FLAVIA JOLLIFFE

ASSESSING PROBABILISTIC THINKING AND REASONING

Chapter 13

> Because the fundamentals of probability are mathematically rather simple, it is easy to overlook the extent to which the concepts of probability conflict with intuitive ideas that are firmly set and difficult to dislodge by the time students reach secondary school. Misconceptions often persist even when students can answer typical test questions correctly. (Moore, 1990, p. 119)

1. INTRODUCTION

A broad interpretation of assessment is taken in this chapter. Thus both informal monitoring of students' thinking and reasoning, as well as more formal assessment are considered. Assessment in a research based context is not discussed although tasks used in research are considered.

According to Holmes (2002), assessment can be considered as having four purposes. These are formative, diagnostic, summative, and evaluative. Formative assessment is used mainly to give feedback to the student and the teacher, and sometimes to parents and others concerned with a student's progress. Formative assessment tends to mean setting students problems and marking their attempts, but informal monitoring in the classroom is a form of formative assessment too. This includes talking to students while they work on problems, talking to individuals or small groups undertaking practical work, observing students' activities, and group discussions with a whole class. The challenge for the teacher here is in making adequate records of students' contributions. Any grades given as part of a formative assessment are used only as feedback. However, research has suggested that telling students their grades is less effective than giving only other kinds of feedback (Holmes, 2002).

At younger ages summative assessment is used mainly to determine in which class a child is placed, or to which educational institution the child should progress. In general it is used to help determine the future education, and ultimately the career, of students, and grades are important. Diagnostic

Graham A. Jones (ed.), Exploring probability in school: Challenges for teaching and learning
325—344. © 2005

assessment is often used to find gaps in a student's knowledge. To some extent formative assessment is also diagnostic, but we might want to assess a student's background knowledge or innate ability rather than the extent to which the material taught in the course has been learnt. This is particularly relevant to probabilistic thinking and reasoning. Evaluative assessment is concerned more with assessing the effectiveness of the teaching than with the ability of the students, but summative assessment too could be used for this purpose.

Although a distinction is sometimes made between statistics and probability, the term statistics is often taken to include probability, and in this chapter probability is considered to be part of statistics rather than a topic in mathematics. Thus publications which at first sight appear to be concerned mainly with assessment of statistics, such as Gal and Garfield (1997), and Chapter 10 on assessment in Hawkins, Jolliffe, & Glickman (1992), contain much which is relevant to the assessment of probability. Even when research under the heading of the assessment of statistical thinking and reasoning does not include probability, the framework on which the research is based, the methodology, and the findings are still relevant to the assessment of probabilistic thinking and reasoning (Jolliffe, 1991).

We might adapt the passages on statistical thinking and statistical reasoning in Garfield, delMas, and Chance (2003) to explain what is meant by probabilistic thinking and probabilistic reasoning. Probabilistic thinking might be defined as the way people reason with the ideas of probability and make sense of probabilistic information (see also Langrall & Mooney, this volume). Reasoning means understanding and being able to explain and justify probabilistic processes (see also Watson, this volume). More specifically, probabilistic thinking involves understanding how models are used to simulate random phenomena, how data are produced to estimate probabilities, and how symmetry and other properties of the situation enable the determination of probabilities. It also involves being able to understand and use context when solving a problem, and having an appreciation of when subjective probabilities might be used.

Both thinking and reasoning involve understanding. As they are abstract concepts they can only be assessed indirectly, but they underpin the learning and teaching of probability. In general in assessment we assess the product, but not the process. However, it is easier to set and to mark rote calculations and manipulations than it is to devise questions that test understanding. Moreover, assessments requiring grading have in the past been notable for their lack of questions testing understanding, particularly apparent in externally set assessments. As classroom teaching has tended to concentrate on preparing students for summative assessments, the implications for the teaching of probability are a worry. It would be useful to know the effect, on subsequent progress and long term retention, of focussing on getting correct answers in spite of poor understanding. There are, however, signs that both assessments and classroom practice are now changing in line with recommendations for teaching (see Stohl, this volume).

An introduction to probability typically covers different approaches to the measurement of chance, particularly the relative frequency and equally likely approaches, the idea of an event space, properties of probability, addition of probabilities of mutually exclusive events, conditional probability, and independent events. However, facility with probabilistic thinking and reasoning is also needed when considering the theory and application of probability distributions and sampling distributions. Probability is crucial to the theory and practice of classical inferential methods (level of significance, power, p-values, interpretation of confidence intervals), and in Bayesian statistics. This chapter deals mainly with assessing probabilistic thinking and reasoning in the context of an introductory program on probability and in the context of what might be described as a layman's understanding of probability. It should be noted that although many probability problems involve combinatorial reasoning, assessment of such reasoning *per se* is not discussed in this chapter. Useful references for those interested in this topic are Batanero, Godino, and Navarro-Pelayo (1997) and Batanero and Sanchez (this volume).

