Scientific Research Publishing

Open Journal of Social Sciences, 2018, 6, 337-345
http://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

A New Method to Determine Validity of Student Teaching Evaluation Scores

Yanyan Sheng[1], Yougen Lou[2]

[1]Management School of Yangtze University, Jingzhou, China
[2]School of Foreign Studies, Yangtze University, Jingzhou, China
Email: louyougen@163.com

How to cite this paper: Sheng, Y.Y. and Lou, Y.G. (2018) A New Method to Determine Validity of Student Teaching Evaluation Scores. *Open Journal of Social Sciences*, **6**, 337-345.
https://doi.org/10.4236/jss.2018.64026

Received: March 18, 2018
Accepted: April 27, 2018
Published: April 30, 2018

Abstract

Based on Item Response Theory, a theory frame to determine validity of teaching evaluation scores was developed, then rater leniency and rater self consistency from Rasch model were selected to determine validity. An example illustrated how to use rater leniency and rater self consistency of Rasch model to determine validity. Data collected from Rasch model indicated that leniencies of some raters were significantly different and self consistency of some raters were not good, then some student evaluation scores were valid but other student evaluation scores were invalid. Then, contributions and limitations in the paper were discussed.

Keywords

Validity, Student Evaluations of Teaching, Rasch Model

1. Introduction

University student evaluations of teaching popularly was used to finish teaching management, such as evaluating, diagnosing, rewarding or punishing on teaching, the reason was that it was regarded as an important way to improve teaching work by some experts or managers [1]. But there was an implicit assumption that the teaching evaluation scores were valid. Validity means that evaluation scores equal to true value of teaching standard and score gape indicates true value gape of teaching standard [2]. If scores are used in teaching management, they must be valid and reflect true values of teaching standards. On the contrary, if they are invalid, they are needed to be improved even they are abandoned [3]. So, it was very important to judge on validity of student evaluation scores.

Were student evaluation score valid or not? It hadn't been confirmed in China, views in the current literature could be divided into three kinds. One view

thought that university student evaluation scores were valid, because university students were adults and they knew clearly what happed in the class, so they had abilities to evaluate their teachers appropriately. Some evidence showed that reliability among students was higher than 0.8, so did retest reliability [4]; students in some survey said that they understood evaluation standards and evaluated their teachers impersonally, so their scores were equal to their teachers' true standards [5] [6], these behaviors indicated that evaluation was valid. The opposite view was that university student evaluation scores were invalid. Students were disturbed by many factors unrelated to their teachers' teaching, such as students' emotions and attitudes to evaluation, psychological strategies on evaluation, appearance and gender of teachers, course and evaluation procedure, so their scores deflect true standards of teachers [7] [8]. The neutral point of view was that university students evaluation scores were partially valid (some students valid, other students invalid), and teachers' true teaching standards and other irrelevant factors about students, teachers and procedure were the reason [9].

Why were these views different? These researchers had different criterions to determine validity, but these criterions were strong sample dependence or weak stability. For example, some researchers used abilities of students to judge validity, but the abilities of students to correctly evaluate their teachers' teaching standards were uncertain [10], so the criterion was weakly stable. Other researchers explained some factors related to ratings by regression analysis, whether unrelated factors with teachers' abilities significantly affected ratings was used to judge validity, but their conclusions are different [11]. Samples among these papers were different, so their conclusions changed with samples; individuals in these surveys evaluated different teachers, so their conclusions changed with individuals, so these criterions were strong sample dependence. In order to obtain correct conclusion, sample dependence or weak stability must be avoided.

According to current problems, a new method should be developed to determine validity of students' scores. Many researchers divided validity into face validity, criterion validity and construct validity [12], but Rasch model of IRT theory can put them together to deeply analyze validity and had an advantage to avoid sample dependence or weak stability [13] [14], so two indicators in Rasch model were used to evaluate student ratings validity in this paper.

Two contributions were shown in this paper. In theory, validity on student evaluations of teaching by Rasch model was developed, which were different from face validity, criterion validity and construct validity in current research. In practice, validity determined by Rasch model avoided sample dependence or weak stability, results from Rasch model could be used to compare true abilities of teachers because influences of raters and scales which were unrelated to teachers' abilities had been excluded.

