CONSTRUCT VALIDITY OF A DEVELOPMENTAL ASSESSMENT ON PROBABILITIES: A RASCH MEASUREMENT MODEL ANALYSIS

Athanasios Gagatsis, Leonidas Kyriakides & Areti Panaoura
DEPARTMENT OF EDUCATION, UNIVERSITY OF CYPRUS

ABSTRACT

The aim of the study reported in this paper was to obtain construct related evidence of a test on pupils' understanding on probabilities. The theoretical background underlying the design of the test is presented. The test was administered to year 4, year 5 and year 6 Cypriot pupils (n=623). The Extended Logistic Model of Rasch was used and the data were analysed by using the computer programme QUEST. A scale was created for the test and analysed for reliability, fit to the model, meaning and validity. It was also analysed separately for each of five groups (boys, girls, year 4, year 5 and year 6 pupils) to test the invariance of the scale. Analysis of the data revealed that the instrument has satisfactory psychometric properties. Five levels of probabilistic thinking were also identified. Despite the fact that there is a linear sequential hierarchy among the five levels, a big difference between the second and the third level was found. The findings are discussed with reference to intended uses of the assessment. Suggestions for further research are also drawn.

I) INTRODUCTION

In recently published literature on mathematics education there is a movement to introduce elements of probability into the elementary school curriculum. Thus, nowadays probability is one of the major areas of mathematics in primary curricula (NCTM, 2000; DfEE, 1999; Ministry of Education, 1994). The main purposes of the studies which have been developed by educational psychologists and mathematics educators about probability were the investigation of the conceptions of pupils about probability and the understanding of the concept of probability (Jones et al, 1999; Konold et al, 1993), the investigation of the misconceptions of children (Garfield & Ahlegren, 1998; O'Connell, 1999; Ayres & Way, 2000) and the proposition of valid framework which would enable young children's probabilistic thinking to be described and predicted across levels (Jones et al, 1997; Amir & Williams, 1999). It has been shown that early primary pupils are able to develop conception about probability prior to classroom instruction and the experiences of the six years old pupils are useful for the teaching of probability (Ayres & Way, 1999; Ojeda, 1999). Children bring informal knowledge acquired in daily life from their culture which might interfere with their learning of probability. Ayres and Way (1999) reveal that children without any formal probability schooling can make decision based on likelihood. However, Amir and William (1994) argue that it is possible for pupils of

twelve years old to encounter difficulties in predicting the probability of tossing a coin, believing that it depends on how you toss it.

Although there has been considerable research into students' probabilistic thinking, there has been almost no research on the development and evaluation of instructional programmes in probability. Instructional programmes are expected to be flexible and guided by formative assessment of pupils' understanding of the subject, according to the cognitive stages of children. Jones et al. (1999) supported that pupils' thinking about probability could be divided into four stages. At the first stage pupils predict the probability of an outcome on the basis of subjective judgement. At the second stage they predict the probability of an outcome after a combination of quantitative judgements and subjective judgements. At the third stage they compare the probabilities of different types of outcomes on the basis of consistent quantitative judgements and they distinguish "fair" and "unfair" probability generators on the basis of valid numerical reasoning. At the fourth stage they solve different types of problems on probabilities. This kind of theories empowers the construction of appropriate instructional programmes with purposes and school activities which are connected with students' thinking about probability. In the absence of a framework for systematically describing and predicting young children's thinking on probability the instruction of probability in primary education is possible to be inappropriate. According to Shaughnessy (1992) there is a need to develop appropriate tasks to assess students' conceptions of probability, their understanding of probability and their ability to solve problems on probabilities.

In this context, the main purpose of the study was the development of an assessment tool for measuring primary pupils' ability on the understanding of probability and the examination of the construct validity of the test. A test's construct validity is defined by the degree to which a set of items measures the theoretical construct it was designed to measure (Allen & Yen, 1979). Construct validity is an ongoing process whereby a test is evaluated in the light of a specific construct. It is therefore important to collect data to verify that the measured attribute behaves in concordance with the underlying theory (Cronbach, 1990). Eventually, the purpose of the study was not only to construct a valid tool of assessment on probabilities for pupils of primary school but also to identify levels of probabilistic thinking which could be helpful for diagnostic teaching of probabilities at primary education.

