Journal on Efficiency and Responsibility in Education and Science
ISSN: 1803-1617, doi: 10.7160/eriesj.2012.050403

ERIES Journal

STATISTICAL EVALUATION OF EXAMINATION TESTS IN MATHEMATICS FOR ECONOMISTS

Nikola Kaspříková

University of Economics, Prague
nikola.kasprikova@vse.cz

ARTICLE INFO

Article type
Full research paper

doi: 10.7160/eriesj.2012.050403

Article history
Received: August 8, 2012
Received in revised form: November 7, 2012
Accepted: November 9, 2012
Available on-line: December 30, 2012

Abstract

Examination results are rather important for many students with regard to their future profession development. Results of exams should be carefully inspected by the teachers to help improve design and evaluation of tests and education process in general. Analysis of examination papers in mathematics taken by students of basic mathematics course at University of Economics in Prague is reported.

The first issue addressed is identification of significant dependencies between performance in particular problem areas covered in the test and also between particular items and total score in test or ability level as a latent trait. The assessment is first performed with Spearman correlation coefficient, items in the test are then evaluated within Item Response Theory framework. The second analytical task addressed is a search for groups of students who are similar with respect to performance in test. Cluster analysis is performed using partitioning around medoids method and final model selection is made according to average silhouette width. Results of clustering, which may be also considered in connection with setting of the minimum score for passing the exam, show that two groups of students can be identified. The group which may be called „well-performers“ is the more clearly defined one.

Key Words

Education research, examinations, IRT, mathematics for economists, segmentation

Kaspříková, N. (2012) "Statistical Evaluation of Examination Tests in Mathematics for Economists", Journal on Efficiency and Responsibility in Education and Science, Vol. 5, No. 4, pp. 203-211, ISSN 1803-1617, [on-line] www.eriesjournal.com/_papers/article_180.pdf [2012-12-30]. doi: 10.7160/eriesj.2012.050403

Introduction

A lot of data is quite naturally collected within process of teaching. It is not only data obtained by automated data registration related to usage of modern web-based e-learning systems, even though this is nowadays probably the most data-intensive area in education and also heavily utilised for analyses (for overview of data mining methods related to analysis of e-learning systems see e. g. Romero and Ventura (2007). Another (and more traditional) sources of data are sample surveys and data about examinations. A number of analyses are being performed and many descriptive or prediction models are built using data related to education process, see Kotsiantis, Pierrakeas and Pintelas (2004) for details.

Evaluation of performance of students is an important part of education process and it may also provide useful information for assessment of the effects of education process and results of an analysis may suggest how to improve the teaching process. Data about tests represent valuable source of information if used for further analysis.

Data analysis can provide answers to questions such as:

1. Are there any dependencies between performance in test and other, e.g. behavioural or demographic characteristics of students?
2. What are the dependencies between performance in particular problem areas covered in the test and what are the relations of items to total score or even to assumed latent ability trait?
3. Are there any groups of students who are similar with respect to performance in test?

Analytical question (1) is rather basic one and may easily be answered using standard statistical framework of hypothesis testing. Evaluation of performance in basic mathematics course with respect to gender and major field of study is discussed in (Kaspříková, 2012).

Analysis of relations between particular problem areas covered in test and correlation of items with total test score was discussed in (Kaspříková, 2011) and the analysis is further extended in this paper to cover characteristics of items with respect to assumed ability level.

Analytical task (3) may be solved using clustering methods and the results may be used e.g. for setting optimal cut-off points for passing the exam (see Sireci and Robin (1999)).

This paper addresses just subset of possible analytical questions. Many other questions may arise within the scope of analysis of exams, such as a search for common factors influencing performance in tests - there is usually a general factor which may be interpreted as general ability to solve tasks. Another common analytical task is examination of reliability of tests and analysis of sources of variability, including investigation of effect of examiners - see Cronbach (2004), Holland and Hoskens (2003) and Harik et al. (2009).

This paper shows application of both basic data analysis techniques and more advanced data mining tools for analysis of tests taken by students of basic "all-in-one" mathematics course at University of Economics in Prague, with the aim to get answers to the questions (2) and (3) stated above. We will investigate if there are some significant dependencies between test items and total score or ability level. Then a cluster analysis will be performed to learn if there are some groups of students homogeneous with respect to performance in test and if so, what is the most suitable value for cut-off point for passing the exam.