The paper was divided into five parts. Part one was introduction, value of the research was presented by papers about validity on student evaluations of teaching. Part two was frame of ratings validity; two indicators of Item Response

Theory were chosen to evaluate validity of student scores. Part three was methods, measure tools, participants and an example were presented. Part four was results, Rasch model was a good tool to evaluate validity, some student scores were valid, but other student scores were not valid. Part five, discussion, limitations and future research were presented.

2. Frame of Ratings Validity

Item Response Theory supposes that the subject had a latent trait which can't be observed by eyes, but it can be reflected by some special behaviors which can be observed by eyes or other assessment methods. If the relationship between these behaviors and the latent traits can be described by functions or models, we can use these behaviors to infer the latent traits. For example, accuracy rate in verbal items was presumed a function of verbal ability, a soldier's tenth ring hitting probability was supposed as the function of the pistol marksmanship, the probability that workers agreed with "work was attractive" was assumed to be a function of "satisfactory degree" [15].

Rasch model was developed from Item Response Theory. In the function of $Log\left(P_{ni,jk}/P_{ni,j(k-1)}\right)=B_n-D_i-C_j-F_k$, $P_{ni,jk}$ was a rate that the candidate N was evaluated as K levels by rater J on the item I; $P_{ni,j(k-1)}$ was a rate that the candidate N was evaluated as $K-1$ levels by rater J on the item I; B_n was competency parameters for candidates (n = 1,2,3…); D_i was difficulty parameters for task i(i = 1,2,3…); C_j was a rating leniency degree of rater j(j = 1,2,3…); F_k was a difficulty that a candidate was evaluated from the grade $K-1$ to K in the quantitative table model or in the step score model.

Through Rasch model and some software, ratings divided into four parts, candidates items and raters can be put in the same scale, and changes on rating behavior of raters can be described by parameters from Rasch model [16]. Facets software are very popular to analyze data in IRT models, it can get rater leniency degree and self consistency to describe rating behaviors and supplied some control parameters to judge raters. If values of rater leniency are in a certain range, rater leniencies is very good, ratings difference between raters come from random factors; If values of rater self consistency are in a certain range, rater self consistencies are very good, raters can distinguish candidates with higher abilities from ones with lower abilities, the scores reflect true values of the candidates.

According to meaning of validity, scores should be equal to true abilities of candidates; unrelated factors (including items and raters) with their abilities should be eliminated. Rasch model can not only separate abilities of candidates from items and raters, but also supply control parameters to explain rating behaviors of raters, these parameters can't be affected by different samples and individuals in surveys like other papers, they are certain and don't have sample dependence. So, rater leniency and self consistency of Rasch model would be good indicators to explain validity of student ratings.

3. Methods

Based on the validity of Rasch model in this paper, we collected some data in a Chinese university to present how to use the two indicators to explain validity of student ratings.

3.1. Measure Tools

Limited by difficulties in collecting data, we selected a teaching evaluation scale including 8 items from Yangtze University, the scale was presented in Table 1.

5 testers read the scale. They completed it in 25 seconds to 50 seconds. They needed at least 16 seconds to complete it, because 8 items including 16 short sentences, they took at least 1 seconds to read a sentence excluding short thinking time, it is the fastest reading speed. So, they offered a standard to choose data.

Another tool used in this study was Facets 3. 71.4 software developed by Linacre. If data fit the model, data could be input in the software, then rater leniency and self consistency can be obtained, they are two indicators of validity.

3.2. Participants Distribution

Data were collected from 196 students in three social science schools of Yangtze University. According to sampling proportion, 20% is the minimum sampling ratio, but sample size need not to be added if samples are beyond 200 [17], response rate of the questionnaire beyond 70% is very good [18] ID of 256 students were chosen by Lottery software at random to fill the questionnaire, but only 196 students would complete the questionnaire, accounting for 76.6% of 256 students. In classroom, paper questionnaires were brought to the students who were sucked, then students were told the purpose of our study to eliminate their worries and methods to fill the questionnaire, 196 questionnaires were received. All the items in the questionnaire were completed in the 20 seconds to 60 seconds, values in all the items were not beyond scope, 196 questionnaires received are effective.