A variety of frameworks for assessment is discussed in the next section of this chapter in the context of the assessment of probabilistic thinking and reasoning. Then, after a short section on types of assessment tasks, some specific examples of probability questions are given. These are considered in some depth with comments as to whether they assess thinking and reasoning; suggestions are also made as to how they might be modified to ensure they do. This is followed by a section on research studies into the understanding of probability concepts and similar matters. Such research is relevant to the design of good assessment instruments. The chapter concludes with a section on some assessment methods which might be used as alternatives to the more traditional methods, followed by a short section on the role of the teacher.

2. FRAMEWORKS FOR ASSESSMENT

Assessment cannot be considered in isolation from teaching and learning, although they each have different emphases, that of assessment being to enable pupils to show what they know (Holmes, 2002). In considering assessment it is therefore useful to look at frameworks and models for teaching and learning as well as those for assessment. These include official statements of curricular goals and objectives. Assessment tasks might be compared with frameworks to check what dimensions are being assessed.

Many schemes stem from Bloom's taxonomy of educational objectives in the cognitive domain. Wood (1968) suggested one for *mathematics* based on this. Wood's scheme can be adapted easily to statistics (Jolliffe, 1991) and to probability. It has as teaching objectives, (a) Knowledge and information, (b) Techniques and skill, (c) Comprehension, (d) Application, and (e) Inventiveness. These objectives are considered to be ordered along a simple to complex dimension, and higher objectives build on lower ones.

Of Bloom's five objectives, comprehension corresponds most closely to thinking and reasoning, but thinking and reasoning might also be considered to be part of inventiveness. The Wood scheme allows for three types of comprehension: translation, interpretation, and extrapolation. Translation is an activity requiring the change of form of a communication, for example explaining in words what is meant by a conditional probability written as P(A|B) and relating it to particular events A and B. Interpretation involves a rearrangement of material, for example changing frequencies in a table to estimates of probability. Extrapolation is an extension of interpretation and could include statements about the consequence of a communication, for example, when an estimate of the probability of an event A is applied to estimate the number in a sample expected to experience event A.

Wood (1968) defines inventiveness as assembling elements and parts to form a pattern or structure which was not previously clearly visible. It involves students in making discoveries and perhaps improvising, and might require an approach which is new to the student. As a probability example, the students might look at the lengths of runs in a sequence of coin tosses. Thinking and reasoning are clearly essential for inventiveness. This scheme might well be criticised for placing comprehension at a higher level than knowledge and techniques. Is it really possible to have knowledge without comprehension? Similarly some would argue that teaching of knowledge and information should build on what students find out for themselves (part of inventiveness). In teaching probability, particularly to young children, playing games, as described in the booklets of practical exercises, games and experiments in probability published by the Royal Statistical Society Centre for Statistical Education (http://science.ntu.ac.uk/rsscse/), can be a useful first step (see Pange, 2002, for further discussion).

Although proposed in relation to introductory statistics courses, the model of statistical reasoning in Chervany, Collier, Fienberg, Johnson and Neter (1977) is useful as a framework for assessing probabilistic reasoning. The three main stages are (I) Comprehension, (II) Planning and Execution, and (III) Evaluation and Interpretation. Here comprehension, which is to do with the student understanding the problem posed and having knowledge of the concepts given in the statement of the problem, is placed as the first stage where it logically belongs. Planning and execution is concerned with the student knowing how to solve, and solving, the problem. Evaluation and interpretation is broken down into verifying the solution to the problem from knowledge of similar problems, and stating the results using paraphrases.

Nitko and Lane (1991) give a framework to help instructors generate assessment tasks based on theoretical conceptualisations of statistical activities and understandings. They divided statistical activities into three related domains – statistical problem solving, statistical modelling, and statistical argumentation. They focussed on five interrelated ways of describing a person's understanding: understanding as representation, as knowledge structures, as connections among types of knowledge, as entailing active construction of knowledge, and as situated cognition.

Understanding as representation means that the student can move within and between internalised ideas, symbols, and systems, similar to translation in Wood's system. Understanding as knowledge structures is to do with being able to access and organise the knowledge needed in order to solve problems. Symbolic, formal and informal knowledge are examples of types of knowledge. In probability a student might understand the connection between informal knowledge that a fair coin falls heads down about half the time that it is tossed, and the formal knowledge based on equally likely theory that the probability of a head is ½. The active construction of knowledge involves expanding one's own knowledge structures and ways of thinking to incorporate concepts and principles. This would typically be assessed by setting students tasks at intervals over time. Understanding as situated cognition relates to being able to put learning in a real world context.

The assessment framework proposed by Garfield (1994) arises from the different aspects of assessment and has five dimensions. These are as follows: what to assess, the purpose of the assessment, who will do the assessment, the method of assessment, and the action to be taken following the assessment coupled with the nature of the feedback to be given to students. Assessment activities can be classified by all dimensions simultaneously, but some intersections of categories within dimensions are less meaningful than others. The dimension which has most in common with Wood's (1968) scheme and the model of Chervany et al. (1977) is *what to assess*. This dimension is broken down into concepts, skills, applications, attitudes, and beliefs. Thinking and reasoning are not mentioned explicitly, but would not be out of place in the "what to assess" dimension. Elsewhere in the paper, and in other work by Garfield and colleagues, it is clear that they consider understanding, thinking and reasoning to be important teaching goals and important indicators of what to assess.