Materials and Methods

Data description

Sample of N=45 test papers in mathematics is available for analysis. For all the tests the following structure holds: there are 8 tasks in test, covering the topics of basic lectures of mathematics for economists. Tests included namely one item which can be classified as basic linear algebra calculations (henceforward denoted LA), then matrix algebra task (denoted MA), limit calculation (LF), item focused on rather straightforward derivatives application (D1), item including a more difficult application of derivatives (D2), integral calculation (IN), optimization of a function of two variables (F2) and solving a differential equation (DE). If an item was perfectly solved by the student, it was evaluated by 5 points, otherwise an evaluation between 0 and 5 points could by assigned for a partially correct solution. Following the evaluation of particular items, an overall score (denoted SC) was calculated as sum of all eight evaluations for particular test items, i.e. SC = LA + MA + LF + D1 + D2+ IN + F2 + DE, thus giving nine variables for analysis in all.

Basic univariate analysis, data preparation and statistical methods

Basic univariate sample statistical summary characteristics of variables investigated are given in Table 1.

	LA	MA	LF	D1	D2	F2	IN	DE	SC
Min	0	0	0	0	0	0	0	0	0
Median	5	5	5	5	3	2	3	2.5	27
Mean	3.9	3.8	3.5	3.5	2.8	2.4	2.7	2.6	25.3
Max	5	5	5	5	5	5	5	5	40

Table 1: Basic sample summary statistics

The distribution of score assigned for particular items and total score is illustrated in Figure 1.

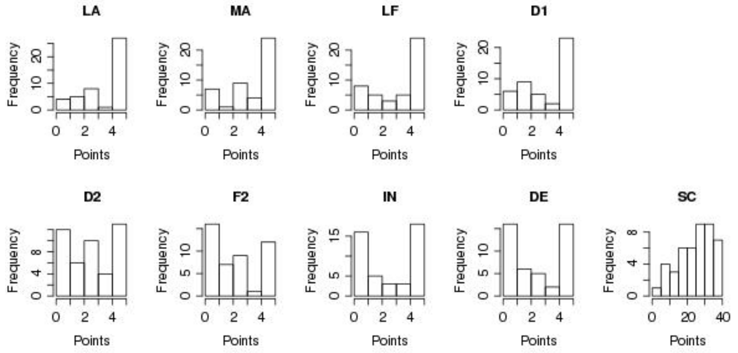

Figure 1: Histograms of scores for particular test items and overall test score

Based on histograms it can be observed that obviously the variables do not follow normal distribution and indeed with formal tests of normality based on sample kurtosis and sample skewness, except for D1, D2, F2 and SC variable the null hypothesis is rejected at 0.05 significance level. It is more often the case that a student either solves the task completely (and obtains 5 points for the item) or gets 0 score for the item, than it is the case that there is a partial solution of the task.

Non-normality of variables to some extent limits the range of statistical methods applicable to this data.

For general multivariate analysis of dependencies between overall score and particular problem areas covered in test we first perform simple bivariate analysis. Dependency between every two variables is assessed with Spearman correlation coefficient, i.e. the analysis is actually based on ranks of original variables. We are considering multiple evaluations (there are 9*8/2=36 pairs of variables) in the analysis and significance level of 0.001 is used for assessment of statistical significance.

For data analysis within item response theory (IRT) framework we convert each variable to binary using value 3 as cut-off point – assigning value 1 if the original value is at least 3 points and assigning 0 otherwise. This way of recoding is supported by rather high frequencies of 0 and 5 points (see histograms in Figure 1) and provides us with dichotomous variables which have clear interpretation – we get 1 if the performance of the student in particular item was "good" and we get 0 if the performance was "bad". Such simplified classification, if applied directly when evaluating the tests, could be also easier for the teacher in comparison with assigning 0 to 5 points for every task. We use two-parameter logistic model described and implemented by Rizopoulos (2006). In this model, probability of correct answer in particular item *i* by the student is given by the ability level of the student and by item characteristics. It holds

$$\log\left(\frac{p_i}{1-p_i}\right) = a_i + z_m b_i$$

where

p_i is probability of correct answer in particular item *i* by the student,

a_i is the easiness parameter of the item,

b_i is discrimination parameter of the item (showing how well the item discriminates between students with low and high level of latent ability trait),

z_m is ability level of the student.

