A FRAMEWORK FOR ASSESSING QUESTIONS INVESTIGATING THE UNDERSTANDING OF PROBABILISTIC CONCEPTS

John M. Truran
Private Practice, Adelaide, South Australia

Research has suggested that two well-regarded tests of probabilistic understanding measure different constructs. Here a framework is proposed for classifying questions on probabilistic understanding and used to analyse the structure of the tests. It provides a way of predicting some differences found by post factum *statistical analysis. This suggests that it may have more general validity for evaluating instruments assessing probabilistic understanding.*

Probability has come into many school curricula at much the same time as schools have been expected to raise the quality of their assessment procedures. Here a framework is proposed for classifying questions on probabilistic understanding, and is used to help explain some research findings about two well-regarded tests.

Some mathematics educators argue, with some validity, that written assessment instruments are inappropriate for assessing complex concepts like probability. Alarcon (1982), for example, has shown that, when probabilistic concepts are examined in discussions and then in a written questionnaire, many students change their responses. These arguments are not addressed here because the external demands on schools for increased assessment mean that we need to be able assess the quality of those instruments which are likely to be used, and the principles outlined are equally applicable to more interactive modes of assessment.

Two Instruments Assessing Probabilistic Understanding

Green (1982a, 1982b) reported a study of 3000 children from Years 7 to 11 in the East Midlands of England. He hoped that his carefully developed assessment instrument could be related to Piagetian levels as had been done for other school topics by the Concepts in Secondary Mathematics and Science Project. He obtained informative responses to a wide variety of questions, but was unable to link all of them to a Piagetian stage-model. Only eighteen of the original, quite diverse, fifty items satisfied the requirements for validity and reliability, so these constituted the sub-test he used to establish his Piagetian stage-model.

Soon after, Fischbein & Gazit (1984) developed two instruments to assess the effect of a programme of instruction in probability with about 300 Israeli pupils from Years 5 to 7 in the experimental and control groups. The first (A) contained procedural questions directly related to the instruction, the second (B) comprised eight questions specifically designed for the experiment to reveal several well-known probabilistic misconceptions which were potentially present. The authors found that the questions were able to detect changes in understanding, although sometimes the effect of instruction seems to have led to decreased understanding.

Subsequent Use of These Instruments

Both tests have been used by subsequent researchers (e.g., Watson, Collis & Moritz, 1994; Glencross & Laridon, 1994) as source of questions, sometimes with minor modifications. Green (1986) used his experience to develop a new test to assess understanding of randomness, but has not re-used his original test. However, Izard (1992) administered Green's eighteen items to about 1100 students in Hungary, Brazil, and francophone Canada and found a general confirmation of Green's work but with some variations. He considered that the test had acceptable test reliability and item fit, but that the decision rules for grading the open-ended questions needed further examination. Surprisingly, he did not make comparisons between the different countries.

All these researchers seem to have believed that their questions were valid tests of the concepts being investigated, although Green was concerned that his statistical analysis had removed some of what he saw as his more interesting questions. Later Godino, Batanero & Cañizares (1994) presented both these tests of primary probabilistic reasoning to the same group of 251 Spanish children in Years 6 to 8, so the opportunity arose to examine just what the questions were assessing.

Godino et al. constructed a detailed analysis of the skills and understandings being tested in both tests (using all fifty of Green's items). They argued that if Green's test were a test of probabilistic reasoning then the whole test, and perhaps its reduced form would have high predictive value for results from some other such test. They found significant correlation between some of the Israeli questions and Green's test, but not for all comparisons, and concluded, *inter alia,* that the Israeli test contained components of probabilistic reasoning not included by Green. A factor analysis of the Israeli test found two factors (one based mainly on qq. 6, 7 & 8, and the other on qq. 2, 3, 4 & 5), the second of which did not correlate well with Green's results. A factor analysis of Green's test produced fifteen factors, thus confirming its diversity of coverage. They concluded:

> The multitude of factors included in Green's test and the low predictive value of the "probabilistic level" and of the other scores in the said test, with respect to the success in Fischbein & Gazit's test, suggest that a critical review is needed to consider the probabilistic knowledge of the subjects as a linear structure.
>
> More research is needed to explore in-depth the nature of probabilistic reasoning and its structure. At the same time it shall be necessary to compile and analyse the banks of items that make up a representative sample of [Primary Probabilistic Reasoning] and of the universe of appropriate contextual variables … .