Another framework to have in mind when considering the assessment of probabilistic thinking and reasoning is that proposed for assessing young children's thinking in probability (Jones, Langrall, Thornton, & Mogill, 1997; Jones, Thornton, Langrall, & Tarr, 1999). This was developed for situations in which probabilities can be determined by considering symmetry, number, or simple geometric measures. There are four constructs in this framework — sample space, probability of an event, probability comparisons, and conditional probability. The children's thinking, as shown by their responses to probability tasks, was classified into one of four levels. Level 1 was associated with subjective thinking, Level 2 was transitional between subjective and naïve quantitative thinking, at Level 3 the child used informal quantitative thinking, and in Level 4 numerical reasoning. The difference between this framework and the others discussed is that it is specific to probability. Although designed for young children it could be applied to the probabilistic thinking of older students (see Tarr & Jones, 1997; Tarr and Lannin, this volume).

3. TYPES OF ASSESSMENT TASKS

There are many different types of tasks which can be given to students to assess different dimensions of learning (Garfield, 1994), some being more appropriate to a particular skill than others. The method of assessment also needs to be appropriate for the age and stage of the student who is being assessed, and the purpose of the assessment. There is general agreement that a range of assessment methods is needed in order to get a comprehensive picture of a student's understanding.

It is easy to see what is unsatisfactory in assessment tasks set by others, but less easy to design tasks. However, help is becoming available through the web-based project ARTIST — Assessment Resource Tools for Improving Statistical Thinking (Garfield et al., 2003; http://www.gen.umn.edu/artist/). This is targeted at introductory statistics courses, and probability is one of the topics included in the project. In addition to resources, such as references to relevant publications, some of which can be downloaded from the web page, a collection of high quality assessment items is being developed. The plan is that these will be coded according to content and the type of cognitive outcome (Garfield et al., 2003). The cognitive outcomes include thinking and reasoning. Assessment items and tasks will be in a variety of formats, including items which require students to match concepts or questions with explanations, and longer written tasks such as projects and portfolios, as discussed in Section 6 of this chapter. At the time of writing the ARTIST project is at an early stage, but is likely to develop into an important resource well worth exploring.

In line with Jolliffe (1997) we can break assessment into assessment of factual knowledge, of computational ability, and on the ability to use computers. Questions could be posed in a multiple choice form, or could be open-ended. Projects and practicals are often open-ended. Students might be assessed on written or on oral answers, and might work on problems either as individuals or in a group. Probabilistic thinking and reasoning could be assessed in all of these types of tasks, although not all such tasks are designed to do so.

Tasks where students are able to associate the words in a problem with the equations in the course, and know where in the equation to substitute the numbers given in the problem, were called "pluginski" tasks by students at Berkeley (Freedman, Pisani, & Purves, 1978). Such questions lead to a tendency to shortcut thought and ideas and have little to commend them. They test little more than recognition of a type of problem and the ability to perform a computation, and do not assess thinking or reasoning.

4. TYPICAL PROBABILITY QUESTIONS

In this section some examples of fairly standard questions on elementary probability, set in a real-life context, are discussed. The primary aim of these questions appears to be to test knowledge of, and skills in, the rules of probability, rather than to test thinking and reasoning. Short-comings of the questions are pointed out, and suggestions are made as to how students' answers might indicate probabilistic thinking and reasoning, although such answers might not be what the assessor expected. Ways in which the questions could be modified to test probabilistic thinking and reasoning more directly are also suggested.

These questions are suitable for use with students who have been taught the elements of probability theory, but the context of some might need adapting for younger pupils. The questions could be difficult, and in that sense unfair as assessment questions, if students had not already seen solutions to similar types of questions. They involve understanding the problem, moving from the words in the problem to an alternative representation (*translation* in Wood, 1968, *understanding as representation* in the framework of Nitko and Lane, 1991), as well as knowing the rules needed to solve the problem.

Example 1

> In a residential area where there are 1,000 households, 800 have a computer, 600 have a video recorder, and 102 have a fax machine. Sixty per cent of the households with computers also have video recorders, but only 9 percent of households with computers also have fax machines. Forty-five households have all of a computer, video recorder, and a fax machine. Twenty households have a video recorder and a fax machine, but no computer. If a household is chosen at random from this area, find the probability that it has (a) none, (b) exactly one, (c) all three, of computer, video recorder, and fax machine.

This question can be solved fairly easily by drawing a Venn diagram and showing on it the numbers of households with different combinations of ownership for the three items. The fact that numbers are mentioned rather than probabilities suggests this method of solution. It is then more an exercise in arithmetic and logic than in probability. The question can also be solved by the extension of the additive law of probabilities to more than two events which are not mutually exclusive. It does not test probabilistic thinking or reasoning, and questions of this type rarely do. There is an attempt at making the situation realistic, but the numbers have clearly been chosen carefully to work out nicely. A market researcher might be interested in the probabilities, but if the numbers were known it seems likely that ownership by households might also be known. Would anyone choose just one household? Students reflecting on the given numbers might also consider the possibility that households might move, and that they might buy or sell the mentioned items, that is that the numbers given would change over time.