Answer in particular item is manifest variable and ability level is latent variable in this model and values of variable are estimated for every student. The model which we have chosen allows that items may be different with respect to their difficulty and discrimination power. We do not use three parameter model with guessing parameter, because the tasks in the test can not be solved by pure guessing. Item characteristic curve and item information curve (see Rizopoulos (2006)) for particular items will be evaluated in the analysis. Item characteristic curve (ICC) shows probability of correct answer as a function of ability level and item information curve (IIC) can be used for assessment of how much useful the item in test is for various levels of ability trait.

Cluster analysis is performed using partitioning around medoids method, implemented in cluster package in R computing environment (for detailed description of the method see Struyf, Hubert and Rousseeuw (1997)). Partitioning around medoids has a couple of advantageous properties, one of them is no need for initial guess of the cluster centres. Another property which makes this method suitable to be used in our analysis is that it works with medoids, which are real representatives, as opposed to centres (used e.g. in standard k-means clustering), which may often be artificial. Silhouette width, described by Rousseeuw (1987), is used for formal evaluation of quality of clustering model and for model selection.

R environment for statistical computing (Hornik and Leisch (2004) and R Development Core Team (2012)) is used for the calculations.

Regarding elementary characteristics of the test, Cronbach's alpha, commonly used for assessment of reliability, is 0.81; difficulty of the test is 0.61 and discrimination value is 0.63.

Results

Correlation analysis

Dependencies between items and overall score are depicted in Figure 2. There is an edge between two nodes if the null hypothesis of zero correlation coefficient has been rejected. Edges in the graph are then evaluated with sample Spearman correlation coefficient.

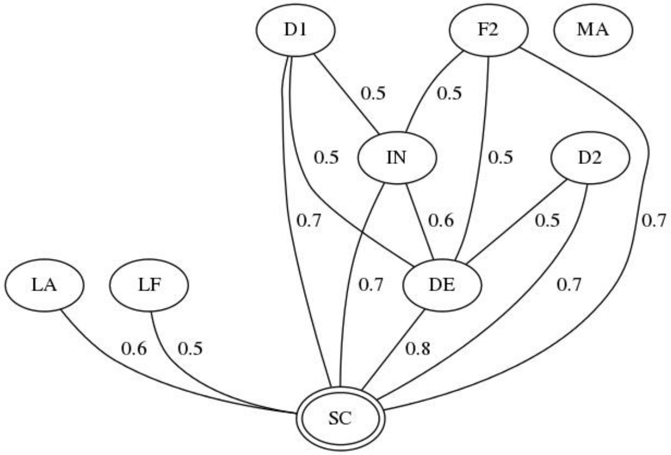

Figure 2: Dependency graph for test scores with Spearman correlation coefficient

IRT analysis

Results obtained for particular items using two-parameter logistic IRT model are shown in Figure 3 and Figure 4, which show item characteristic curves and item information curves respectively. Three-parameter logistic IRT model has also been fitted, but it did not bring better fit compared with the two-parameter model at 0.05 significance level so it is not further reported here.

Cluster analysis

Two groups of students were identified in cluster analysis of data about performance in test. Number of clusters was selected according to the highest average silhouette width. Number of students included in particular clusters was 25 for Group 1 and 20 for Group 2 respectively. For basic characteristics of the resulting clusters see Table 2. Average silhouette width of the resulting clustering model is 0.26.

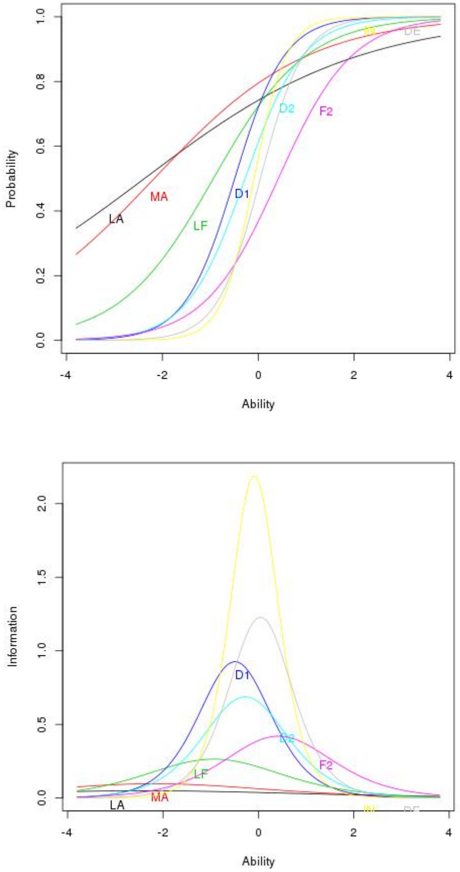

Figure 3 and 4: Item characteristics curves and Item information curves

	Number of students	Mean score in test	Average silhouette width
Group 1	25	18.4	0.14
Group 2	20	33.8	0.4