A Framework for Classifying Questions on Probabilistic Understanding

At the same time as Godino et al. were doing their work, J. Truran (1994) was preparing a framework, of which a modified version is presented in Table 1, for class-

Table 1
A Framework for Classifying Random Probability Functions and Ways of Encountering Them

		1	2	3	4
A	Type of RG	Disc	Coin	Die	Urn
		5	**6**	**7**	**8**
		Cards	Contiguous Spinner (†)	Non-contiguous Spinner	Electronic
		9	**10**	**11**	**12**
		Human	Asymmetric Solids	Many interacting forces	Miscellaneous

		1	2	3	4
B	Place of RG in Culture	Own Culture	Unusual	Different	
C	Previous Practical Experience with RG	> 7 Days	2–7 Days	< 1 Day	
D	Previous Theoretical Experience with RG	> 7 Days	2–7 Days	< 1 Day	None
E	Operator of RG	Not Mentioned	Self	Other—Present	Other—Absent
F	Style of Response	Oral	Written	Multiple Choice	Non-Linguistic
G	Number of Elementary Events	2	Small	Large	Very Large
H	Number of Events	2	3–5	6	> 6
I	Structure of RG	Symmetric	Slightly Asymmetric	Very Asymmetric	Deceptive
J	Knowledge of Structure of RG	Known	Unknown		
K	Reward	None	Hypothetical	Actual	

	Type of Question		**Type of Probabilistic Situation**
α	Prediction of Outcome	I	Single Trial
β	Prediction of Set of Outcomes	II	More than One Trial
γ	Selection of Outcome	III	Previous Results
δ	Statement of "Likely" Outcome	IV	Previous Predictions of Results
ε	Comparison of RGs	V	Concurrent Operation of Another RG
ζ	Fair Allocation of Payout for Bets	VI	Previous Experience with Similar RGs
η	Examination of Sequences of Outcomes	VII	Changes in RG from Trial to Trial
θ	Linguistic Questions of Technical Knowledge	VIII	Long Term Reward Maximisation
ι	Listing of Outcomes		

(†) Spinners may be further divided in three categories:
(a) those where only the pointer is free to move;
(b) those where only the sectors are free to move;
(c) those where both pointer and sectors are free to move (as in a roulette wheel).

ifying random generators (RGs) and questions about them. He saw this as useful for assessing the comprehensiveness of a test or unit of work, and it now also seems to be able to fulfil some of the gaps identified by Godino et al. It has three parts—the first describes the nature of the RG and its relationship to the student, the second the type of question being asked, and the third the type of probabilistic situation in which the question is used. For simplicity the classification deals only with outcomes from a single RG, not with compound RGs. It has been based on many research findings about the types of probabilistic situations which seem to influence students' responses. There is not space to list all these findings here, but, for example, Category E, which considers who actually operates the RG, follows in part from Zaleska's (1974) finding that responses may differ according to whether the subjects or the experimenters actually draw the balls from an urn. It has been tested for its ability to classify more than fifty questions developed by a wide range of researchers, including Green, but not Fischbein & Gazit, and the modification presented here summarises the experiences gained in this preliminary testing.

To help to clarify the complex detail of Table 1, the classification of one of Green's questions is presented in Table 2. The question is:

> When an ordinary 6 sided dice is thrown which number or numbers is it hardest to throw, or are they all the same?
>
> Answer..

The right hand column summarises of the meaning of the alpha-numeric terms to clarify the links with Table 1. The only potential ambiguity here is the classification "δ"—statement of "likely" outcome. Here the student is being asked for an "unlikely" outcome, and this meaning is implied by the use of quotation marks.

Table 2
Analysis of Green's Question 4

A	3	Die
B	1	Own Culture
C		Unknown
D		Unknown
E	1	Operator: not mentioned
F	2	Written
G	2	Small Number of Elementary Events
H	3	6 Events
I	1	Symmetric
J	1	Known
K	1	No Reward
	δ	"Likely" Outcome
	Ι	1 Trial

Analysis of the Two Tests

All the items of Green (G) and Fischbein & Gazit (FGB) concerned with a single RG have been classified using this framework, and the results are summarised in Table 3 with Green's items first. His test seems to have 53 items, not 50 as mentioned above, and it has not been possible to explain all the reasons for the discrepancy. The term “u” refers to “unknown” and “na” to “not applicable”—cases where the questions are not dealing with operations of a single random generator. Those questions which comprised Green’s final statistically reliable form are in bold, apart from G10 and G26 (a) which do not fit in the framework. This makes nineteen items, not eighteen: both parts of G3 were probably taken as one item.

