Classical Test Theory and Item Response Theory in Measuring Validity of Peer-Grading in Massive Open Online Courses

Daria Kravchenko

Received in July 2018

Daria Kravchenko
Analyst, Centre for Psychometrics in eLearning, National Research University Higher School of Economics. Address: Bld. 1, 21/4 Staraya Basmannaya St, 119607 Moscow, Russian Federation. Email: dakravchenko@hse.ru

Abstract. The article presents the results of research on validity of peer-review assignments in massive open online courses within the framework of classical test theory (CTT) and item response theory (IRT). CTT-based analysis yielded data on convergent validity of the peer-review assignment, the low level of its criterion validity, and rater disagreement. IRT-based analysis revealed rater bias and established that experts largely tend to be lenient and overrate their peers. The findings are used to discuss the advantages and disadvantages of the psychometric theories in question and the opportunities for combining the two.
Keywords: massive open online courses, peer grading, classical test theory, item response theory, peer-review assignments.

DOI: 10.17323/1814-9545-2018-4-99-115

Massive open online courses (MOOCs) as a form of distance learning have been growing more and more popular among students as well as universities. In 2016, 6,850 courses from over 700 universities were available worldwide. Coursera was the largest MOOC platform in 2016 with over 23 million registered users [Shah 2016]. In 2017, over 800 universities were offering more than 9,400 MOOCs, and Coursera crossed the milestone of 30 million users and 2,700 courses [Shah 2017].

MOOCs provide open access to learning materials online, thus being able to enroll an unlimited number of students. An online course consists of video lectures, readings, hands-on activities, quizzes, and discussion forums. MOOCs are usually developed by universities and offered through providers, or platforms, such as Coursera, EdX, XuetangX, FutureLearn, Udacity, National Open Education Platform, Stepik, or Universarium. Coursera and EdX are the two largest provid-

The author thanks her academic supervisor Dmitry Abbakumov for assistance with this article. *Translated from Russian by I. Zhuchkova.*

ers of MOOCs with around 30 and 14 million registered users, respectively [Shah 2017].

When colleges started accepting MOOCs for credit on equal terms with conventional offline courses, stricter requirements began to be applied to validity and reliability of assessment tools. MOOCs most often use automated and peer grading to test knowledge and skills. Peer assessment implies that at least three students provide feedback on an answer constructed by a peer. Submissions to be evaluated are selected randomly.

Peer grading allows using open-ended assignments (e. g. essays and design projects) and has a high educational potential, as students improve their analytical skills by reviewing and commenting on their fellows' works. However, there is substantial bias in peer ratings, which are largely subjective, so their validity and credibility are questionable.

Peer assessment validity research findings are dubious. A number of works revealed a strong positive correlation between peer grades, instructor grades and tests [Kaplan, Bornet 2014; Dancey, Reidy 2017]. Other researchers found validity of peer ratings to be low due to raters' unawareness of the principles of objective assessment [Admiraal, Huisman, van de Ven 2014], their lack of expertise in the subject [Falchikov, Goldfinch 2000] and the fact that objective assessment criteria are not provided for every course [Falchikov, Goldfinch 2000].

This article explores classical test theory and item response theory as two approaches toward research on validity of peer grading in MOOCs, illustrates using two online courses how these approaches can be applied, discusses their advantages and disadvantages as well as the opportunities for combining the two.

1. Research on validity of Peer-review assignments

Psychometrics offers two approaches to studying validity of assessment tools: classical test theory and item response theory[1]. The two approaches do not exclude each other, so it is proposed to combine them.

1.1. Classical test theory

A valid test, according to Anne Anastasi, measures reliably the quality that it was designed to measure. In this article, validity of peer ratings is taken as accuracy of the scores that students award to one another. In terms of classical test theory, researchers usually measure construct and criterion validity as well as classical reliability [Anastazi, Urbina 2007].

Construct validity is one of the fundamental theoretical types of validity reflecting the degree to which the stated property is repre-

[1] The National Council on Measurement in Education: *A Professional Organization for Individuals Involved in Assessment, Evaluation, Testing*. Philadelphia, PA. http://www.ncme.org/home

sented in test results [Shmelev 2013]. This study measures convergent validity, which is understood as positive correlation between results obtained using different tools measuring the same construct. For instance, several tests are available that measure intrinsic motivation. In order to establish convergent validity, it makes sense to collect data from every test and compare the results. If results of different tests show a strong correlation, one can talk about their convergent validity.