Example 2

> The probabilities that a parcel posted in central London arrives 1 day, 2 days, or 3 days after posting are 0.4, 0.5, and 0.1 respectively. The corresponding probabilities for a parcel posted in the suburbs are 0.3, 0.4, and 0.3. Two parcels are posted independently of one another, one in central London, and one in the suburbs. What is the probability that the parcel posted in central London arrives before the parcel posted in the suburbs?

Given these probabilities, the student has to think of a way to represent the events such that the parcel posted in the centre arrives before the one posted in the suburbs, that is, has to have in mind a suitable sample space, even if this is not written out in full. In this example this is more a matter of common sense than probabilistic thinking. There are three events of interest and the answer to the question can be obtained by multiplying probabilities to find the probability of each of these and then adding the three probabilities.

The word independently in the question is used in a general sense, but it is reasonable to assume statistical independence. Would we expect students to mention that? Would they do so if asked to state their assumptions? Some students, especially those who are not sure how to find the answer, might question whether the probabilities were reasonable, and might wonder what is meant by arrival. Where does it arrive? Is arrival the same as delivery? Are the two parcels being sent to similar destinations? One posted in central London to a central London address might well arrive more quickly than one posted in the suburbs to an address in a remote part of India.

The question could test probabilistic thinking and reasoning if it also included a part asking for a comment on whether the probabilities could be expected to hold for all parcels and all days, and a part asking how such probabilities might be estimated. A touch of realism could be added to the problem if it were set in the context of a firm dispatching orders by post. The dispatch manager might wish to know whether parcels posted in central London were more or less likely to arrive before parcels posted in the suburbs. Alternatively, students could be asked to suggest situations where the probability requested might be of practical interest. In both cases this is easier to do in formative assessment than in timed summative assessment. In the latter, care has to be taken not to overwhelm students with information which is not strictly relevant, and not to expect students to think up situations in a relatively short time.

Example 3

> A blind woman picks up two socks from an unsorted pile of 9 blue socks and 5 green socks. What is the probability that the two socks are the same colour as one another? What is the probability that the two socks are of different colours from one another?

In order to obtain the "obvious" answers to this question the student has to relate the situation to one where events, here picking up socks, occur at random. The student might then question whether the events "picking up a single sock" are equally likely. Is the woman more likely to pick up a sock from the top of the pile than from the bottom, or to pick up larger socks than smaller ones? Are woolly socks easier to pick up than silk ones? Is it possible than some socks are caught up with other socks? A student who considered these matters would be demonstrating probabilistic thinking as would a student who mentioned an assumption of equally likely events.

The question does not state whether the woman puts back the first sock before taking out the second, although it would seem reasonable to assume that she is taking socks out without replacement. On the other hand she might realise that the first sock was the wrong size or too thick for her purpose and so put it back before taking a second sock.

Conceptually the woman can be pictured as picking up two socks at once, or picking them out in turn. Students who use the "two socks at once" approach perhaps show a higher level of thinking than those using the latter. However if the woman picks up the socks together, it is more likely that the socks are lying together in the pile, so that socks which are near one another have a greater probability of being chosen than those which are further apart. Students who were worried about this might give up at this point unless prompted for an explanation of why they could not continue.

There are many variations of this question, but note how difficult it is to think of a situation which can easily be modelled according to the "rules" of probability. Moreover, it is also very difficult to word a question in such a way that it steers students toward using the rules.

It can be argued that students who perform well on this question in "pluginski" mode might well have little probabilistic thinking and reasoning ability. One good point is that the second question can be answered by noting that it refers to the complementary event to that in the first question. The question itself would be greatly improved if the students were asked to explain how they arrived at their answers or why they felt that they could not calculate answers.

All the difficulties in context discussed above can be avoided by asking essentially the same question in terms of balls in an urn. The set-up can then be made clear, and the correct answers can be obtained by applying the rules of probability; however, there is almost no test of probabilistic thinking and reasoning. The question could read:

An urn contains 9 blue balls and 5 green balls. You select two balls at random and without replacement from the urn. What is the probability that the two balls are the same colour as one another? What is the probability that the two balls are of different colours from one another?

Example 4

> Suppose that 35% of the mugs in a coffee shop are blue, 25% are, red and 40% are brown. Suppose further that 10% of the blue mugs are cracked as are 5% of the red and 7% of the brown. A woman buys a mug of coffee and finds that the mug is cracked. What is the probability that the mug is blue?

At face value this is a straight-forward question involving conditional probabilities and can be solved as an application of Bayes' theorem by plugging numbers into a formula. An alternative method of solution is to represent the sample space as a unit square. The square is divided into three rectangles of width 0.35, 0.25, and 0.40 to represent the mugs of different colours, and then each rectangle is subdivided to show the proportion cracked. The area representing blue cracked mugs divided by the area representing cracked mugs is the required probability. A student using this method to find the solution exhibits more probabilistic reasoning than a student solving the question by pluginski.