Table 1. Teaching evaluation scale.

NO	Rating scale	grade
(1)	a model of virtue for others, conscientious in scholarship	0 - 15
(2)	clearness in instruction and highlight in key and difficult points	0 - 10
(3)	richness and proficiency in content	0 - 10
(4)	Fullness in lesson preparation and large information in lecture	0 - 10
(5)	formals in language expression and bi-direction exchange between of students and teachers	0 - 15
(6)	proper teaching methods and vivid presentation	0 - 15
(7)	good organization in teaching	0 - 10
(8)	applying theory well to reality and focusing on capacity building	0 - 15

3.3. Data Analysis as an Example

Data for Rasch should be connectivity. That is to say, some raters evaluate some same candidates or some candidates are evaluated by some same raters. If the consumption can't be met, data should be divided into several parts, one part for Rasch at one time. Data were divided into 12 groups, 5 to 23 raters in one group.

Because of limited length, the paper can't present all data from Rasch, so an example was presented as follow.

3.3.1. Overall Rasch Model Fit

Data for Rasch model should be unidimensional, which meant all items should be used to measure a common latent trait, data for these items can be added to compare abilities among different individuals, so unidimension was very important.

Eigen-value can be used to judge unidimension by Winstep. If the first elgen-value in raw variance unexplained by Rasch model was smaller than 3, the data was unidimensional [19]. **Table 2** was one part data from Rasch, the first elgen-value in raw variance unexplained by Rasch model was 0, far smaller than 3, so the data was unidimensional and Rasch model can be used to analyze it.

3.3.2. Rater Leniency

Rater leniency was relatively high or low among scores of raters. If we use original score to evaluate a candidate's ability and the candidate was evaluated by a rater whose leniency was high, the candidate will get a higher score. Otherwise, the candidate will get a lower score. If leniency was significantly different among raters, but we must use original scores to distinguish abilities of candidates, the scores are invalid, so rater leniency was a very important indicator to judge validity.

Data from Rasch model was presented in **Table 3**. Seven students evaluated their teachers 48 times, total score was from 1936 to 2012, average score was from 40.29 to 41.92, Fair(M) was from 38.13 to 40.04. Measure was an indicator of rater leniency from Rasch model, it was from 0.18 to −0.28. No 7 student was the most severe, but the student No 3 was the most lenient. Strata was an index that rater leniency can be divided into several parts, the value of strata was 2.99, so their rater leniency can be divided into statistically 3 significant parts. So, rater leniency among them was not from random, scores were not all valid.

3.3.3. Rater Self Consistency

Rater self consistency was an indicator of rating pattern. All raters have their own special rating patterns. For example, comparing with other raters, rater A

Table 2. Eigen-value and variation.

	Eigenvalue	Empirical
Raw variance explained by measures	9.57	82.7%
unexplained variance in 1st dimension	0	0%
unexplained variance in 2nd dimension	0	0%

Table 3. Rater leniency and consistency.

Student ID	Total score	Total count	Obsvd Average	Fair(M) Average	Measure	Infit
7	1936	48	40.33	38.13	0.18	1.52
1	1934	48	40.29	38.18	0.16	0.99
6	1942	48	40.46	38.26	0.14	0.72
4	1940	48	40.42	38.31	0.13	1.3
2	1989	48	41.44	39.42	-0.14	0.51
5	1998	48	41.63	39.68	-0.2	0.85
3	2012	48	41.92	40.04	-0.28	0.8

Note: Total scores doubled.

proffers to give higher scores to all candidates and gives different candidates larger score difference, but rater B proffers to give lower scores to all candidates and gives different candidates smaller score difference, so rater A and rater B have different patterns. As long as rating pattern of a rater isn't changed, scores equal to true abilities of candidate, score differences are appropriate to ability gapes among candidates. If rating pattern of a rater is changed, candidates with higher abilities are given lower scores, but candidates with lower abilities are given higher scores, candidates with the same ability differences sometimes are given larger score gapes, sometimes smaller gapes, these scores are don't indicate true abilities of candidates, so they are invalid.