II) THE DEVELOPMENT OF THE TEST: SPECIFICATION OF THE CONSTRUCT DOMAIN

The construction of the test was guided by existing research and theory in the following two areas: a) current philosophy on Mathematics Education and b) research and theory on developmental assessment. Moreover, a key requirement in designing the test was its alignment with the mathematics curriculum that was operative in the area where the study is conducted. It was therefore taken into

account that probabilities consisted a major part of the Cyprus primary mathematics curriculum. A content analysis of the national textbooks in Mathematics was conducted which helped us to identify seven main aims of teaching probabilities at primary schools and the emphasis which is given to each of them. Moreover, a documentary analysis of the National Standards of the USA (NCTM, 2000) and of the English National Curriculum (DfEE, 1999) was conducted. It was found that these seven aims are also implied in the national standards of USA (NCTM, 2000) and the English national curriculum (DfEE, 1999). Thus, the test tasks were constructed according to these seven aims (see Table 1). As far as the influence of research and theory on developmental assessment, the applications of developmental assessments for measuring proficiency in cognitive abilities and content areas (Brown et al, 1992) were taken into account. Two essential concepts derived from these works were as follows: a) the developmental ordering of tasks on a continuum of difficulty and b) the provision of controlled, interactive support to examinees during the testing process. It was therefore decided to include in the test assessment tasks related with each aim on pupils' skills in probabilities which will cover a range of item difficulties. Moreover, in the instructions given to teachers who were asked to administer the test information were provided on the kind of support that pupils could have.

Table 1: Specification table of the Probability Test (PT)

Aims	**Tasks of the test at different levels**				
Pupils should be able to:	**1**	**2**	**3**	**4**	**5**
Describe events as impossible, likely or certain.	1, 2				
Compare events according to their degree of likelihood.	5, 8, 9	6, 10			
Estimate the probability of an event by using the formula of Laplace.		3, 4, 11			
Predict the change of the probability of an outcome of type A if the conditions of the experiment are changed.			12, 13		
Compute probabilities for simple compound events, using such methods as organised lists or tree diagrams.			7, 16,17		
Use Laplace formula of estimating the probability of an outcome of type A to compute the total number of possible outcomes or the number of outcomes of type A.				14, 15	
Examine the fairness of a game.					18, 19

III) METHODS

Once the final version of the test was developed, a table which indicated the relations between the tasks of the test and the aims of the teaching of probabilities was created (Table 1). The specifications and the tasks were content validated by two experienced primary teachers, the authors of the national textbooks, two postgraduate students of Mathematics Education, and two members of the Educational staff of the

Department of Education at the University of Cyprus. The "judges" of the content and the face validity of the test were asked to mark-up and to make comments on the items. In the light of their comments minor amendments were made. The final version of the written test was administered to 623 Cypriot primary pupils of year 4 (220), year 5 (218) and year 6 (185). Additionally, 318 of the subjects were girls and 305 were boys.

The Extended Logistic Model of Rasch (Rasch, 1980) was used and the data were analysed by using the computer programme Quest (Adams & Khoo, 1996) to create a scale satisfying the seven measurement criteria set out by Wright and Masters (1981) which have to be met in order to claim that the items form a valid and reliable scale. The scale is based on the log odds (called logits) of pupils' abilities to answer correctly the 19 items of the Probability Test (PT). The items are ordered along the scale at interval measurement level from easiest to hardest. The Rasch measure produces scale-free measures of pupils' abilities and sample free-item difficulties (Wright & Masters, 1981). This implies that the differences between pairs of measures of pupils' abilities and item difficulties are expected to be sample independent.

IV) FINDINGS

The data were analysed initially with the whole sample (n=623) and all the 19 items together. There were two items (16 and 17) which did not fit the model. Thus, the analysis was repeated with the whole sample and the 17 remaining items. Then, the analysis was repeated with each of the five groups of the sample. This was done to investigate whether the test is used consistently by boys, girls, year 4, year 5 and year 6 pupils and is part of the measurement criteria set out by Wright and Masters (1981).