Table 3
Classification of Questions of Green and Fischbein & Gazit

Question	RG Type	RG Nature A	B	C	D	E	F	G	H	I	J	K	Quest-ion	Situ-ation
G1	disc	1	1	u	u	1	3	1	1	1	1	1	δ	I
G2	**urn**	**4**	**1**	**u**	**u**	**4**	**3**	**3**	**1**	**1**	**1**	**1**	**δ**	**I**
G3 answer	**spinner**	**6**	**1**	**u**	**u**	**1**	**3**	**2**	**2**	**1**	**1**	**1**	**ε**	**I**
G3 reason	**spinner**	**6**	**1**	**u**	**u**	**1**	**2**	**2**	**2**	**1**	**1**	**1**	**ε**	**I**
G4	**die**	**3**	**1**	**u**	**u**	**1**	**2**	**2**	**3**	**1**	**1**	**1**	**δ**	**I**
G5	**coin**	**2**	**1**	**u**	**u**	**1**	**3**	**1**	**1**	**1**	**1**	**1**	**δ**	**III**
G6 (a) answer	urn	4	1	u	u	2	3	2	1	1	1	1	ε	I
G6 (a) reason	urn	4	1	u	u	2	2	2	1	1	1	1	ε	I
G6 b answer	**urn**	**4**	**1**	**u**	**u**	**2**	**3**	**2**	**1**	**1**	**1**	**1**	**ε**	**I**
G6 b reason	**urn**	**4**	**1**	**u**	**u**	**2**	**2**	**2**	**1**	**1**	**1**	**1**	**ε**	**I**
G6 c answer	**urn**	**4**	**1**	**u**	**u**	**2**	**3**	**2**	**1**	**1**	**1**	**1**	**ε**	**I**
G6 c reason	**urn**	**4**	**1**	**u**	**u**	**2**	**2**	**2**	**1**	**1**	**1**	**1**	**ε**	**I**
G6 d answer	**urn**	**4**	**1**	**u**	**u**	**2**	**3**	**3**	**1**	**1**	**1**	**1**	**ε**	**I**
G6 d reason	**urn**	**4**	**1**	**u**	**u**	**2**	**2**	**3**	**1**	**1**	**1**	**1**	**ε**	**I**
G6 e answer	**urn**	**4**	**1**	**u**	**u**	**2**	**3**	**2**	**1**	**1**	**1**	**1**	**ε**	**I**
G6 e reason	**urn**	**4**	**1**	**u**	**u**	**2**	**2**	**2**	**1**	**1**	**1**	**1**	**ε**	**I**
G7 (a) (i)	language	-	1	u	u	-	3	-	-	-	-	-	θ	-
G7 (a) (ii)	language	-	1	u	u	-	3	-	-	-	-	-	θ	-
G7 (a) (iii)	language	-	1	u	u	-	3	-	-	-	-	-	θ	-
G7 (a) (iv)	language	-	1	u	u	-	3	-	-	-	-	-	θ	-
G7 (b) (i)	language	-	1	u	u	-	2	-	-	-	-	-	θ	-
G7 (b) (ii)	language	-	1	u	u	-	2	-	-	-	-	-	θ	-
G7 (b) (iii)	language	-	1	u	u	-	2	-	-	-	-	-	θ	-
G7 (b) (iv)	language	-	1	u	u	-	2	-	-	-	-	-	θ	-
G7 (b) (v)	language	-	1	u	u	-	2	-	-	-	-	-	θ	-
G8	coin	2	1	u	u	1	3	1	1	1	1	1	β	I
G9	**die**	**3**	**1**	**u**	**u**	**4**	**2**	**2**	**3**	**1**	**1**	**1**	**ζ**	**II**
G10 – G12	na													
G13	language	-	1	u	u	-	2	-	-	-	-	-	θ	-
G14	language	-	1	u	u	-	2	-	-	-	-	-	θ	-
G15	language	-	1	u	u	-	2	-	-	-	-	-	θ	-
G16	language	-	1	u	u	-	3	-	-	-	-	-	θ	-