Is the problem believable? Why would anyone want to know the probability that if a mug is cracked it is a blue mug? In any case surely it is likely that, if a mug were seen to be cracked, coffee would not be put in it, and that cracked mugs would be discarded or kept for emergency use only. Are the given percentages such that the numbers of mugs of different colours and the numbers which are cracked are all integers? A student who tried to answer the question by finding numbers of mugs to satisfy the constraint, for example, by assuming there were 1000 mugs in total would quickly come unstuck. As with the other examples discussed, the attempt to introduce realism has not succeeded. However, the question could be improved if students were asked to state the assumptions they make when obtaining an answer and to discuss whether these assumptions are reasonable. Bayes' theorem does have useful applications of course, and good questions can be set in terms of medical and legal examples. Knowing that the probability of having an illness given a positive result on a test is not the same as the probability that a result is positive given that one has the illness, can be reassuring when hearing that the result of a test is positive.

The examples in this section have been chosen to cover standard techniques taught in an introductory course on probability. They have been discussed in some depth to illustrate how what at first sight might appear to be an interesting real-life question is actually an unrealistic situation. Further than this, in order to calculate a probability it might be necessary to make assumptions which are very unlikely to hold in practice. It is therefore always important, when writing an assessment question, to think carefully about the situation to which the question relates. With this in mind, suggestions have also been made in this section as to how questions which might be unsatisfactory as regards exercises in calculating probabilities could be suitable for assessing probabilistic thinking and reasoning.

5. TASKS USED IN RESEARCH STUDIES

Research into the understanding of probability concepts has shown that both children and adults have misconceptions concerning the outcomes of probabilistic events (Batanero & Sanchez, this volume; Kahneman, Slovic, & Tversky, 1982; Green, 1983, 1988, 1991; Jolliffe, 1994a; Jones & Thornton, this volume; Konold, 1995; Metz, 1997; Watson, this volume). At younger ages these might be related to the development of ideas of chance (Fischbein, 1975; Piaget & Inhelder, 1951/1975). At older ages the position is less clear, but some research studies suggest that there might be a tendency for these age groups to favour equally likely outcomes (Konold, Pollatsek, Well, Hendrickson, & Lipson, 1991; Jolliffe, 1994a). Konold (1995) reports that a 50% chance is interpreted as lack of knowledge about the outcome in outcome oriented individuals whom he defines as those who think of probabilities in terms of yes/no decisions. Other misconceptions which have been observed, and are well documented, include representativeness where subjects believe that a sample should exhibit the same distribution as the population from which it has been taken, and availability which is to do with how easy it is to think of particular instances of an event. These, and others, are discussed in Hawkins and Kapadia (1984) and Shaughnessy (1992).

In designing instruments to monitor and assess probabilistic thinking and reasoning of school students we need to be aware of this body of research and of the methodology used. This background can help us to ensure that the instruments are appropriate for the stage of development of the students and that they test what is intended. Some research tasks, perhaps with some adaptation, could be suitable for use in assessment.

Problems posed in research studies are sometimes based on scenarios which test probabilistic thinking and reasoning. They do not rely on knowledge of rules or the application of techniques, and are more concerned with intuitions. For example, a question in Nisbett, Krantz, Jepson and Kunda (1983) asked subjects to imagine that they were explorers who had landed on a little known island and had encountered a new bird, a shreeble, that was blue in colour. Subjects were asked what percent of all shreebles on the island they expected to be blue, and were then asked why they guessed this percent. In other versions of the question three or twenty blue shreebles were observed. This question was designed to explore beliefs about homogeneity and reliance on the law of large numbers.

In the experiments described in Fong, Krantz, and Nisbett (1986) there were three major types of problems. In *probabilistic* problems subjects had to draw conclusions about a population from a sample drawn at random; in *objective* problems the sample was objective but it was not clear that randomness was involved; and in *subjective* problems the sample data were clearly subjective. Six different underlying problem structures were used in each type of problem; for example, a large sample versus a small sample, a large sample from a population that was similar but not identical to the target population. Responses were coded as entirely deterministic, poor statistical, and good statistical. In that randomness and probability are closely linked, the questions can be considered to assess probabilistic thinking and reasoning.

One of the Fong et al. (1986) probabilistic problems described a procedure for deciding which 5,000 out of 10,000 students would be allowed to live on campus. Students picked, over a 3-day period, a number from a box containing numbers from 1 to 10,000. If the number picked was 5,000 or under the student could live on campus. Joe talked to five students on the first day of the draw and four of them had picked low numbers. He thought that the numbers were not properly mixed so he rushed over to pick a number and found that it was low. He later talked to four people who picked numbers on the second or third day and they all had high numbers. This confirmed his belief that the numbers were not properly mixed. Subjects were asked what they thought of Joe's reasoning and to explain their answers.