Table 2: Profiles of groups resulting from cluster analysis

Average total score SC (this variable was not included as input variable to clustering algorithm) in Group 1 is considerably lower (18.4) than in Group 2 (33.8). Density estimates of total score by cluster are illustrated in Figure 5. Average score for every particular test item was higher for Group 2 than for Group 1, so Group 2 can be called "well-performers". The cut-off value for total score to distinguish between clusters is at some 26 points, which represents 58 % of score which can be reached in the test.

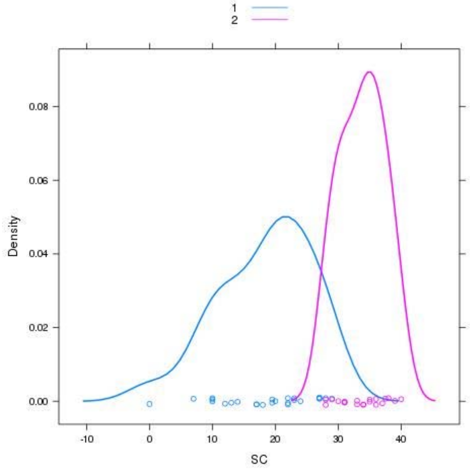

Figure 5: Density estimates of score by cluster, Group 2 on the right-hand side

Discussion

All the dependencies shown in Figure 2 are positive, i.e. with positive correlation coefficient. This is in agreement with our expectation and may be a demonstration of a general ability concept referred to in other analyses and captured formally in IRT theory (Sheng and Wikle 2008), as if a student performs well in one part of mathematics, it is likely that performance in solving problems related to another part of mathematics will be good as well.

Clearly (and as could have been expected) SC represents the node with highest degree (7 connections) in the graph in Figure 2, followed by DE (5 connections) and IN (4 connections). On the other hand, matrix algebra task seems to be rather independent of other items in test and even of overall score. LA and LF are dependent just on overall score.

The highest dependency, according to the correlation coefficient value, was observed between overall score and a test item focused on differential equations. Rather high correlation was observed between overall score and D2, D1, F2, IN. High correlation is also between DE and D1.

Evaluation of items in the test within the IRT framework shows that F2 task may be considered as rather difficult one (see ICC curves in Figure 3). According to item information curves (see Figure 4) it seems that most items in the test are most useful for assessment of students with medium level of ability. F2 task may be useful for evaluation of students with higher ability level, performance in MA and LA tasks does not seem to reflect ability level well and perhaps these tasks should be removed from the test. IN or DE task may be used for quick assessment of students with medium ability level. In case that there was a need to better distinguish ability of students with rather high ability level or similarly for increasing information value of the test on ability levels of low ability students, the test would have to be redesigned. But taking into consideration the fact that the test is designed for a course in basic mathematics for students whose major field of study is economics, it may be appropriate that the test is calibrated for medium ability level students.

Regarding cluster analysis results, average silhouette width of the resulting clustering model is 0.26. This value suggests that assumption about existence of the clustering structure in the data is not groundless, even though the classification structure is not very strong.

Average silhouette width was higher in Group 2 (0.4) than in Group 1 (0.14), suggesting that Group 2 represents the more clearly defined group.

The cut-off value discovered in cluster analysis is quite close to the value which is usually set for passing the exam (60%), so the analysis has confirmed that the 60% cut-off is not set at some groundless level and that such value is based on some sort of a natural split. Note that the cut-off value for passing the exam should be communicated to students in advance, so there is definitely a need for some sufficiently stable cut-off value. It may not be a good idea to use different cut-off values based on every particular test, even though such cut-off values may have better characteristics form analytical point of view.