G17 answer	spinner	7	1	u	u	1	3	2	1	1	1	1	ε	I
G17 reason	spinner	7	1	u	u	1	3	2	1	1	1	1	ε	I
G18	**urn**	**4**	**1**	**u**	**u**	**1**	**3**	**2**	**2**	**1**	**1**	**1**	**δ**	**VII**
G19 (a)	**spinner**	**7**	**1**	**u**	**u**	**1**	**3**	**2**	**1**	**3**	**1**	**1**	**δ**	**I**
G19 (b)	**spinner**	**7**	**1**	**u**	**u**	**1**	**2**	**2**	**1**	**3**	**1**	**1**	**δ**	**I**
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**		
G20 (a)	coin	2	1	u	u	4	2	1	1	1	1	1	η	III
G20 (b)	coin	2	1	u	u	4	2	1	1	1	1	1	η	III
G21 (a)	channels	12	2	u	u	1	3	2	4	3	1	1	δ	II
G21 (b)	channels	12	2	u	u	1	3	2	2	3	1	1	δ	II
G21 (c)	channels	12	2	u	u	1	3	2	2	3	1	1	δ	II
G21 (d)	channels	12	2	u	u	1	3	2	2	3	1	1	δ	II
G22	channels	12	2	u	u	1	2	2	4	2	1	1	δ	I
G23	drawing pin	10	1	u	u	4	3	1	1	3	1	1	β	III
G24	births	12	2	u	u	4	3	1	1	?	?	1	δ	II
G25	urn	4	1	u	u	4	3	u	1	u	2	1	δ	III
G26 (a)–(d)	na													
FGB1	urn	4	1	u	u	1	2	2	2	1	1	1	β	II
FGB2	many	11	1	1	u	4	2	4	1	4	1	2	δ	VIII
FGB3	urn	4	1	u	u	4	2	4	4	1	1	2	δ	III
FGB4	urn	4	1	u	u	4	2	4	4	1	1	2	δ	II
FGB5	urn	4	1	u	u	4	2	4	4	1	1	2	δ	III
FGB6	urn	4	1	u	u	4	3	3	1	1	1	1	ε	I
FGB7	urn	4	1	u	u	4	2	3	1	1	1	1	ε	I
FGB8	urn	4	1	u	u	4	2	3	1	1	1	1	ε	II

Probably not all workers would agree with all these classifications. For example, the channels in G21 & G22 have been treated as single RGs, and for FGB2, which deals with a child's view that entering a classroom right foot first will increase his chance of gaining good marks, it is certainly arguable whether good marks may be seen as an outcome of a random generator. However, the discussion below does not require agreement for each and every classification, so any debate about a small number of difficult classifications should not affect the general conclusions.

Discussion

Table 3 shows that Green's final set of questions are predominantly about comparison of urns, only three items deal with situations more complex than that of considering just one trial, and all but one deal with "likely" outcomes or comparison of RGs. Furthermore, using the classification as a guide to comprehensiveness, shows that although the test covers four different RGs, it does so within restricted contexts and does not address many ideas necessary for a good understanding of probability. So it does not seem appropriate for Izard's international survey unless his concerns were more to do with statistics than assessing probabilistic understanding.

The framework shows clearly just why Green felt that his final test omitted many interesting questions. For example, most of the questions involving asymmetric RGs have been omitted, as well as all the questions directly addressing language and almost all of the questions incorporating outcomes of previous trials

When we consider the FGB questions, we see that the first factor identified by Godino et al. involves all the questions concerned with comparison of random generators (ε), and the second involves questions about "likely" outcomes (δ). Godino et al. see the second factor is concerned with the students' "biases and deep-rooted beliefs"—ideas more likely to be elucidated by questions of this type.

Godino et al. state that Green's test does not contain questions like these second factor questions. But the framework is able to show that G21 does have some similarities, but also differs in not involving hypothetical rewards, large numbers of elementary events, or considering more than one trial. Furthermore, G21 deals with an RG—channels—which may well have been less familiar to students, and is not very clearly explained in the test. So the framework taken with the statistical analysis can assist in clarifying the nature of a statistical claim.

The first six factors identified in Green's test each contain items from different parts of particular questions (in Factor 3, the item "5dr" must be a misprint for "6dr"), sometimes with other items as well. Factor 3 contains G18, which is clearly of a different structure from the other three items, and Factor 4 contains two items of a quite different type. On the other hand, the markedly similar items arising from the comparison of urns in G6 have not all grouped themselves into one factor.