Table 2: Statistics relating to the test for the whole sample and the five groups

Statistics	Whole (n=623)	Boys (n=305)	Girls (n=318)	Year 4 (n=220)	Year 5 (n=218)	Year 6 (n=185)
Mean (items)	0.00	0.00	0.00	0.00	0.00	0.00
(persons)	-1.18	-1.21	-1.16	-2.55	-1.17	-0.22
Standard deviation (items)	1.67	1.84	1.61	1.84	1.32	1.58
(persons)	1.39	1.35	1.36	0.96	1.22	1.17
Separability* (items)	0.99	0.99	0.99	0.96	0.99	0.99
(persons)	0.89	0.86	0.89	0.86	0.91	0.90
Mean Infit mean square (items)	0.99	1.00	0.99	1.00	0.99	1.00
(persons)	0.99	0.99	0.99	1.00	0.99	1.00
Mean Outfit mean square (items)	1.02	1.00	1.03	1.03	1.02	1.01
(persons)	1.02	1.01	1.04	1.02	1.03	1.01
Infit t (items)	-0.04	-0.05	-0.05	0.03	-0.04	-0.06
(persons)	-0.01	-0.01	-0.01	-0.03	-0.01	-0.04
Outfit t (items)	-0.07	0.06	0.11	0.10	0.11	0.09
(persons)	0.04	0.11	0.08	0.15	0.02	0.04

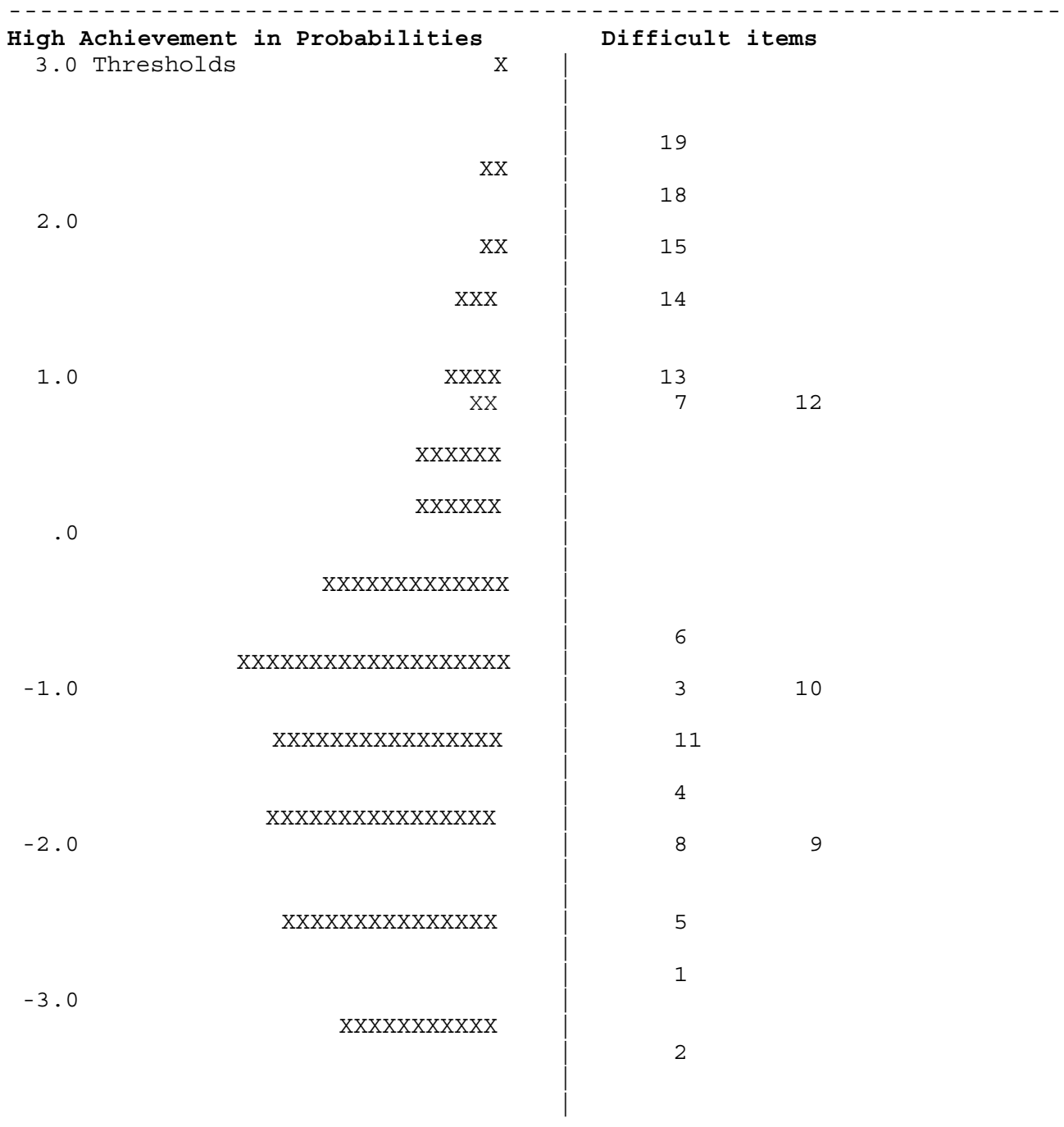
* Separability (reliability) represents the proportion of observed variance considered to be true. A value of 1 represents high separability and a value of 0 represents low separability.

