Results and Comparison of Different Complementary Assessment Methods of Science Learning Outcome.

Mikael Lönn, Ann Mutvei, Jan-Eric Mattsson
Södertörn University (Sweden)
mikael.lonn@sh.se, ann.mutvei@sh.se, jan-eric.mattsson@sh.se

Abstract

To assess the quality of different aspects of the learning outcomes in relation to knowledge requirements as results of teaching several assessment methods have to be used. For most teachers it is also obvious that students differ in their ability to demonstrate the learning outcome depending on the assessment method used. In order to compare different assessment methods of the learning outcome of pre-school teacher students' different types of tasks were evaluated and compared in order to identify the potential of each method to give the students fair chances of showing their skills. Thus, assessments based on multiple choice questionnaires of different types, long answer questions, practical laboratory experiments, experiment construction and the students ability to evaluate experiment plans were compared. Having Swedish as mother tongue also was included as an explanatory variable since we suspected that some of the assessment methods in reality rather evaluates the linguistic skills in interpreting texts rather than evaluating the content knowledge of the subject. The results for each student when different methods were used were compared in order to evaluate if some of the methods for assessment gave similar results or if the methods induced differences in the results for the same student. We use ordination techniques to assess and visualize main trends in the data and linear models and classification trees to evaluate specific associations. There is correlation between results from several assessment methods, there are positive correlation between combinations of results from long answers, experiment and experiment construction, meaning students who showed good results with one method did so also with the others - but in some comparisons like long answer questions and multiple choice questions good results were independent of each other. There was a negative effect of having a non-Swedish mother tongue on the results in multiple choice questionnaires, but a positive effect of a non-Swedish mother tongue on the combined scores on experimental construction and experiment. Linear models show that good achievements in experimental construction are explained by high summed scores of Doll´s criteria, the four R's richness, recursion, relations, and rigor.

1. Introduction

A primary school teacher class was studied during their second semester which comprised of natural science education. Learning was assessed in several ways, here we want to compare results from written exams with questions of several types - multiple choice questions, long answer questions, experimental skills, experiment construction and evaluation of experimental plans. We also wanted to examine the effect of having a non-Swedish mother tongue, results from a nursing class in the United States suggest that linguistic problems are explaining variation in test scores using multiple choice exams [1]

There are specialized statistics used for studies of learning, e.g. used by e Silva et al. [2] that manipulate the raw data prior to analysis. The problem with those approaches is that it is difficult to go back from the statistical analysis to the kind of data collected, e.g. by making predictions from the models. Instead we used several general purpose statistical methods that we earlier have applied in ecology [3] and in medicine [4].

Our research questions are:

Are the results from different kinds of assessments uncorrelated or can they be grouped?

How does the students mother tongue explain performance in different assessments?

Can results from different kinds of assessments be explained by properties of the students, measured using Doll´s R categories [5]

2. Methods

2.1 Data

The data is collected from pre-service teacher training classes. Data on scores from examinations using multiple choice questions, long answer questions, experimental work, and construction of a experimental scientific test and evaluation of other student's experimental scientific test are collected from one classes in their second semester, in total 47 students. The scores from multiple choice questions were maximum 17 (mean 9.8), for the long answer questions 10 (7.3), for the experiment 9 (7.2), for the experiment construction 10 (7.1) and the experiment plan evaluation 5 (4.3). All students were classified into mother tongue categories "Swedish" and "other".
During their second semester the students did a fieldwork with repeated observations of the same place in nature or in town. Those places were revisited in the fifth semester and the student reflections made after this visit were analyzed. The quality of the reflections in each category was quantified using Doll´s R: richness, recursion, relations, and rigor.[5] and the summed scores were used in analyses (max = 55, mean = 31.5)

2.1 Statistical methods

We used the statistical package R 2.15.2 [6] and within the R environment applied ordination methods from the package vegan [7], and the package effects [8] to make plots of predictions from linear models. The classification tree was mad using the R-package rpart [9]. Linear models were constructed for each of the assessment scores using the summed Doll´s R score as explanatory variable.

3. Results

First we made a Principle component analysis (PCA) using the 5 assessment methods to form the ordination. Each point is a student and those with similar profiles are close to each other. The arrows point in the direction of the highest scores for each assessment.

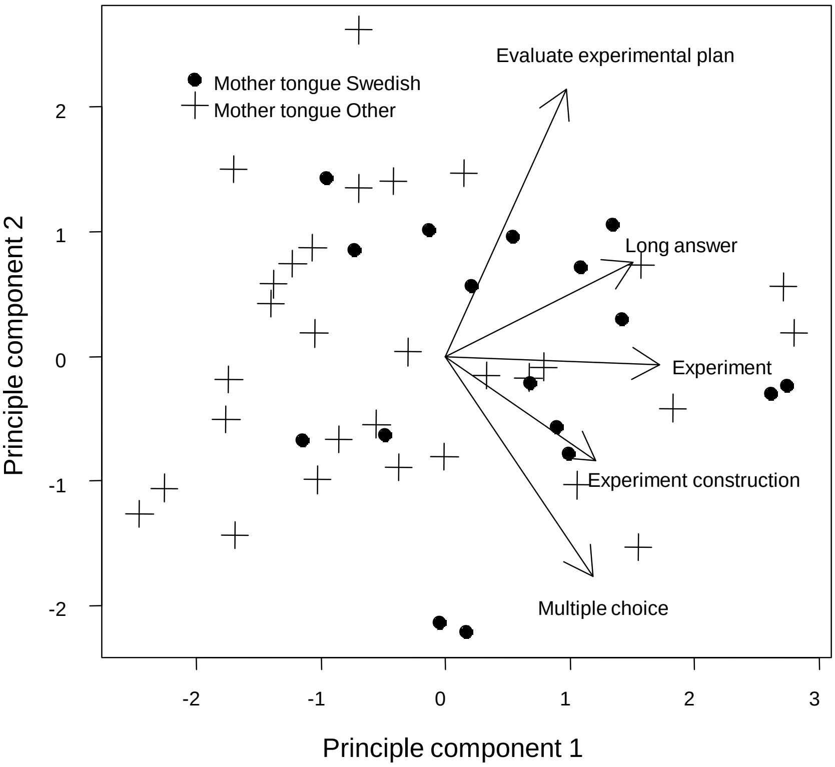

Fig. 1. *Results from a Principle component analysis using the 5 assessment methods to form the ordination*.

Student with similar score profiles group together in the graph. The arrows shows the main trends in the ordination, e.g., Experiment students to the right in graph have high scores and students to the left have low scores. Arrows that go in the same direction indicate that students that have high scores in one respect also have it in the other respect, e.g., multiple choice questions and experimental construction. Arrows that are perpendicular to each other indicate that scores are independent, e.g., high scores in multiple choice questions are independent on high scores in long answer questions. Mother tongue of the students are just plotted and do not influence the ordination.

To evaluate the effect of mother tongue we made a classification tree. The student are divided into groups that are further subdivided according to the explanatory variable that explains most of the variation in each division. When a group is separated out it is no longer part of the subsequent analysis. The most important variables are highest up in the tree and the score limit given is for going to the left branch in the tree. We can for example see that a group of students with a non-Swedish mother tongue get low scores on multiple choice question assessments, but also that there is a group of students with non-Swedish mother tongue that higher scores than those with Swedish as a mother tongue on experimental construction and experiment.

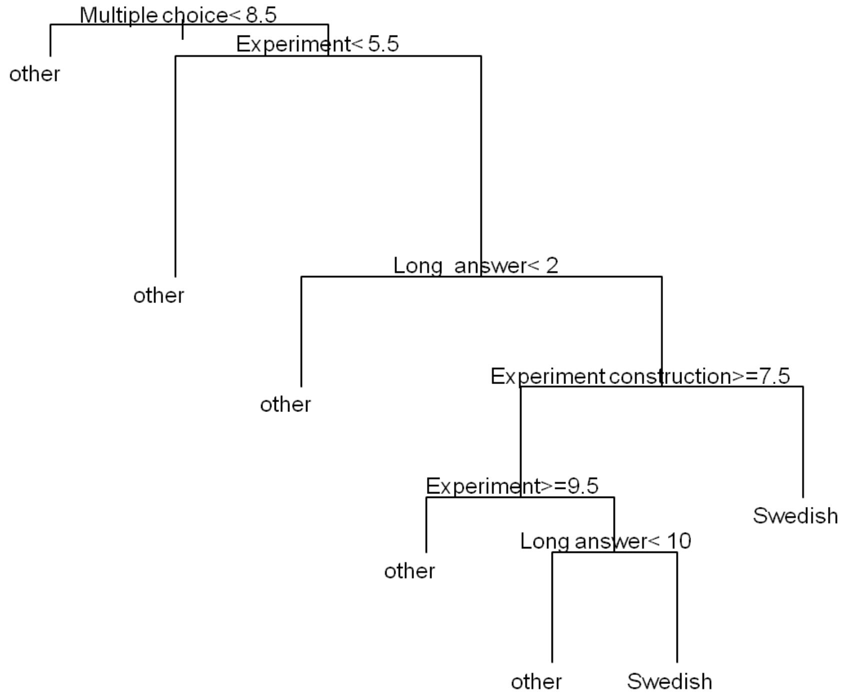

Fig. 2. *Results from a classification tree analysis where the scores from the five assessment types were used to classify the students according to mother tongue. 13 out of 17 students with Swedish as mother tongue are correctly classified and 28 out of 30 with other mother tongue.*

We used linear models to evaluate the association between the scores from the five assessment categories as response variables and the total score the R:s of Doll as explanatory variable. One of the variables was significantly affected by the total score the R:s of Doll. The experimental construction score was positively associated, the regression parameter was 0.072 with SE= 0.03 (F1,45= 4.7, P= 0.035).

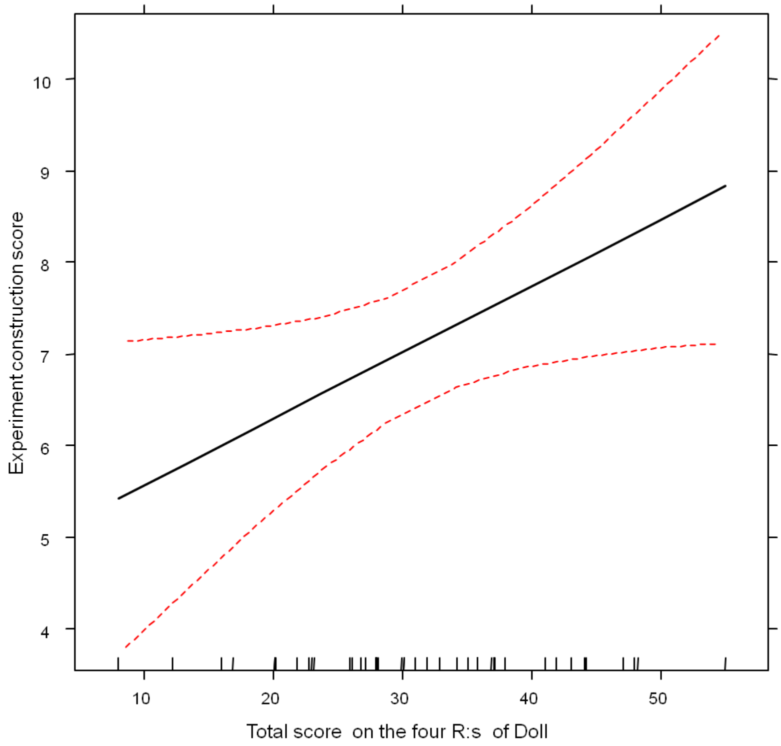

Fig. 3. *Prediction from a linear model using Experimental construction score as response variable and the total score on the R:s of Doll as explanatory variable.*

4. Discussion

The main picture from the PCA (Fig. 1) is that students with high scores in general comes out to the right in the graph. This is obvious for the assessment using Long answers, Experiment and Experiment construction. The scores for Experiment evaluation and Multiple choice questions are to a large extent following Principle component 2 but pointing in opposite directions. In can be noted that Long answer scores and Multiple choice scores are perpendicular to each other, meaning that the results from those two assessments are independent from each other. The interesting result here is that students with high scores in one type of assessment do not necessary get high scores in another type of assessment.

