A developmental scale for assessing probabilistic thinking and the tendency to use a representativeness heuristic

Thekla Afantiti Lamprianou and Julian Williams
University of Manchester

We report a study of children's probability conceptions and misconceptions due to the representativeness heuristic. Rasch measurement methodology was used to develop a 13-item open response instrument with a sample (N=116) of 12-15 year olds. A hierarchy of responses at two levels is confirmed for this sample, and a third level is hypothesised. Each level is characterised by the ability to overcome typical 'representativeness' effects, namely 'recency', 'random-similarity' (at level 1), 'base-rate frequency' and 'sample size' (at level 2-3). Our interpretations were validated and anomalies identified through clinical interviews with children making the errors (n= 8), suggesting another measure, which we named the 'representativeness tendency' from 11 multiple-choice errors.

Introduction

This study builds on previous work on children's understandings, intuitions, use of heuristics and misconceptions in their probabilistic thinking (Fischbein, 1975, 1997; Kapadia & Borovcnik, 1991; Shaughnessy, 1992) and especially the significance of the representativeness heuristic (Green, 1982; Kahneman, Slovic & Tversky, 1982; Amir & Williams, 1999; Amir et al, 1999). The misconceptions based on the representativeness heuristic are some of the most common errors in probability: children tend to estimate the likelihood of an event by taking into account how well it represents its parent population and how it appears to have been generated.

In this study we aim to contribute to teaching by developing an assessment tool which can help teachers diagnose inappropriate use of the representativeness heuristic in responses to questions relevant to the probability curriculum. Williams and Ryan (2000) argue that research knowledge about students' misconceptions and learning generally needs to be located within the curriculum and associated with relevant teaching strategies if it is to be made useful for teachers. This involves a significant transformation and development of research knowledge into pedagogical content knowledge (Shulman, 1987), which requires its own study. The development of the assessment instrument involved tuning of, or development of diagnostic items from the research literature: thus the instrument provides a 'boundary object' between the research practice and innovative practice of assessment for teaching and learning.

Thirteen items were used to construct the instrument (the instrument can be seen in full on the web at http://www.education.man.ac.uk/lta/tal). The items identify four effects of the representativeness heuristic; the *recency* effect, the *random-similarity*

effect, the *base-rate frequency* effect and the *sample size* effect. Most of the items have been adopted with slight modifications of those used in previous research by Green (1982), Kahneman, Slovic & Tversky (1982), Konold et al (1993), Batanero, Serrano & Garfield (1996), Fischbein & Schnarch (1997) and Amir, Linchevski & Shefet (1999). Other items were developed based on findings of previous research.

Items called *recency1, 2 and 3* tested for the negative recency effect and the gambler's fallacy. According to this effect, a long sequence of one outcome must be followed by the other outcome in order to equilibrate the proportions. Items called *random-similarity 4, 5, 6 and 8* tested for the effect which expects a sample to appear similar in proportion to the parent population and apparently randomly-generated. These items were developed from Kanheman, Slovic & Tversky (1982), Fischbein & Schnarch (1997), Green (1982) and Shaughnessy (1992), respectively.

Items called *base-rate 10, 11 and 12* were written to examine the effect of prior probability or the base-rate frequency of the outcomes in contexts appropriate to this age group. According to this effect, prior probabilities are effectively ignored when misleading irrelevant but stereotypical information is introduced. As Kahneman, Slovic and Tversky (1982, p.5) mentioned, "when no specific evidence is given, prior probabilities are properly utilised; when worthless evidence is given, prior probabilities are ignored". Finally, items called *sample size 7, 9, 13* tested for the tendency to neglect sample size in estimating probability: the first from Fischbein & Schnarch (1997), the second from Shaughnessy (1992), the last ours. These items examined the belief that the probability of a certain proportion in a sample is independent of the sample size, contradicting the central limit theorem, i.e. the probability of getting a certain empirical result tends to approach the theoretical prediction as the sample gets larger.

Method

In order to be able to administer more items to the same sample of pupils two separate test-forms with common linking items were constructed. Test A, designed to be easier, consisted of eight items - items 1 to 6, 9, 10 - and Test B, intended to be more difficult, consisted of ten items - items 3, 4, 6 to 13. Five of the items were included in both tests. Test A was administered to pupils in Year 7 and Test B was administered to pupils in Year 8 and 9.

The tests were administered to 116 pupils from two schools in the North West of the United Kingdom. Before administering the tests to the pupils, the teachers of the six classes were asked to read and comment on the suitability of the tests for their classes. They found the wording of the items acceptable for the pupils' age, but they commented on the degree of difficulty of question 13 (*sample size 13*).

