference between ‘strongly agree’ and ‘agree’. For the relationship between reliability and choice of scale, some scholars believe that the 5-point or 7-point item has the highest reliability, and some scholars believe that the reliability and the number and classification item are mutually independent. It is nessary to describe the characteristics of rating scales for such uncertain views. For most evaluation, indicator has the higher scores means the greater impact on evaluation object, therefore thresholds are monotonically increasing.

Test results of category calibration are shown in Table 2, the different of adjacent thresholds should be greater than 1.4 and less than 5, the best range should be approximately between 1.4 and 3.5. If the difference is less than 1.4, then the level is hard to provide sufficient information, greater than 3.5 the portion information of the level will be lost, if more than 5, the amount of information will be lost more than 50% [6]. Most of the thresholds are in the optimum range.

By testing the fit and category calibration, some indicators have been chosen which may be deviation from the target characteristics or inappropriate definition or insufficient level information. However, simply delete the bad fit indicators should not be the best approach. Therefore, when Rasch model has been used to estimate the parameters, simultaneously apply the three-parameter model to estimate the item questionnaire, to assess the information of different indicators. Because of binary items the three-parameter model used, Likert scales need to be converted into a binary scales. The questionnaire data converted into 0 to1 score and apply the three-parameter model to estimate the indicators.

Through three types of parameters estimation, selected indicators are closely related to evaluation object. unsatisfactory performance of indicators their pseudo-guessing, indicators difficulty, and indicators discrimination are relatively small, indicating that evaluators cannot be convinced of its importance, therefore evaluator have focused on the negative result on the scales.

Conclusion

In this paper, the indicators selection model was constructed based on IRT. Through evaluated of the questionnaire data, estimated characteristics of their ability, combined with expert analysis, unfit evaluators and the questionnaire items with ineffective evaluation factors were screened. This selection method is suitable for ensuring objectivity in the case of less demanding on data amount, it can selects the indicators effectively when the multivariate analysis failed.

Acknowledgements

This research was supported by the Scientific Research Foundation of Guilin University of Technology (Grant No.002401003439) and the Guangxi Natural Science Foundation (Grant No. 2014GXNSFCBA118287), all support is gratefully acknowledged.

References

[1] Yue Liu, Hongyun, Comparison of MIRT linking methods for different common item designs, Acta Psychologica Sinica, 4504(2013):466-480. In Chinese.

[2] Yan Kong, Fan Zhang, Chinese citizen science research quality evaluation method based on item response theory , Science and Technology Management Research, 2009, 29(4): 280-283. In Chinese.

[3] Adam Smith, Robert Rush, Lesley Fallowfield, Galina Velikova and Michael Sharpe. Rasch fit statistics and sample size considerations for polytomous data, BMC Medical Research Methodology, 2008, 8(33): 78-90.

[4] Xiaomin Sun, Houcan Zhang. An IRT analysis of rater bias in structured interview of national civilian candidates, Acta Psychologica Sinica, 2006, 38(4): 614-625. In Chinese.

[5] J.M. Linacre. Investigating rating scale category utility. Journal of outcome measurement, 1999, 3(2): 103-122.

[6] T. Bond, C. Fox, Applying the Rasch Model: Fundamental Measurement in the Human Sciences. London:Routledge, 2007.