One of the objective problems concerned the psychology department at the University of Michigan. The admissions committee was considering whether to admit a particular student from a small nonselective college. One member of the committee argued against admission as their records showed that students from such colleges performed at a substantially lower level than Michigan students as a whole. Another member remarked that two years previously they had admitted a student from this college and that student was now among the three best students in the department. Subjects were asked to comment on the arguments put forward by the two committee members and to state their strengths and weaknesses.

One of the subjective problems in their study describes a man talking about his three-year-old son and saying that he thinks that the son will, like him, not have much interest in sports. He justifies this by referring to two occasions when he has observed the son playing ball with other children but quickly losing interest in the game. Subjects were asked whether they agreed with the father's reasoning and to say why they agreed or did not agree.

These problems, and others given by these authors, test an innate and general understanding of probability and in this sense probe into probabilistic thinking and reasoning. Moreover, in cases where the context is not immediately suitable for use with students at younger ages, the problems could easily be adapted. For example, rather than allocation to housing on campus, the problem could be related to a lottery where the 5,000 prizes are of a nature which would appeal to children of the age group concerned. Such problems could be used as part of monitoring, diagnostic or formative assessment, but lend themselves less readily to assessment where a grade is required, as grading is to some extent subjective.

There have been a number of studies building on those done by Green (1983, 1988, 1991) on school children aged 7-16 in the UK. These were concerned with the investigation of chance and probability concepts. The tests used contained questions on randomness and on the comparison of odds and could be used in nonresearch situations in the classroom. One of the randomness questions asked pupils to generate a pseudorandom sequence of 50 Hs and Ts to simulate the tossing of a fair coin (Green, 1991). A comparison of odds question (Green, 1983) read "A small round counter is *red* on one side and *green* on the other. It is held with the *red* face up and tossed high in the air. It spins and then lands. Which side is more likely to be face up, or is there no difference?" The choice of answers given was (a) the red side is more likely, (b) the green side is more likely, (c) there is no difference, and (d) don't know. With this question there was a tendency for younger pupils to show negative recency, opting for green being more likely next time because the counter was held with red face up, perhaps suggesting that it was green's turn. It is important to give a don't know option as this suggests to pupils that they are not necessarily expected to know the answer and might prevent some pupils guessing. Probing into the reasons as to why a particular answer has been given is also important and sometimes reveals that a correct answer is given for a wrong reason (Konold et al., 1991; Jolliffe, 1994b). Probing is particularly helpful when wrong answers have been given to fairly simple questions. It has the potential to be more successful in an oral than in a written assessment as the assessor can query responses when these are unclear and can give prompts if the student is having difficulty in making a response. However, probing can inhibit the student if the assessor is attempting to record responses.

6. ALTERNATIVE METHODS OF ASSESSMENT

In recent years some educators have developed methods of assessment as alternatives to the more traditional pen and paper methods that are based on standard questions found in the text-books published up to the last years of the 20th century. These alternative methods include authentic assessment (Colvin & Voss, 1997), use of portfolios (Keeler, 1997), oral assessment, assessment of group work, assessment based on using a computer, and assessment by projects or other investigations. Computer-based assessment typically consists of multiple choice questions, and often is programmed to give immediate feedback; it is not considered in this chapter.

In authentic assessment students are assessed on tasks that are relevant to them outside of school or college. Thus the context needs to be real, or at the very least realistic. With younger pupils probability questions might be set in terms of their chances of winning various games, or of getting a complete set of cards such as are sometimes included in packets of breakfast cereals. Older pupils might be more interested in their chances of winning a lottery, applications in risk or medicine, or election results. The Chance newsletter available at http://www.dartmouth.edu/~chance is a useful source of current examples and Everitt (1999) discusses interesting applications.

In portfolio assessment a selection of the student's work is collected into a portfolio for evaluation. The teacher and student usually agree on the selection, which is meant to represent what the student has learnt. The portfolio shows the student's progress and achievements over time. It is particularly suited to project work, enables the students to construct their own meanings for what they are learning, and can involve them in keeping a reflective journal which forms part of the portfolio. A portfolio for probabilistic thinking might include tasks involving the modeling of a random phenomenon such as the sex of first-born children, looking at data to estimate the probability that the first-born is male for different countries, and the implication of the results for a society where inheritance of certain privileges goes only to first-born males. Keeler (1997) gives a full discussion of the different issues involved in portfolio assessment.

Oral assessment of probabilistic thinking and reasoning has been used in many studies on the understanding of probability concepts, so there is much useful experience here. Some of the tasks used in a research setting could also be used in the classroom. Getting students to talk through their solutions to problems while they write them could form the basis of an assessment. One of the advantages of this mode is the interaction between teacher and students, making it easier for the teacher to explore a student's thinking. The method is particularly suitable for questions involving visual or physical representation; for example, young children could be asked about probabilities associated with segments of a spinner or could be asked about situations presented via a story (Kafoussi, 2004). When grading is important, questions and the way in which they are asked need to be standardised. A broad partly subjective grading scheme might work well, such as a score of 0 if the student displayed no or almost no understanding, a score of 2 for excellent understanding, and a score of 1 for something intermediate between 0 and 2. Some groups of students might be disadvantaged by oral assessment, for example those being assessed in other than their mother tongue, and shy students. Further comments on oral assessment are given in Hawkins et al. (1992, pp. 209-10) and in Jolliffe (1997, p. 202).