In this study, convergent validity is measured by computing Pearson's correlation coefficients between average peer grade and test scores as well as between every individual peer's rating and test scores (since the course contains both peer-review assignments and automatically graded quizzes).

Linear correlation formula:

$$r_{xy} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2(Y-\overline{Y})^2}}, \quad (1)$$

where X, Y are observations, i. e. sample units, $\overline{X}$, $\overline{Y}$ are sample means.

Criterion validity is understood as positive correlation between outcome and an empirical criterion. Possible criteria may include, for example, final grades in the subject in which students' knowledge and skills are tested. This study uses final course grade as a criterion. To measure criterion validity, Pearson's correlation coefficients were computed (see equation (1)) between every individual peer's rating and final course grade as well as between average peer grade and final course grade.

Reliability is normally calculated as the correlation coefficient between peer and professor grades, implying that the professor is able to provide an accurate and objective assessment of students' works. In this study, classical reliability is taken as the degree of rater agreement based on a comparison of scores awarded by the raters. If all the three peers award the highest score, one can talk about rater agreement, unlike when different scores are awarded.

Rater agreement was measured using Kendall's coefficient of concordance (W):

$$W = \frac{12S}{n^2(m^3-m)}, \quad (2)$$

where S is the sum of squared deviations in all the ranks given to every object from the mean; n is the number of judges; and m is the number of objects.

1.2. Item response theory

Validity of peer ratings and the very grading procedure have been discredited a number of times [Charney 1984; Gere 1980; Huot 1990]. Even if judges specialize in the area assessed and are able to provide equipollent evaluations, interpretation of the assessment scale leaves questions: it cannot be a linear scale, and two points in one assignment cannot be equipollent to two points in another. This and other

characteristics of the CTT assessment scale make ensuring validity and reliability of peer-review assignments challenging. Item response theory (IRT) offers a metric scale with no lower limit, and the sum of all assignment difficulties is zero. This approach allows measuring assessment validity more accurately and identify bias in peer ratings.

Research in expert ratings mainly focuses on their reliability. John M. Linacre [Linacre 1989] states that true-score theory is key how variance in expert ratings and undesired judge-dependent "error" variance become a measurement challenge, so these variances should be reduced as much as possible. Another approach to expert ratings is applied in a multifaceted model designed by Linacre, who took the Rasch model as a basis. In this model, variance in expert ratings is seen as an inevitable part of the rating process; moreover, it is regarded not as a barrier but as conducive to measurement as it provides variability sufficient to estimate the probability of judge severity, item difficulty and examinee ability on a linear scale.

Adherents of the Rasch model argue for the importance of giving judges the understanding of the rating scale that they will be using to assess students [Lunz, Wright, Linacre 1990]. In fact, the use of the Rasch model eliminates the need to ensure rater agreement, since examinee ability ratings do not depend on severity of individual judges.

Within an IRT framework, the scores awarded to students in peer grading are approached as a function of three variables—examinee ability, item difficulty and judge severity or lenience [Lunz, Wright, Linacre 1990]—and students' test scores are regarded as a function of two variables, examinee ability and item difficulty.

A multifaceted Rasch model was used [Lunz, Wright, Linacre 1990]:

$$(3) \quad log\left(\frac{Pnijk}{Pnij(k-1)}\right) = Bn - Di - Cj - Fjk,$$

where *Pni* is the chance of examinee completing item *i* successfully; examinee *n* has ability *Bn* and item difficulty *Di*; and *Cj* is severity of judge *j*, who awards rating *k* to examinee *n* for item *i*.

The low validity of this model manifests itself in the high level of unexpected ratings and values differing from statistical criteria. Unexpected ratings occur when judges give ratings that differ from the ones that are expected, i. e. predicted by the model.

2. Applying classical test theory and item response theory to validity research using online courses

2.1. Classical test theory

Data from 1,308 learners (total registered users) in the course *Philosophy of Culture*[2] was analyzed. Sixty-six percent of the students were female and 34 percent were male. The age varied between 15 and 50

[2] National Research University Higher School of Economics. *Philosophy of Culture*. https://www.coursera.org/learn/filosofiya-kultury

years (M=30 years). Forty-six percent of the enrollment had an undergraduate degree (Bachelor's/Specialist's). The majority (67 percent) had been born and lived in Russia.

The focus was on students who completed the course successfully, took part in peer assessments and were rated by at least three judges. The resulting sample was thus comprised of 188 people.