So while factor analysis has identified some factors in both tests which framework analysis might have predicted, it has also highlighted some questions which are surprising linked or not linked, and the framework provides a starting point for examining why these discrepancies might have occurred. For example, although the comparison of urns questions have a standard format they form a single factor in the Israeli test, but separate into different factors on Green's, as well as indicating different Piagetian levels. Green (1982b, p. 338) suspected that children did not consider the problems as being essentially similar. J. & K. Truran (1999) summarised research into such questions and showed that children's responses to slightly different sets of numbers are often idiosyncratic, subconscious, and unpredictable. We do not yet understand the reasons, but it would clearly be unwise, for example, to deduce much from an assessment instrument containing just one "comparison of urns" question, because the development of understanding seems to be non-linear.

The framework approach rests predominantly on an analysis of question format; the statistical approach more on understanding of meaning. Neither is an all-purpose tool, although, as we have seen here, their strengths do overlap to some extent.

Conclusion

Researchers often report new tests for assessing probabilistic understanding, e.g., Batanero, Serrano & Garfield (1996), Reading & Shaughnessy (2000). The framework and statistical analysis can be complementary partners for analysis and building up the item-bank mentioned above. The framework is simple, and applicable *before* administering a test as well as after obtaining statistical results. It can provide a guide to the comprehensiveness of a test or unit of work, which may be of value for analysing the complexity of probabilistic knowledge and deciding whether its acquisition is linear or holistic. While it needs a little modification, further testing, and much extension, this paper has shown how useful a research tool it can be.

References

Alarcon, Jésus (1982) *L'Appréhension des Situations Probabilistes chez des élèves de 12–14 ans: resultats d'une enquête proposée à des élèves de 4ème et de 5ème.* Thesis: Louis Pasteur University of Strasbourg, France

Batanero, Carmen; Serrano, L.; Garfield, Joan B. (1996) "Heuristics and Biases in Secondary School Students' Reasoning about Probability" Pp. 2-51–2-58 in Puig, Luis; Gutiérrez, Angel (eds) *Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education* Valéncia Spain: Universitat de Valéncia

Bell, Gary; Wright, Bob; Leeson, Neville; Geake, John (eds) (1994) *Challenges in Mathematics Education: constraints on construction* Proceedings of the Seventeenth Annual Conference of the Mathematics Education Research Group of Australasia. Lismore, New South Wales, Australia: Mathematics Education Research Group of Australasia

Fischbein, Efraim; Gazit, A. (1984) "Does the Teaching of Probability Improve Probabilistic Intuitions?" *Educational Studies in Mathematics* 15 (1): 1–24

Garfield, Joan (ed.) (1994) *Research Papers from the Fourth International Conference on Teaching Statistics (ICOTS 4).* Minneapolis Minnesota, United States of America: International Study Group for Research on Learning Probability and Statistics [This document is unpaginated.]

Glencross, Michael J.; Laridon, Paul E. (1994) "Understanding of Probability Concepts among South African Children" In Garfield (1994)

Godino, Juan D.; Batanero, Carmen; Cañizares, M.J. (1994) "A Comparative Study of Two Instruments for Evaluating Probabilistic Reasoning" In Garfield (1994)

Green, David R. (1982a) *Probability Concepts in 11–16 Year Old Pupils* Loughborough, United Kingdom: Centre for Advancement of Mathematical Education in Technology, Loughborough University of Technology

—(1982b) *Probability Concepts in School Pupils Aged 11–16 Years* PhD Thesis: Loughborough University of Technology

— (1986) "Children's Understanding of Randomness: report of a survey of 1600 children aged 7–11 years" Pp. 287–291 in Davidson, R.; Swift J. (eds) *Proceedings of the Second International Conference on Teaching Statistics* Victoria, British Columbia, Canada: University of Victoria

Izard, John F. (1992) "Patterns of Development with Probability Concepts: assessment for informative purposes" Pp. 355–367 in Stephens, Max; Izard, John (eds) *Reshaping Assessment Practices: assessment in the mathematical sciences under challenge* Proceedings from the First National Conference on Assessment in the Mathematical Sciences Geelong, Victoria, 20–24 November 1991. Hawthorn, Victoria: Australian Council for Educational Research

Reading, Chris; