Item-response theoretical analysis of students’ achievement in mathematics word problems

Csíkos, Cs. A., Józsa, K.
Department of Education, Attila József University, Szeged, Hungary

and Kontra, J.
Mihály Munkácsy High School, Kaposvár, Hungary
This study was a part of a larger research project which focused on mathematics competence (see Csíkos, 1999; Józsa, 1999). The data collection took place in April and May, 1998.

Subjects were 2572 students from 31 schools across Hungary. The subjects’ ages ranged from 11 to 17 years (5th, 7th, 9th and 11th graders).

Among the tests one measured students’ problem solving. 8 out of 20 the tasks measuring problem solving skill can be considered as mathematics word problems. A lot of investigations have addressed the question of mathematical word problems. The evaluation of students’ achievement in some types of word problems calls forth interesting questions. In many cases the solution of this type of problems depends solely on the presence or
absence of some kind of ‘insight’. It is the case in the famous horse-trading or water-lily problems.

One of the 8 mathematics word problem from the problem solving test was an adopted version of the famous horse-trading problem (Maier & Burke, 1967, p. 307):

A man bought a horse for $60 and sold it later for $70. Then he bought it back for $80 and sold it for $90. How much did he make in the horse-trading business?

There can be some evaluation methods used to score students’ solution on these kinds of tasks. One possible evaluation method is to give 1 point to the correct solution and give 0 to any other. The lack of any algorithm for solving these tasks speaks in favor of the dichotomous (0-1) evaluation
method. However the tasks are not equally difficult, so scaling validity problems can easily take place. The use of OPLM item-response theoretical software (Verhelst, N. D., Glas, C. A. W. & Verstralen, H. H. F. M, 1995) can be a powerful tool in the evaluation of students’ achievement in mathematics word problems.

The problem solving test had two versions: one of them contained 8 problems that could be considered as purely mathematical ones. This test-version was administered to 575 9th and 11th grade students. 99 out of the 575 students had 0 point, and none of them scored the 8 point maximum. Consequently, 476 students were involved both in the classical and item-response theoretical analysis.
By means of using classical unweighted scoring methods there is no possibility to distinguish between students who equally had - for example - 3 points in this subtest.

Since 3 points out of 8 points can be reached in 56 different ways, it is not reasonable to presume that reaching 3 points indicates reaching a „3-point level of ability”.

The distribution of students’ results can be seen in the next diagram.
We may use evaluation methods based on weighted scores. In this case, each item has a ‘weight’: the 0-1 score is modified by multiplication by the weight of the item. Among the 8 items there are 4 relatively easy and 4 relatively difficult ones. The easier ones may have a weight of 1, the more
difficult ones will have a weight of 2. Using this scoring method the following distribution pattern can be drawn:

![Students' achievement (weighted scores)](image)

The two distributions are not only very similar to each other but there is a very high correlation between weighted and unweighted scores. \((r=0.96, p<0.001)\)
If we use scores of OPLM computed on the basis of latent trait abilities, a radically different distribution of performance emerged:
We use metric scores of OPLM instead of latent ability scale, because Lord (see Hambleton and Swaminathan, 1985) suggested that the metric score may be easier to understood by test score users. Furthermore, he suggested that the maximum should be the number of the items.

As it can be seen from the diagram, OPLM allows for much more detailed analysis of students’ performance in a subtest containing mathematics word problems. This result may be generalizable for many kinds of tasks used to measure students’ performance in a given field.

Using item-response theoretical models in the evaluation of students’ performance on different types of tasks encounters its first initial difficulties in Hungary. Spreading the practice of
using IRT-models requires demonstrating the power of probabilistic test models. This investigation can be considered as one of our efforts towards this aim.