Data on peer grades, test scores and final course grades in Coursera's *Philosophy of Culture* was obtained from the final report on a student survey run by the Centre for Institutional Research, Higher School of Economics.

Philosophy of Culture includes five multiple-choice quizzes and two peer-review assignments. CTT was used to analyze one peer-review assignment with assessment criteria. Students were asked to write a short essay on a particular topic. Analysis involved only data from the students whose essays were rated by at least three judges. Performance was assessed using four criteria, on a scale from 0 to 3 points for each criterion. Thus, the highest total score that could be awarded by a judge was 12.

The peer-review task was the following: "Please choose a specific moment or event in history (it may be the one analyzed by the lecturer) and find typical examples of "nature vs. culture", "nature vs. spirit" and "culture vs. spirit" dualisms. If desired, you can map them into an Euler diagram". Students were given model diagrams to perform the task. One of the criteria is described below.

Criterion 1. What elements can be found in the diagram? The elements the presence of which is assessed: name of the diagram, two examples of categories, and the dualism between them.

3 points: name of the diagram, two examples of categories, the dualism between them;
2 points: three out of four elements;
1 point: two out of four elements;
0 points: only one element.

The assignment provided examples to make assessment easier, which could also be referred to when performing the task.

Final grade was calculated as follows:

Final grade = average score for tests and peer-review assignments (performed during 7 weeks) × 0.5 + final exam score × 0.4 + active participation in the discussion forum × 0.1

Coefficients of contribution were assigned to each type of activity by the course developer. In this particular course, peer-review assignments account for 50 percent of the final grade, so it is vital to ensure that there is no bias in peer ratings.

Ratings based on the four criteria were used to estimate the score awarded by each of the raters (the median). Next, every student was awarded a score from each of the three judges. Those scores were used to calculate the coefficient of concordance. The overall score for the peer-review assignment, which contributed to the final grade, was calculated as the arithmetic mean of the three judges' ratings. Those overall scores were used to measure correlations.

2.2. Item response theory

The sample included 1,483 student works (868 in *Philosophy of Culture* and 615 in the English-taught course *Understanding Russians: Contexts of Intercultural Communication*[3]). All in all, 4,449 peer grades were obtained, as every work was rated by three judges.

The peer-review assignment in *Understanding Russians: Contexts of Intercultural Communication* also consisted in writing an essay. Students were free to choose between two topics. The essay instructions explained how to structure an essay, mentioned the keywords to use, and provided length requirements.

Judges were instructed to rate essays based on six criteria. One of the criteria implied awarding the highest score in case the essay provided an answer on how to bridge cultural gaps in cross-cultural communication, specified cultural barriers and discussed them from the perspective of cultural dimensions. Other requirements included length of 500–1000 words, novelty, and references to external sources or course resources. Depending on whether the essay featured all the required content elements, it was awarded the relevant score.

Every student has an ID, for which every action on the platform is recorded. IDs of examinees and raters were used for analysis. The data was exported to the FACETS control file, which captured student's ID, IDs of the three judges, and the scores based on six criteria. In other words, the file contained comprehensive information on the students and the grades that they received from the judges.

This analysis provided information on rater bias, i. e. extreme severity or lenience in peer ratings.

3. Peer grading validity measurement results

3.1. Classical test theory

Table 1 presents the results of convergent validity evaluation.

Correlations among tests 2, 3, 4, 5, 6 and the peer-review assignment are weak and insignificant. Multiple-choice tests and peer-review assignments differ in their content. The coefficients thus do not have to be significant, since the tasks measure knowledge in different subdomains of philosophy of culture. However, the correlation coefficient of 0.57 between test 1 and the peer-review assignment is significant, so it can be concluded that peer grading is characterized by con-

[3] National Research University Higher School of Economics. *Understanding Russians: Contexts of Intercultural Communications*. https://www.coursera.org/learn/intercultural-communication-russians

Table 1. **Correlations Between Peer-Review Assignments and Multiple-Choice Tests in MOOCs**

	Peer-review assign`ment
Test 1	0.57**
Test 2	0.04
Test 3	0.26
Test 4	0.18
Test 5	0.02
Test 6	0.01

* $p \leqslant 0.05$. ** $p \leqslant 0.01$

vergent validity as the first peer-review assignment and test 1 measure knowledge about the same constructs.

The correlation coefficient between the final grade and the peer-review assignment is 0.73 ($p \leqslant 0.01$), i.e. significantly high. It demonstrates that peer assessment contributes a lot to the final grade and has a high predictive value. One can also talk in this case about criterion validity of peer reviews in *Philosophy of Culture*, final grade serving as the evaluation criterion.