Table 2 provides a summary of the scale statistics for the whole sample and the five groups. The following observations arise from this table. First, we can observe that for the whole sample and for each group the indices of cases and item separation are higher than 0,85 indicating that the separability of the scale is satisfactory. Second, the infit mean squares and the outfit mean squares are approximately 1 (0,99 up to 1,03) and the values of the infit t-scores and the outfit t-scores are approximately zero (-0,07 up to 0,11). And since the mean squares are within 30% of the expected values, calculated according to the model, it can be claimed that there is a good fit to the model. It is also important to note that all the items have difficulties which could be considered invariant across the 5 groups, within the measurement error (0,15). Thus, an important aspect of creating a scale (sample-free item difficulties) has been achieved. Third, the mean scores of pupils' performance indicate that an increase of performance by age can be identified. However, even the mean of year 6 pupils is relatively low (-0,22) and thereby the mean of the whole sample is very low (-1,18). Fourth, the standard deviations of the abilities of each year group but year 4 are relatively high. This implies that there is a big variation among the responses of year 5 and year 6 pupils whereas the performance of year 4 pupils was generally very low.

Figure 1 illustrates the scale for the remaining 17 items of the test with item difficulties and the whole group of pupils' measures calibrated on the same scale. The following observations arise from Figure 1. First, the items are well targeted against pupils' abilities in probabilities. More specifically, pupils scores range from -4,18 to 3,04 logits and the item difficulties range from -3,34 to 2,68. It can be, however, claimed that the targeting of the items at pupils' abilities in probabilities could be improved by adding some very easy items (Thresholds ≈ -4,00). Second, the most important weakness of the test is the absence of moderately easy items to moderately hard (i.e. from -0,76 to 0,91 logits). Thus, although the psychometric properties of the test seem to be satisfactory, the Probability Test could be improved by adding items which are neither easy nor hard. Third, five levels of probabilistic thinking can be identified. These levels are very similar to the levels mentioned at the specification table of the test. More specifically, pupils who are at the first level (i.e. below –2,00 logits) are able to describe events as impossible, likely or certain. They are also able to find which of two events is more likely to happen and their decision is based on the fact that they have realised that the probability of an event A depends on the number of outcomes of type A. However, pupils who are at the second level (-2,00 up to –0,75) are able to use the formula of Laplace for computing a probability and to compare events according to their degree of likelihood. After the second level, there is a relatively big area where none item is included. This could be attributed to weaknesses of the test either in including neither easy nor difficult tasks or in the fact that another level of probabilistic thinking should be included in the design of the specification table. However, this finding may reveal a gap between the

second and third level and that pupils have to make an important progress in order to move from the second to the third stage. At the third level (0,90 up to 1,30), pupils are able to compute probabilities for simple compound events (e.g. throwing two dices), using such methods as organised lists or tree-diagrams. They are also able to predict the changes of the probability of an outcome of type A when the conditions of the relevant experiment are changed. Pupils who are at the fourth level (1,30 up to 2,00) they can not only find the probability of an outcome of a specific type but are also able to use Laplace formula in order to compute the total number of possible outcomes or the number of outcomes of a specific type. Finally, pupils at the fifth level (above 2,00) are able to examine the rules of a game and find out whether the game is fair.