Projects and practical work might well be done by groups rather than individuals, and might involve use of the computer; hence, these methods of assessment can conveniently be considered together as in Hawkins et al. (1992, pp. 205-9). Clearly practical work is as useful in probability as in statistics more generally, and projects in probability topics could for example, be on modelling applications such as the spread of AIDS, or a queuing system. An obvious use of the computer is simulation. Successful projects in probability depend on students being able to think and reason probabilistically. Practical work might help them to think and reason in this way. If students have to work out details of the problems for themselves, it will be easier for the teacher to assess these qualities. For example, suppose students were asked to decide whether a table was composed of random digits. Rather than suggest that students looked at the proportions of each digit and pairs of digits and at runs of digits, they might in the first instance be left to decide how to examine the table. It is important to have a framework of objectives against which to assess projects and practical work, but as in the case of oral work, assessment is partly subjective. In assessing group work one difficulty is in assessing the contributions that individuals make. One possibility is to ask the different members of the group to rate the contributions made by others.

7. TEACHERS AND PROBABILITY ASSESSMENT

Teachers themselves might have a poor understanding of probability and their own misconceptions (Fischbein, 1990; Pratt, this volume; Stohl, this volume). This could make it difficult for them to recognise that their students' understanding is flawed. Some teachers also lack confidence when teaching topics involving numeracy and this could affect many aspects of their teaching, including the development and implementation of assessment tasks. Research into attitudes and beliefs could be useful in discovering, and helping to overcome, this problem.

Teachers will need training in how to assess their students. There are several strands to this: they need to know how to develop and to use instruments, how to organise their teaching in order to incorporate assessment into the time available, and how to record the outcomes of monitoring and assessment. As already mentioned, recording outcomes is particularly difficult for teachers in the case of observation of students. To date more effort has been put into training teachers how to teach (Hawkins, 1990) than how to assess, with the possible exception of projects. Teachers involved in marking examinations set by others might receive guidance and training in the implementation of the intended marking scheme. This certainly occurs in the case of public examinations such as the General Certificate of Education which is taken by pupils in many parts of the world.

Monitoring and formative assessment should ideally inform classroom instruction, giving the teacher the opportunity to spend further time on concepts and methods which have been misunderstood. Monitoring can also be made part of the process of teaching, for example by asking students to test their predictions of random events and by confronting them with their misconceptions (Chance, 2002).

8. CONCLUSION

This chapter has approached the assessment of probabilistic thinking and reasoning from several different angles. Such assessment has been considered against a background of the purposes of, and frameworks for, assessment. Definitions of probabilistic thinking and reasoning based on suggestions for what is meant by statistical thinking and reasoning have been given. Types of assessment tasks and different methods of assessment have been described, illustrated by examples of assessment tasks, including some used in research studies.

The role of the teacher in devising and implementing assessment has been examined. In particular, some examples have been discussed in depth to help the teacher ensure that tasks do indeed assess probabilistic thinking and reasoning. Exciting methods of assessment are beginning to be used in schools and colleges, and with the increased interest in teaching methods and emphasis on the importance of understanding that is occurring world-wide, assessment can only improve.

REFERENCES

Batanero, C., Godino, J.D., & Navarro-Pelayo, V. (1997). Combinatorial reasoning and its assessment. In I.Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 239-252). Amsterdam, The Netherlands: IOS Press.

Chance, B. (2002). Concepts of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3).[Online:www.amstat.org/publications/jse/v10n3/chance.html]

Chervany, N. L., Collier, R. O. Jr., Fienberg, S. E., Johnson, P. E., & Neter, J. (1977). A framework for the development of measurement instruments for evaluating the introductory statistics course. *The American Statistician, 31*(1), 17-23.

Colvin, S. & Vos, K. E. (1997). Authentic assessment models for statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 27-36). Amsterdam, The Netherlands: IOS Press.

Everitt, B.S. (1999). *Chance rules: an informal guide to probability, risk, and statistics* .New York, USA: Copernicus.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children.* (C. A. Shepherd, Trans.) Dordrecht, The Netherlands: Reidel.

Fischbein, E. (1990). Training teachers for teaching statistics. In A. Hawkins (Ed.), *Training teachers to teach statistics.* (pp. 48-57). Voorburg, The Netherlands: International Statistical Institute.

Fong, G .T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253-292.

Freedman, D., Pisani, R., & Purves, R. (1978). *Instructor's manual for statistics.* New York: Norton.

Gal, I. & Garfield, J. B. (Eds.). (1997). *The assessment challenge in statistics education.* Amsterdam, The Netherlands: IOS Press.