Reliability of peer grading is determined by the coefficient of concordance, which is 0.53 (p=0.000). This level of rater agreement is considered to be medium, which means that judges may differ in their opinions when it comes to criteria-based ratings. Rater disagreement may result from the lack of understanding of the assessment criteria or such criteria being inadequately defined. Kendall's coefficient of concordance is a simple and comprehensible statistic to assess agreement among raters, that is why this study only analyzes one example of a peer-review assignment.

Analysis of the ratings awarded for each criterion revealed that the raters tended to give extremely high or low grades, avoiding the middle categories of the rating scale. Research literature also describes the effects of rater severity or lenience, the findings being obtained within an IRT framework [Falchikov 1986; Orpen 1982; Ueno, Okamoto 2016; Lunz, Wright, Linacre 1990].

The most important CTT-yielded findings in research on validity of peer assessment in the specified course are as follows:

1. The assignment has a medium level of convergent validity.
2. The contribution of the peer-review assignment to the final grade

must be considered significant. The level of criterion validity is just below average.

3. The level of criterion reliability is medium, i. e. experts may disagree in their criteria-based ratings. Insufficient reliability can be explained by inaccurate wording. The four criteria proposed for assessment allowed for subjective interpretation, hence considerable disagreement among raters. Criteria should be made simpler and more accurate. Grading instructions also should be more detailed, enabling students to evaluate performance of their peers more adequately.

When using these findings, it is important to consider the study's gross limitations. First, it only analyzed peer grading in terms of a single peer-review assignment in a humanities online course. There was no chance of comparing peer reviews in this task with those in other MOOCs (whether in humanities or in science). Another essential limitation consists in the sample size of under 1,000. Such limitations can be mitigated by reproducing the study in other different MOOCs (in humanities and science) that use peer grading.

Data analysis in terms of CTT also has some limitations. In particular, it provides no possibility of assessing measurement error and rater severity. These limitations were overcome by framing the analysis into item response theory.

3.2. Item response theory

Results of evaluating peer grading validity in *Philosophy of Culture* in terms of IRT are presented in Figure 1 as graphic measures of examinees, raters and assignment (with criteria). The left-hand side of the map displays a logit scale (log probability), which is the same for all the three facets (examinees, raters, criteria). The map is scaled using asterisks, one for every four examinees/raters.

All the facets are ranked top down: examinees from the best to the worst performers, criteria from the highest to the lowest comprehensibility, and raters from the most lenient to the most severe ones.

The far right column contains the most probable indicators for each level of examinee ability. Differences in the figure are presented as a difference between the facet elements.

In this particular case, data is ranged between –8 and +10 logits. As can be seen from the rater column, 28 raters are extremely lenient, i. e. their ratings are higher than those of other judges for all the criteria. It follows from the relative position of raters and students in the map that raters tend to award higher scores than deserved: most of them nestle between 0 and +4 logits, while examinees are ranged between –2 and +2 logits, which means that the raters were not severe in assessing students' abilities. The distribution of examinee ability is skewed negatively, i. e. most of the students have an average level of ability which is lower than the ratings awarded by their peers. The distribution of rater severity is skewed positively, i. e. raters tend to be le-

Figure 1. **Data Map for Assessing Validity of Peer Grading in the Course *Philosophy of Culture***