Figure 1: Scale for the Probability Test (N=623, L=17)



```
-4.0                          XXXXXX       |
  Weak achievement in Probabilities       Easy items
-------------------------------------------------------------------
```

Note: Each X represents 5 pupils

V) DISCUSSION

The Extended Logistic Model of Rasch was useful in creating a good interval level measure of the Probability Test identifying primary pupils' abilities in probabilities and for investigating its validity and reliability. The Rasch model was also helpful in analysing the conceptual design of the test. The findings of this study reveal that the Rasch analysis supports the conceptual design of the instrument. The underlying trait, that is primary pupils' abilities in probabilities, seems to be an overarching concept comprised of five different levels of probabilistic thinking. Thus, the Probability Test and its Rasch scale may help teachers decide how to identify and meet pupils' learning needs in relation to the five levels of probabilistic thinking and how to use their teaching time and their resources. An important implication of the identification of learning needs is that decisions about the next learning steps follow from it and pupils could be helped to improve their abilities and move from a lower level of thinking to a higher level. However, teachers should be aware of the fact that although the five levels follow a linear sequential hierarchy, there are pupils who are at the same level but their abilities may differ. Moreover, there is no clear distinction between the levels but between the second and the third level. It is important to note that acceptable fit was also obtained using structural equation modeling procedures for the theoretical five-factor first-order structure. A two second-order factors structure was also supported revealing that factors 1 to 2 (i.e. levels 1 and 2) were explained better by a second-order factor variable which was substantively different from that which was explaining factors 3 to 5. However, further research regarding the levels of probabilistic thinking is needed in order to examine whether a new level covering aims of teaching probability which are not mentioned in the Cyprus curriculum should be included in order to cover the area between the second and the third level of probabilistic thinking. Finally, the analysis leads to suggestions for improving the targeting of items against pupils' measures through the addition of one very easy item and some neither easy nor very difficult items (-0,5 up to 1,0 logits). Thus, further validation studies of a new version of the Probability Test may be needed in order to obtain a better targeting against primary pupils' abilities in probabilities.

References

Adams, R.J. & Khoo, S.T. (1996). *Quest: The Interactive Test Analysis System.* Camberwell, Victoria: ACER.

Allen, M.J. & Yen, W.M. (1979). *Introduction to Measurement Theory.* Monterey: Brooks and Cole.

Amir, G. & Williams, J. (1994). The influence of children's culture on their probabilistic thinking. In Joao Pedro de Ponte & Joao Filipe Matos (Eds.), *Proceedings of 18th Conference of the PME, 2,* 24-31. Portugal: University of Lisbon.

Amir, G. & Williams, J. (1999). Cultural influences on children's probabilistic thinking. *Journal of Mathematical Behavior, 18* (1), 85-107.

Ayres, P. & Way, J. (2000). Knowing the sample space or not: the effects on decision making. In T. Nakahara & M. Koyama (Eds.), *Proceedings of 24th Conference of the PME, 2,* 33-40. Hiroshima: University of Hiroshima.

Brown, A.L., Campione, J.C., Weber, L.S. & McGilly, G. (1992). Interactive learning environments: A new look at assessment and instruction. In B.R. Gifford & M.C. O'Connor (Eds). *Changing assessments: alternative views of aptitude, achievement and instruction (121-211).* Boston: Kluwer

Cronbach, L.J. (3rd Ed) (1990). *Essentials of Psychological Testing.* New York: Harper & Row.

DfEE (1999). *The National Curriculum for England.* London: DfEE.

Garfield, J. & Ahlegren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research, *Journal for Research in Mathematics Education, 19* (1), 44-63.

Jones, G., Langrall, C., Thornton, C., & Mogill, T. (1997). A framework for assessing and nurturing young children's thinking in probability, *Educational Studies in Mathematics, 32*, 101-125.

Jones, G., Thornton, C., Langrall, C., & Tarr, J. (1999). Understanding student's probabilistic reasoning. In L. Stiff & F. Curcio (Eds.), *Developing Mathematical Reasoning in Grades K-12* (146-155). NCTM Yearbook.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A.(1993). Inconsistencies in students' reasoning about probability, *Journal for Research in Mathematics Education, 24* (5), 392-414.

National Council of the Teachers of Mathematics (2000). *Principles and Standards for School Mathematics*. Reston, V.A.

O'Connell, A. (1999). Understanding the nature of errors in probability problem-solving, *Educational Research and Evaluation, 5* (1), 1-21.

Ojeda, M. (1999). The research of ideas of probability in the elementary level of education. In O. Zaslavsky (Ed.), *Proceedings of 23rd Conference of the PME, 4*, 1-8. Haifa: Institute of Technology.

Rasch, G. (1980). *Probabilistic Models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Shaughnessy, M. (1992). Research in probability and statistics: reflections and directions. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (465-495). USA: NCTM.

Wright, B. & Masters, G. (1981). *The Measurement of Knowledge and Attitude (Research memorandum no. 30).* Chicago: Statistical Laboratory, Department of Education, University of Chicago.