Garfield, J. B. (1994). Beyond testing and grading: Using assessment to improve student learning. *Journal of Statistics Education* (Online), *2(1).* [Online:www.amstat.org/publications/jse/v2n1/garfield.html]

Garfield, J., delMas, R. C. & Chance, B. (2003, April). *The Web-based ARTIST: Assessment resource tools for improving statistical thinking.* Paper presented at the annual meeting of the American Education Research Association, Chicago, USA. Available: http://www.gen.umn.edu/artist/

Green, D. R. (1983). A survey of probability concepts in 3,000 pupils aged 11-16 years. In D. R. Grey, P. Holmes, V. Barnett & G. M. Constable (Eds.), *Proceedings of the First International Conference on Teaching Statistics* (Vol. II, pp. 766-783). Sheffield, UK: Teaching Statistics Trust Sheffield University.

Green, D. R. (1988). Children's understanding of randomness: a survey of 1,600 children aged 7-11. In R. Davidson & J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics.* (pp. 287-291). Victoria, Canada: University of Victoria.

Green, D. R. (1991). A longitudinal study of pupils' probability concepts. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. 1, pp. 320-328). Voorburg, The Netherlands: International Statistical Institute.

Hawkins, A. (Ed.) (1990). *Training teachers to teach statistics.* Voorburg, The Netherlands: International Statistical Institute.

Hawkins, A., Jolliffe, F. & Glickman, L. (1992). *Teaching statistical concepts.*
. London: Longman.

Hawkins, A., & Kapadia, R. (1984). Children's conceptions of probability – A psychological and pedagogical review. *Educational Studies in Mathematics, 15, 349-377.*

Holmes, P. (2002). Teaching, learning and assessment: Complementary or conflicting categories for school statistics. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*[CD-ROM]. Hawthorn, VIC: International Statistical Institute.

Jolliffe, F. R. (1991). Assessment of the understanding of statistical concepts. In D.Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. 1, pp. 461-466) Voorburg, The Netherlands: International Statistical Institute.

Jolliffe, F. (1994a). Proportions, probability, and other matters. In L. Brunelli & G. Cicchitelli (Eds.), *Proceedings, International Association for Statistical Education First scientific meeting, Italy 1993* (pp. 377-383). Perugia, Italy: University of Perugia.

Jolliffe, F. R. (1994b). Why ask why? In, *Proceedings of the Fourth International Conference on Teaching Statistics.* (Vol. 1, pp. 57-64), Morocco. (Reprinted in Research Papers from the 4th International Conference on Teaching Statistics, edited by Joan Garfield. The International Study Group for Research on Learning Probability and Statistics, 1995.)

Jolliffe, F. (1997). Issues in constructing assessment items for the classroom. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 191-204). Amsterdam, The Netherlands: IOS Press.

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

Jones, G. A., Thornton, C. A., Langrall, C. W., & Tarr, J. E. (1999). Understanding students' probabilistic reasoning. In L. V. Stiff & F. R. Curcio (Eds.), *Developing mathematical reasoning in Grades K-12: 1999 Yearbook* (pp. 146-155). Reston, VA: National Council of Teachers of Mathematics.

Kafoussi, S. (2004). Can kindergarten children be successfully involved in probabilistic tasks? *Statistics Education Research Journal, 3(1), 29-39.* [Online:www.stat.auckland.ac.nz]

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Keeler, C.M. (1997). Portfolio assessment in graduate level statistics courses. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp.165-178). Amsterdam, The Netherlands: IOS Press.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education* *3*(1). [Online:www.amstat.org/publications/jse/v3n1/konold.html]

Konold, C., Pollatsek, A., Well, A, Hendrickson, J., & Lipson, A. (1991). The origin of inconsistencies in probabilistic reasoning of novices. In David Vere-Jones (Ed.), *Proceedings* of *the Third International Conference on Teaching Statistics* (Vol. 1, pp. 357-362) Voorburg, The Netherlands: International Statistical Institute.

Metz, K. E. (1997). Dimensions in the assessment of students' understanding and application of chance. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 223-238). Amsterdam, The Netherlands: IOS Press.

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday statistical reasoning. *Psychological Review, 4,* 339-363.

Nitko, A. J., & Lane, S. (1991). Solving problems is not enough assessing and diagnosing the ways in which students organise statistical concepts. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics.* (Vol. 1, pp. 467-474) Voorburg, The Netherlands: International Statistical Institute.

Pange, J. (2002). Can we teach probabilities to young children using educational material from the internet? In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics* [CD-ROM]. Hawthorn, VIC: International Statistical Institute.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in students* (L. Leake, Jr., P. Burrell, & H. D. Fischbein, Trans). New York: Norton (Original work published 1951)

Shaughnessy. J. M. (1992). Research in probability and statistics. In D. Grouws (Ed.), *Handbook of research in mathematics education teaching and learning* (pp. 465-494). New York: Macmillan.

Tarr, J. E., & Jones, G. A. (1997). A framework for assessing middle school students' thinking in conditional probability and independence. *Mathematics Education Research Journal, 9*, 39-59.

Wood, R. (1968). Objectives in the teaching of mathematics. E*ducational Research, 10, 83-98.*