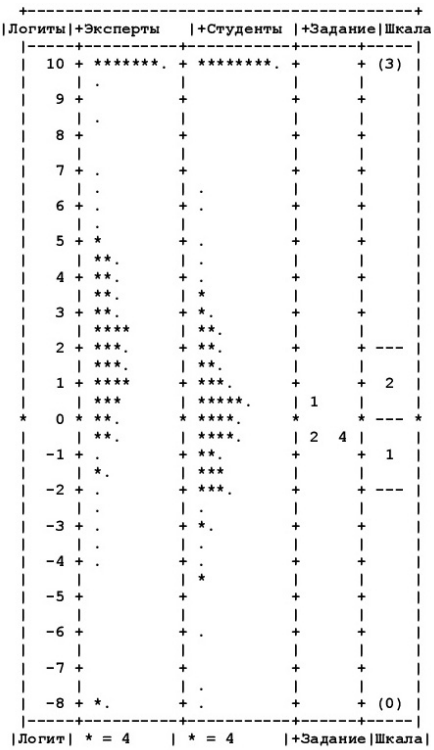
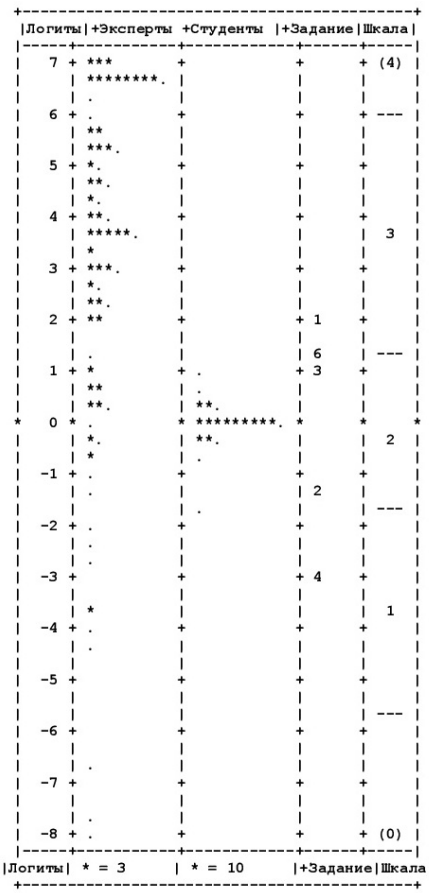


Figure 2. **Data Map for Assessing Validity of Peer Grading in *Understanding Russians: Contexts of Intercultural Communication***



nient. Such disagreement between the ratings and the levels of examinee ability indicates low validity of peer grading in this assignment.

Therefore, it was found that judges tend to rate their peers higher than deserved and the levels of examinee ability are lower than rated.

Another course was analyzed to demonstrate the opportunities of the multifaceted Rasch model in detecting rater bias.

Figure 2 presents the results of evaluating validity of peer grading in an assignment from the course *Understanding Russians: Contexts of Intercultural Communication.* The map is scaled using asterisks, one for every three examinees and every ten raters.

Data is dispersed here between –8 and +7 logits. The rater column shows that nine of the raters were the least severe.

Most raters are ranked between 0 and +6 logits and examinees between –1 and +1 logits. Obviously, the raters were not severe in this assignment either. It follows from the relative position of raters and students in the map that judges tend to award higher scores than deserved. Such disagreement between the ratings and the levels of examinee ability indicates low validity of peer grading in this assignment and thus confirms the findings obtained for the assignment in the first MOOC.

Therefore, analysis of data on the second assignment also shows that raters tend to rate their peers higher than deserved. The grades that they award do not correspond to the levels of examinee ability.

The most important IRT-yielded findings in research on validity of peer assessment are as follows:

1. In both MOOCs, ratings do not correspond to the levels of examinee ability, i. e. judges are largely lenient and tend to give higher ratings than deserved.
2. In both MOOCs, unexpected ratings are observed. Unexpected ratings occur when raters award scores that differ greatly from the ones predicted by the model. Despite the overall tendency toward leniency, there are experts who give lower ratings than deserved. When students with high levels of ability are underrated, it brings inequality into the conditions of task performance and course completion as such. We believe that such ratings should be discarded and factored out when computing the average assignment score and the final grade to maximize assessment objectivity.
   Analysis in terms of IRT also has a number of limitations:
   - It is impossible to determine whether experts overrate or underrate their peers on purpose or just award random scores;
   - The model does not make allowance for student gender, age, motivation, or time spent on a task;
   - Analysis involved only two peer-review assignments in humanities courses.
3. For these limitations to be mitigated, further research is needed that would involve rater surveys and apply other models with more parameters (gender, age, country, etc.).

4. Discussion and conclusion

Validity and reliability of peer grading in two humanities MOOC assignments was measured using two approaches, classical test theory and item response theory. Table 4 shows the advantages and disadvantages of both.

The analysis results obtained with both CTT and IRT are comparable. Still, each of the two theories has its advantages and disadvantages.