In the PCA there seems to be a tendency for high scores which point to the right in the graph to coincide with students with a Swedish mother tongue to be more common to the right in the graph. A more elaborate way to study the effect of mother tongue is through the classification tree (Fig. 2). The first division sorts out a group of students with non-Swedish mother tongue that gets low scores on multiple choice questions. We suggest that this may be that multiple choice questions demand an exact understanding of each word and of the construction of the sentence. Next two groups of students with non-Swedish mother tongue that have high scores at multiple choice questions but low scores on assessments from experiments a long answer questions are sorted out.

In the next two steps a group of students with non-Swedish mother tongue that have high scores at experiment construction and on experiment performance, meaning that those activities are better performed by a group of students with a non-Swedish mother tongue. Experiment construction is a creative act and we suggest that it can be useful to have a broader linguistic and maybe also cultural background or at least that the tasks are neutral with respect to mother tongue.

Two groups of students with a Swedish mother tongue are constructed – they have high scores in multiple choice questions and long answer questions but intermediate scores in experiments and in experiment construction, confirming the picture that multiple choice questions and long answer questions are dependent on linguistic skills and that they do not necessarily measure the learning outcomes they were designed for.

The summed scores for an evaluation according to Doll´s four R was positively affecting experimental construction score (Fig. 3). The estimation of the level of Doll´s R were made independently from the

learning outcome assessments and we see them as measurements of individual development. High scores on Doll´s R are connected with an assessment that contains creativity and we suggest that Doll´s R measurements reflect personal properties that are useful in creative tasks. There are suggestions that being multi-linguistic in general give positive effects for learning, see [10].
Our main conclusion is that different kinds of assessments give different results for different students and that especially multiple choice questions may be problematic for students with a non-Swedish mother tongue, possibly because they test language skills rather than the content knowledge of subject.

References

[1] Bosher S, Bowles, M. 2008. The effects of linguistic modification on ESL students' comprehension of nursing course test items. Nursing Education Perspectives*,* 29*:* 165-172.
[2] e Silva FAR, Mortimer EF, Coutinho FÂ. 2014. Investigation the evolution of conceptual profiles of life among university students of biology and pharmacy: the use of statistical tools to analyze questionnaire answers. In: Mortimer EF, El-Hani CN (eds). Conceptual profiles: a theory of teaching and learning scientific concepts. Contemporary trends and issues in science education 42. Springer Science + Business Media Dordrecht. 2104.
[3] Kolseth A, Lönn M. 2005. Genetic structure of *Euphrasia stricta* on the Baltic island of Gotland, Sweden. Ecography 28: 443-452.
[4] Dumanski JP, Rasi C, Lönn M, Davies H, Ingelsson M, Giedraitis V, Lannfelt L, Magnusson PKE, Lindgren CM, Morris AP, Cesarini D, Johannesson M, Tiensuu Janson E, Lind L, Pedersen NL, Ingelsson E, Forsberg L. 2015. Smoking is associated with mosaic loss of chromosome Y Science 2015: 81-83.
[5] Dolls Jr WE. 1993. A post-modern perspective on curriculum. Teachers College Press. Teachers College, Columbia University, New York and London.
[6] R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
[7] Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, M, Stevens HH, Wagner H. 2012. vegan: Community Ecology Package. R package version 2.0-5.
[8] Fox J. (2003). Effect Displays in R for Generalised Linear Models. Journal of Statistical Software, 8(15), 1-27.
[9] Therneau T, Atkinson B, Ripley B. 2012. rpart: Recursive Partitioning. R package version 4.0-3.
[10] Torpsten, A-C. 2102. Rights and Multilingualism. US-China Education Review B 4.