For the analysis of the results of the tests, a Rasch common calibration was used. Since all items had both a multiple-choice and an open-ended question, a common item Partial Credit analysis (Wright and Matsers, 1982) was run. One mark was given for the correct multiple-choice answer and another one for the correct explanation of the open-ended question for each of the 13 items. The result is a single scale consisting of a 'difficulty' estimate for each scored point and an 'ability' estimate for each child consistent with the Rasch measurement assumptions. Item 13 fell outside a model infit statistic value of 1.3 (see Wright & Stone, 1979) reflecting the difficulty of this item for the sample. The term 'ability' is defined by the performance of the pupil in this particular test, which we consider to imply an ability to avoid inappropriate representativeness effects, and so a particular ability with probabilities: it is NOT a measure of general ability, or even of mathematical ability. We will use the term *ability* italicised in this way throughout this paper.

In addition to the test analyses, we drew on structured clinical interviews with 8 children about the test items to gain insight into the cause of the effects described above, to confirm the literature, validate the items and identify anomalies.

Results

According to the table below, the test and sample can be interpreted as falling into a hierarchy of three levels. At level 1, (-3.0 to -0.5 logits) children can succeed on questions that tested for the *recency* effect and easy questions that examined the *base-rate frequency* and the *random-similarity* effect. At level 2 (-0.5 to 2 logits) children attain higher performance and they can explain their answers to the easier questions that tested for the *random-similarity* effect, they can manage harder *base-rate* and *random-similarity* questions and they are beginning to answer some *sample size* questions correctly. Very few children manage to attain level 3 by answering the hardest questions on *sample size* or explaining their answers to the harder questions that tested for the *random-similarity* effect. In order to establish level 3, it is suggested that a more able sample would be required.

By averaging the ability estimates of those children who made an error, we are able to plot errors on the same logit scale in the table. Pupils who gave responses indicating the *recency* misconceptions had a rather low *ability*. Answers indicating misconceptions based on the *random-similarity* effect were given by a broader range of *ability* pupils (averages ranged from -1.95 to -1.25 logits). On the other hand, the mean *ability* of the pupils who gave responses based on the *sample size* and *base-rate* effect was near the average *ability* of the sample, reflecting the fact that these errors were made by so many children (36%, 49%, 73%, 78% and 86% !!).