Rasch model supplied infit value to judge validity of rater self consistency. If infit value is from 0.8 to 1.2, the rater self consistency is good and the rater fits his rating pattern. If infit value is bigger than 1.2, the rater has a noisy rating pattern, he sometimes proffers to give higher scores, sometimes proffers to give lower scores, score differences were sometimes too big to the ability gape or too small. If infit value is smaller than 0.8, the rater has a muted rating pattern, the score differences are too small to the ability gape.

Rater self consistency as an example was presented in Table 3. The infit values of two raters were bigger than 1.2, which meant the two raters have noisy rating patterns. The infit value of two raters were smaller than 0.8, which meant the two raters have muted patterns. The infit values of other three raters were from 0.8 to 1.2, which meant the three raters were appropriately consistent with themselves. So, scores of three raters were valid but others were invalid.

4. Results

We collected all the rater leniency and Rater self consistency data from Rasch model, data distribution can be used to determine validity.

4.1. Rasch Model as a Good Method to Analyze Data

Rasch model had advantages in data analysis. Firstly, it was a characteristic of wide application. If data was subjective and fit some conditions for Rasch model, data can be analyzed by Rasch model. Secondly, it can be used to select outlier items. All data in one part were put into Facets software, data in one group was not unidimensional. Why the data from the group was not unidimensional?

When the eighth item was deleted, the data were unidimensional, the phenomenon indicated that this item was not probably understood by students or latent trait in this item was different from other items, the outlier item was selected. Thirdly, it offered other information. Data were divided into four parts, estimated values of examinees describe real abilities excluding influence of raters, items and item difficulty and grade difficulty indicate how items affect raters and how raters use rating scales.

4.2. Rater Leniency from Rasch Model as an Indicator to Determine Validity

Rater leniency from Rasch model offered an indicator to determine validity by comparing scores among raters. If rater leniency was significantly different and separation value was bigger than 1, candidates would get higher scores or lower scores than their real abilities. In our survey, data were divided into 12 groups, separation values in 8 groups were bigger than 1, teachers were evaluated by different raters and the raters had several level of leniency, then scores were not regarded as their real abilities, so some scores were valid but others are not.

4.3. Rater Self Consistency from Rasch Model as an Indicator to Determine Validity

Rater self consistency from Rasch model offered an indicator to determine validity by self-comparison. If infit value was from 0.8 to 1.2, Rater self consistency was good. After collecting data from Rasch model, 75 infit values of raters are smaller than 0.8, the percentage of all raters was 38.3; 72 infit values of raters were from 0.8 to 1.2, the percentage of all raters was 36.7; 48 infit values of raters were bigger than 1.2, the percentage of all raters was 35. so, 36.7% of the scores was valid, students can use scores to distinguish different abilities of teachers, but 63.3% of the scores was not valid, the score difference was too big to the ability gape or the score difference was too small to the ability gape.

5. Discussion

The paper develops a theory frame to determine validity of teaching scores based on Item Response Theory, then rater leniency and rater self consistency from Rasch model were selected to determine validity. Some data analysis indicates that leniencies of some raters are significantly different and self consistency of some raters is not good, then some student evaluation scores are valid but other student evaluation scores are invalid.

The paper gives three contributions to current study. Firstly, a new method to determine validity of teaching evaluation scores was developed, strong sample dependence or weak stability of criterion can be solved. Secondly, frame of validity based on Item Response Theory is established, meaning of validity is clearer. Finally, two indicators from Rasch model are selected to determine validity, which can be used to analyze all subjective data if they fit some conditions.

The paper has two limitations. Firstly, we justly illustrate how to get conclu-

sions through Rasch model, but validity of students scores in other Chinese universities can't be judged because sample was limited. Secondly, the reasons why rater leniency was different and rater self consistency was good or not good have not be discussed; these problems will be discussed in the future research.

Future researchers are suggested to do some work. Firstly, sample size should be expanded by group sampling in order to determine validity of Chinese student valuations scores. Secondly, the reasons of validity should be discussed to propose measures to improve validity.