Results of cluster analysis can be compared with clustering of another, larger sample of 110 tests, described in (Kaspříková, 2012) and coming from another academic period, after slight changes in curriculum of the course. The test consisted of two parts. The first part (mid-term test) covered topics in linear algebra and introductory topics from mathematical analysis (limit calculation and derivatives). The second part (final test) was taken by students at the end of the term and this part covered topics of mid-term test and a couple of other topics form mathematical analysis, namely integral calculation, optimization task and a differential equation. There was not any requirement regarding limit for minimal number of points obtained in the first test to be entitled to take the second part of the test and all students have taken both parts of the test. The best clustering model reached average silhouette width 0.6 and it was again a two clusters solution, in which the group with higher mean score was the more clearly defined one. This may be considered as some sort of validation of results described in this paper and it suggests that the obtained clustering structure may hold over time.

Conclusion

It was shown that statistical analysis of examination papers can provide insights into structure of students' performance and it can also provide some assessment regarding design of the test. Results of our analysis addressing tests in basic mathematics for economists course suggest that for a quick assessment of student performance in the course just subset of tasks in test could be used, focusing mostly on comparatively advanced parts of the course, which are differential equations or integral calculation and optimization of functions of two variables. Also a finding of no strong dependence of performance in matrix algebra task (both on performance in other items and on assumed latent ability trait) is interesting and may lead to considerations resulting in removing this task from the test. But these results should be validated in future research after collecting larger data sample. Evaluation using two-parameter logistic IRT model has shown that optimization of functions of two variables may be considered as rather difficult task in the test and that the tasks in the test are mostly useful for assessment of performance of students with medium ability level.

Results of cluster analysis suggest that two groups of students, covering approximately 60% and 40% of students and with different level of performance in test can be identified. The smaller group, which may be called "well-performers", seems to be the more clearly defined one. The suggested cut-off value is approximately 58% points, which is close to the usually used value, which is 60%.

References

Cronbach, L.J. (2004) 'My Current Thoughts on Coefficient Alpha and Successor Procedures', *Educational and Psychological Measurement*, vol. 64, no. 3, pp. 391-418.

Harik, P., Clauser, B.E., Grabovsky, I., Nungester, R.J., Swanson, D., Nandakumar, R. (2009) 'An Examination of Rater Drift Within a Generalizability Theory Framework', *Journal of Educational Measurement*, 46, Issue: 1, pp. 43-58.

Holland, P.W., Hoskens, M. (2003) 'Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test', *Psychometrika*, 68, Issue: 1, pp. 123-149.

Hornik, K. and Leisch, F. (2004) 'R version 2.1.0', *Computational Statistics*, vol. 20, no. 2, pp. 197-202.

Kaspříková, N. (2011) 'Multivariate Analysis of Examination Papers', *Efficiency and Responsibility in Education, Proceedings of the 8th International Conference*, Prague, pp. 120–127.

Kaspříková, N. (2012) 'Data analysis of students' performance', *Efficiency and Responsibility in Education, Proceedings of the 9th International Conference*, Prague, pp. 213–218.

Kotsiantis, S., Pierrakeas, C., Pintelas, P. (2004) 'Predicting students' performance in distance learning using machine learning techniques', *Applied Artificial Intelligence*, 18, Issue: 5, pp. 411-426.

R Development Core Team (2012) R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Rizopoulos, D. (2006) 'ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses', *Journal of Statistical Software*, 17, Issue: 5.

Romero, C., Ventura, S. (2007) 'Educational data mining: A survey from 1995 to 2005', *Expert Systems with Applications*, 33, pp. 135-146.

Rousseeuw, P.J. (1987) 'Silhouettes - A Graphical Aid to the Interpretation and Validation of Cluster-analysis', *Journal of Computational and applied Mathematics*, 20, pp. 53-65.

Sheng, Y. Y. and Wikle, C. K. (2008) 'Bayesian multidimensional IRT models with a hierarchical structure', *Educational and Psychological Measurement*, vol. 68, no. 3, pp. 413-430.

Sireci, S.G., Robin, F. (1999) 'Using cluster analysis to facilitate standard setting', *Applied Measurement in Education*, 12, Issue: 3, pp. 301-325.

Struyf A., Hubert M., Rousseeuw, P.J. (1997) 'Integrating Robust Clustering Techniques in S-PLUS', *Computational Statistics and Data Analysis*, 26, pp. 17-37.