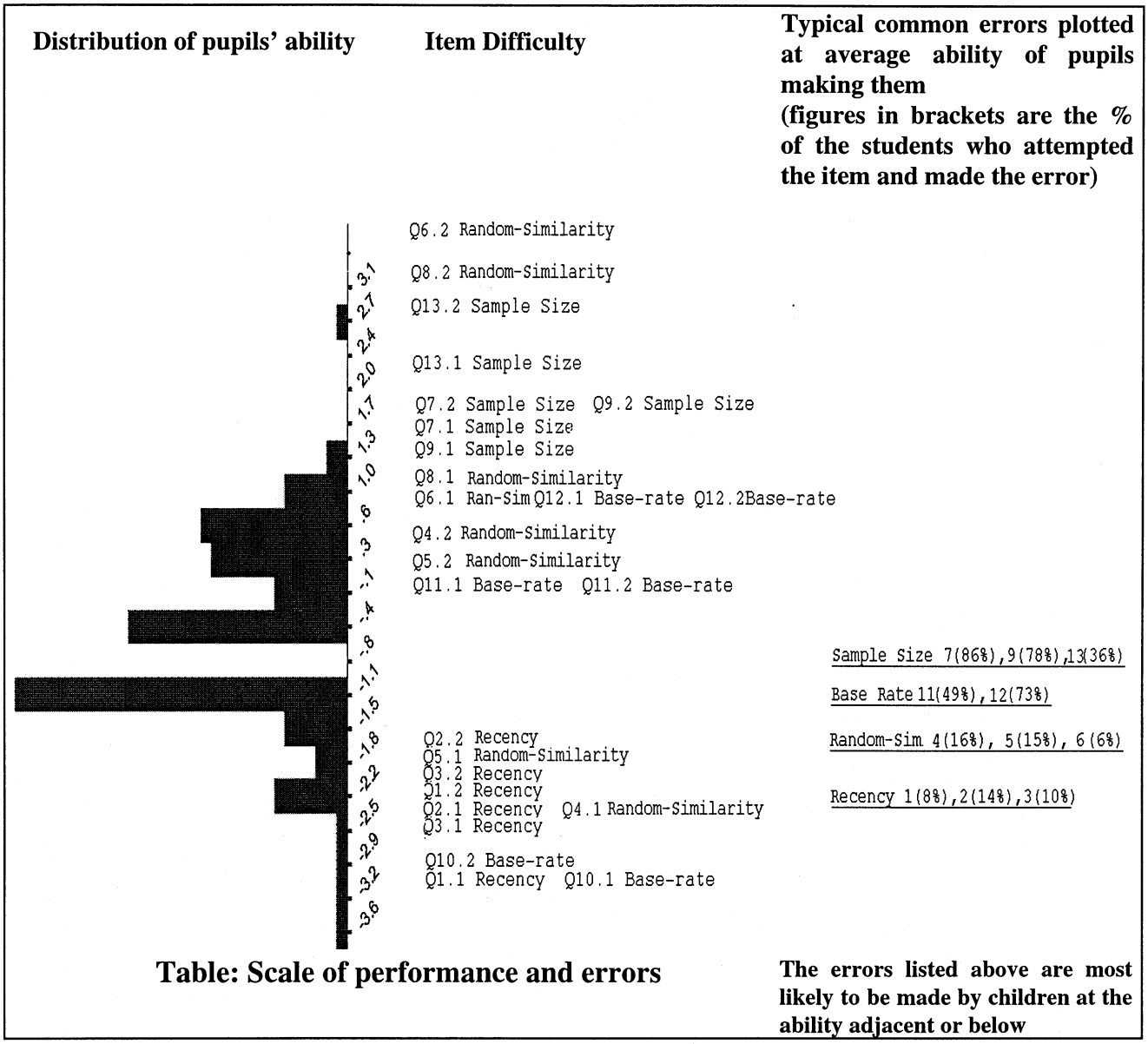


Table: Scale of performance and errors

The errors listed above are most likely to be made by children at the ability adjacent or below

This encouraged us to consider building a 'representativeness tendency' measure from those errors which we can authentically attribute to this heuristic, as a diagnostic measure of tendency to inappropriately apply this heuristic. The main purpose of the pupil interviews was validation of the test, in particular our interpretation that the errors in the test are symptomatic of the representativeness effects discussed in the literature. In this section we illustrate with the interviews of children the *random-similarity* effect, which was examined by *random-similarity 4, 5, 6 and 8*. For example, *random-similarity 4* is illustrated below:

Random-Similarity 4: A fair coin is tossed five times. Which of the following sequence of outcomes is the most likely result of five flips of the fair coin? (H: Heads, T: Tails) (a) HHHTT (b) THHTH (c) THTTT (d) HTHTH (e) All four sequences are equally likely. (Explain why).

Child10: … I'll probably pick (b), because it's a mixture of answers.

Teacher: (b), because it's a mixture of answers and -
Child10: - Because it is more realistic, because it's been a good … em mixed. . . .
Teacher: OK. But we have the same here in answer (d). There are three Heads and two Tails. Why did you choose that answer?
Child10: Because it's a bit more…em… because that's a Head, Tails, Head, Tails, Head, Tails and I think… I don't think it will be … to get that most times …

When this child was asked to solve a similar problem (*random-similarity 5*), his choice of frequency was representative to its parent population (it consisted of equal number of boys and girls) and it appeared to be a random mixture of boys and girls:

Random-Similarity 5: In a family of six children which sequence of births is the most likely? (B: Boy, G: Girl) (a) BBBGGG (b) BGBGBG (c) GBBGBG (d) GBGGGB (e) All four sequences are equally likely. (Explain why)

Child10: I'll probably think (c).
Teacher: Why?
Child10: Because you would more likely to em… pick that one because it's more mixture, but sometimes with different people they have all boys or just all girls. But some people have a mixture.

However, some items proved problematic. *Random-similarity 6 and 8* (the multiple-choice answers were classes of sequences, i.e. 6 Heads and 6 Tails) were developed from an item from Green, in which, as Amir, Linchevski & Shefet found, the majority of the pupils chose a different incorrect answer to the expected *random-similarity*. This was based on the conceptual error which reflected children's failure to discriminate between sequences and classes of sequences (combinations), between ordered sets and unordered sets in probability. An example from the interviews:

Teacher: … So. Why did you answer (e) "All have the same chance"?
Child61: Because they are all likely. You can't guess what … which **one** will be it. But they all have the same chance, because it's half. …

Teacher: OK. … (Child62 writes WWWGGG) … Is there any other sequence for three white and three grey?
Child62: No.