The obvious advantage of CTT is that analysis and interpretation are easier than in IRT. This method is easy to use as a quick diagnos-

Table 4. **Measuring Validity and Reliability of Peer Grading in CTT and IRT**

	CTT (Classical Test Theory)	IRT (Item Response Theory)
1	The level of reliability was assessed as medium due to analysis limitations. Level of reliability may be considered low	Individual assignment reliability was assessed separately from examinee and rater reliability. The level of reliability is high
2	The level of criterion reliability is medium. The lowest values are obtained for criteria 1 and 3. When these two are discarded from analysis, no significant increase in reliability is observed	Criterion analysis shows that scores 1 and 0 are awarded the least often. Task performance should be assessed more severely using the existing criteria. The latter probably need to be improved
3	The level of convergent validity is medium. Peer ratings contribute significantly to the final grade. The level of criterion validity is just below average	The data fits well into the model. However, there is no reason to consider the level of validity to be high, as a number of unexpected ratings and values differing from statistical criteria are revealed
4	Analysis allowed to measure rater agreement and accuracy	Analysis allowed to measure item difficulty, examinee ability and rater severity
5	The need to improve the assessment criteria was revealed	Rater bias was revealed, namely the sample's general tendency to overrate
6	Analysis is quite simple to run	Both analysis and interpretation are more complex than in CTT
7	Measurement error was not assessed	Measurement error was assessed for examinee ability and rater severity

tic method for testing peer-review assignments. However, one should keep in mind that rater agreement depends on the competencies of a specific rater sample and the focus of analysis is limited to measuring rater agreement, providing no possibility of calculating measurement error or assessing rater severity objectively. These limitations are overcome by applying item response theory. This method is more complex but it enables the researcher to spot bias, i.e. over- or underrating, in peer grading.

CTT-based quick diagnostics is an integral part of data analysis. It allows detecting the major weak points and outlining a vector for a more in-depth research using IRT. For this reason, applying a hybrid approach appears to be optimal to fine-tune and improve peer-review assignments.

References

Admiraal W., Huisman B., van de Ven M. (2014) Self- and Peer Assessment in Massive Open Online Courses. *International Journal of Higher Education*, vol. 3, no 3, pp. 110–128. DOI: 10.5430/ijhe.v3n3p119.

Anastazi A., Urbina S. (2007) *Psihologicheskoe testirovanie* [Psychological Testing]. Saint Petersburg: Piter.

Charney D. (1984) The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview. *Research in the Teaching of English*, vol. 18, no 1, pp. 65–81.

Dancey C. P., Reidy J. (2017) *Statistics without Maths for Psychology*. Upper Saddle River: Pearson.

Falchikov N. (1986) Product Comparisons and Process Benefits of Peer Group and Self-Assessments. *Assessment and Evaluation in Higher Education*, vol. 11, no 2, pp. 146–166. DOI: 10.1080/0260293860110206

Falchikov N., Goldfinch J. (2000) Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research*, vol. 70, no 3, pp. 287–322.

Gere A. R. (1980) Written Composition: Toward a Theory of Evaluation. *College English*, vol. 42, no 1, pp. 44–48, 53–58.

Huot B. (1990) The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research*, vol. 60, no 2, pp. 237–263.

Kaplan F., Bornet C. (2014) A Preparatory Analysis of Peer-Grading for a Digital Humanities MOOC. *Digital Humanities: Book of Abstracts.* Lausanne: University of Lausanne, pp. 227–229.

Linacre J. M. (1989) *Many-Faceted Rasch Measurement*. Chicago, IL: MESA.

Lunz M. E., Wright B. D., Linacre J. M. (1990) Measuring the Impact of Judge Severity on Examination Scores. *Applied Measurement in Education*, vol. 3, no 4, pp. 331–345.

Orpen C. (1982) Student versus Lecturer Assessment of Learning. *Higher Education*, vol. 11, no 5, pp. 567–572.

Shah D. (2016) *Monetization over Massiveness: Breaking down MOOCs by the Numbers in 2016*. Available at: https://www.edsurge.com/news/2016–12–29-monetization-over-massiveness-breaking-down-moocs-by-the-numbers-in-2016 (accessed 10 October 2018).

Shah D. (2017) *Coursera's 2017: Year in Review.* Available at: https://www.class-central.com/report/coursera-2017-year-review/ (accessed 10 October 2018).

Shah D. (2018) *A Product at Every Price: A Review of MOOC Stats and Trends in 2017*. Available at: https://www.class-central.com/report/moocs-stats-and-trends-2017/ (accessed 10 October 2018).

Shmelev A. G. (2013) *Prakticheskaja testologija. Testirovanie v obrazovanii, prikladnoj psihologii i upravlenii personalom* [Practical test. Testing in education, applied psychology and human resource management]. Moscow: Maska

Ueno M., Okamoto T. (2016) Item Response Theory for Peer Assessment. *IEEE Transactions on Learning Technologies*, vol. 9, no 2, pp. 157–170.

Wright B. D., Masters G. N. (1982) *Rating Scale Analysis: Rasch Measurement*. Chicago: Mesa.