Supports

The study was supported by Hubei Education Science Program "Evaluation and promotion on validity of College Student evaluations of teaching" (2017GB024) and Humanities and Social Sciences project of Yangtze University "Research on recruitment of full time university teachers based on scientific competency measurement" (2017cszb01).

References

[1] Spooren, P. and Chrwastiaens, W. (2017) I Liked Your Course Because I Believe in (the Power of) Student Evaluations of Teaching (SET): Students' Perceptions of a Teaching Evaluation Process and Their Relationships with SET Scores. *Studies in Teaching Evaluation*, **54**, 43-49. https://doi.org/10.1016/j.stueduc.2016.12.003

[2] Spooren, P., Brockx, B. and Mortelmans, D. (2013) On the Validity of Student Evaluation of Teaching: The State of the Art. *Review of Educational Research*, **8**, 1-45. https://doi.org/10.3102/0034654313496870

[3] Herrin, C.L., Dressel, F.F. and Parsons, C.K. (1990) Application of Item Response Theory in Psychological Measurement. Hubei Education Press, 15-16.

[4] Zhou, T.Y. and Jiang, Y.B. (2015) Analyswas on Validity of University Student' Evaluations-Based on Evaluation Data from S Univesity. *Higher Education Exploration*, **7**, 78-82.

[5] Li, N. and Wang, X. (2012) Analyswas on Influence Factors University Student Evaluations of Teaching. *Population & Economics*, **3**, 27-32.

[6] He, Y.T. (2017) Ethical Perspective of Student Assessment on teaching in Chinese Universities. *Journal of Yangzhou University* (*Higher Education Study Edition*), **6**, 29-33.

[7] Zhou, J.L., Gong, F. and Qin, Y. (2017) Behavior Deviations in College Student Evaluation of Teaching and Their Relationships with Dwascipline Type, College Type and Academic Achievement Self-Evaluation. *Journal of Higher Education*, **10**, 64-74.

[8] Dai, C., Miao, L., Zhu, H., Yu, Q. and Wang, Z.X. (2017) Effects of Nonteaching Factors on Quality of College Cours. *Journal of Higher Education*, **5**, 72-80.

[9] Qiu, K. and Ye, D.Y. (2016) Evaluate on Teaching Ability or Teaching Conditions: Research on the Influence Factors of College Student's Evaluation of Teaching. *Education Science*, **2**, 33-40.

[10] Zhang, X.L. (2011) Problems, Reasons and Strategies in University Student Evaluations. *Education and Careel*, **17**, 28-29.

[11] Mei, P. and Jia, Y. (2013) Research on the Validity of Student Evaluation of Teach-

ing in China's HEWas during the past 10 Years: A Literature Review. *Modern University Education*, **4**, 29-34.

[12] Kane, M. (2000) Current Concerns in Validity Theory. *Journal of Educational Measurement*, **4**, 319-412.

[13] Smith Jr., E.V. (2000). Evidence for the Reliability of Measures and Validity of Measure Interpretation: A Rasch Measurement Perspective. *Journal of Applied Measurement*, **3**, 281-311.

[14] Fan, J.S. (2017) Validity of Self Rating Scale. *Contemporary Foreign Languages Studies*, **3**, 34-40.

[15] Herring, C.L., Dressel, F.F. and Parson, C.K., Consulting Center of Northeast Normal University (1990) Item Response Theory Application in Psychological Measurement. Hubei Education Press, Wuhan, 15.

[16] Linacre, J.M. (2011) A User's Guide to Winsteps Rasch Model Computer Programs. Chicago, 601-602.

[17] Zhang, Z.H., Zhang, J.H., Liu, Z.H., Zheng, Y. and Yang, M. (2016) Application Standard of Questionnaire Survey in Tourism Research. *Progress in Geography*, **3**, 368-375.

[18] Feng, X.T. (2007) Is High Response Rate Better? Another Understanding on Response Rates for Social Survey. *Sociology Study*, **3**, 121-135.

[19] Linacre, J.M. (2013) Old Rasch Form - Rasch on the Run: 2013 January-June. https://www.rasch.org/forum2013a.htm