We therefore came to the conclusion this misconception was distracting from the *random-similarity* effect, which was only given by 6% of children in *random-similarity 6*. Indeed this fell to 0% for *random-similarity 8*, which suffered similarly. Worse still, children chose the correct option but for the wrong reason, arguing from a representativeness perspective, that '6 Heads and 6 Tails' would be most likely, because:

C33: Because, em, there's two sides on the coin and you get, em... even chances of getting six heads and six tails, because if you divide by two... added to two sides of the coin there's six on each, like heads and tails.

This also explains why so few children could give a correct explanation for their answers to these items. We suggest these items may need therefore to be deleted, redesigned or developed and re-scored.

Base-rate 10 might also be discarded when using the test diagnostically for this age group. Almost all pupils answered this question correctly. In addition, all eight interviewees gave the correct answer and justified their responses correctly. From the answers that children gave, it seemed that the irrelevant information did not distract any of the children and they did not ignore the prior probability. This item seemed as if it had no diagnostic value. Analysis of *base-rate 11 and 12*, however, suggested that the more the distracting description is related to a stereotype, the stronger the *base-rate* effect is.

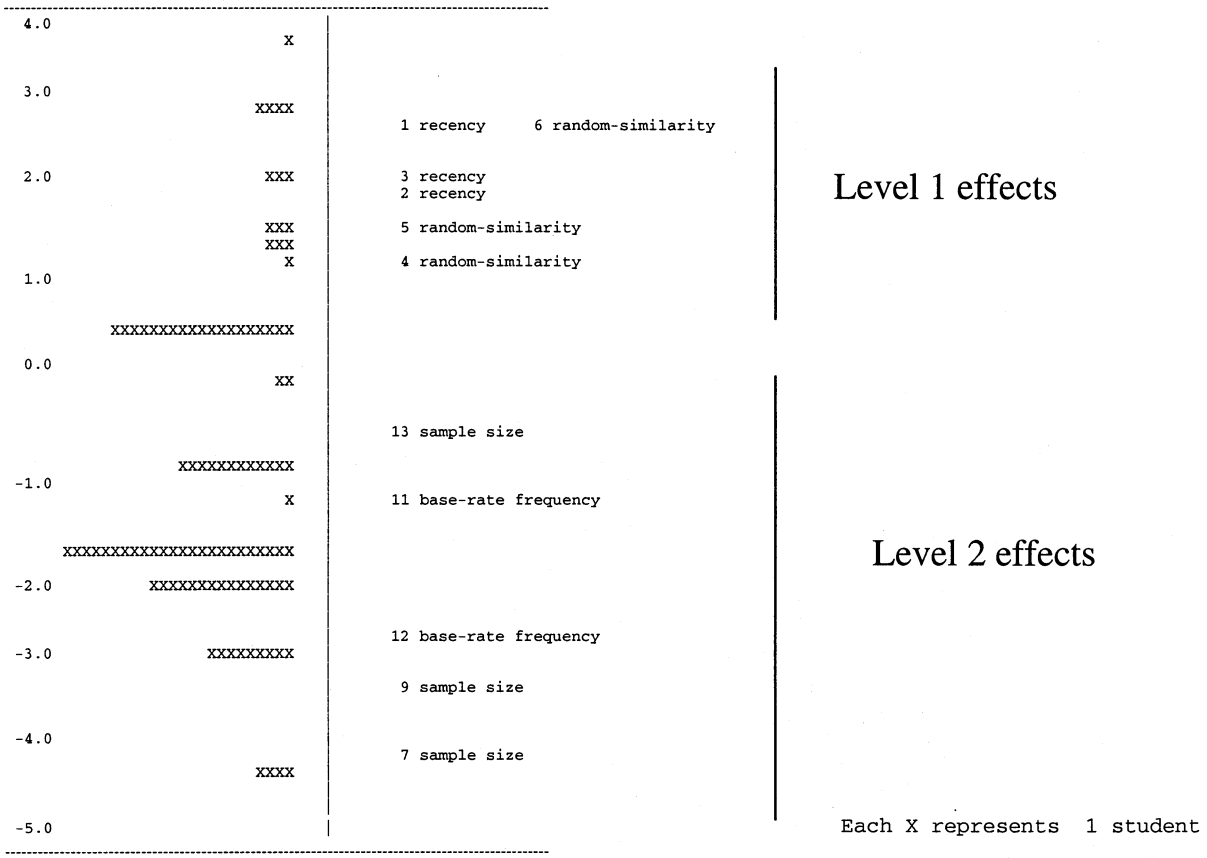
Construction of a representativeness measure

Since the purpose of the diagnostic instrument was to assess whether the representativeness heuristic influenced children's thinking when solving probability problems, a second Rasch model analysis was run. One mark was given only for the multiple-choice answer that indicated the effect of the representativeness heuristic but no marks were given for any other responses. The result was a single scale of items (none of the mark points fell outside a model infit value of 1.3), indicating that the *sample size* effect and the *base-rate* effect were very frequent among the pupils. The *random-similarity* and the *recency* effect influenced a small number of pupils of this sample (see graph below). *Random-similarity 8* and *base-rate 10* were removed from the Rasch analysis because all pupils gave different responses to the expected representativeness effects.

The result is a measure of 'representativeness tendency' for each person, and this naturally correlates negatively with their *ability* as measured previously (rho=- 0.64). However, the outliers are interesting: these represent three children who either found ways of scoring relatively well despite their tendency to use the representativeness heuristic, and vice versa. These might be the focus of further case study.

Children higher up the scale are more likely to make representative-effect related errors, and items higher up the scale are less commonly occurring, i.e. only made by those with a strong 'representativeness tendency'. Note that these fall into the two levels of questions identified previously in the table, with *recency* and *random-similarity* effects generally occurring at level 1 and *base-rate frequency* and *sample size* effects at level 2.

Graph: Item and Person Estimates of the 'representativeness tendency', using 11 multiple-choice responses only.



Conclusions and discussion

We have managed to develop two scales measuring children's responses to the instrument which is revealing about their probabilistic knowledge, especially as regards their inappropriate use of a representativeness heuristic in responding to test questions which are relevant to their curriculum. We have further identified some previously unknown interpretations of children's responses.

While most of the particular items in the scale are not new, the development, validation and calibration of the measures around this heuristic for 12-15 year old children is. We expect these to be useful research tools, but also to impact on teaching practice and teacher education, as discussed in Williams & Ryan (2001).

Having collected responses of some teachers to the instrument, we are doubtful that teachers are aware of these common misconceptions or of the significance of the representativeness heuristic, and we suggest that many teachers might benefit from using such an instrument in their assessment and teaching. The knowledge that teachers would collect from these scales, might enrich teachers' mental models of their learners and help them improve their classroom practice. We will be studying this aspect in the next stage of the work.

References

Amir, G., Linchevski, L. & Shefet, M. (1999). The probabilistic thinking of 11-12 year old children. In O. Zaslavsky (Ed.). *Proceedings of the 23rd Conference of the International Group for the Psychology of Mathematics Education (PME).* Haifa: Israel Institute of Technology.

Amir, G. S. & Williams, J. S. (1999). Cultural Influences on Children's Probabilistic Thinking. *Journal of Mathematical Behaviour,* 18(1), 85-107. 18(1), 85-107.

Batanero, C., Serrano, L. & Garfield, J. B. (1996). Heuristics and Biases in secondary school students' reasoning about probability. In L. Puig & A. Gutierrez (Eds.). *Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education (PME).* Valencia: Universitat de Valencia.

Fischbein, E. (1975). *The Intuitive Sources of Probabilistic Thinking in Children.* Dordrecht: Reidel.

Fischbein, E. & Schnarch, D. (1997). The Evolution with age of probabilistic, intuitively based misconceptions. *Journal of Research in Mathematics Education,* 28(1), 96-105.

Green, D. R. (1982). *Probability Concepts in 11-16 Year Old Pupils. Report of Research, CAMET.* Loughborough, England: University of Technology.

Kahneman, D., Slovic, P. & Tversky, A. (Eds.) (1982). *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Kapadia, R. & Borovcnik, M. (Eds.) (1991). *Chance Encounters: Probability in Education.* Dordrecht: Kluwer.

Shaughnessy, J. M. (1992). Research in Probability and Statistics: Reflections and Directions. In D.A. Grouws (Ed.). *Handbook of Research on Mathematics Teaching and Learning.* Reston: NCTM.

Shulman L.S. (1987). Knowledge and teaching: Foundations of the new reform, Harvard *Educational Review*, 57(1), 1-22.

Williams, J.S & Ryan, J.T (2000). National Testing and the improvement of Classroom Teaching: can they coexist?, *British Educational Research Journal*, 26(1), 49-73.

Williams, J.S & Ryan, J.T (2001). Charting argumentation space in conceptual locales. In M. van den Heuvel-Panhuizen (Ed.). *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education (PME).* Utrecht University.

Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Wright, B.D. & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.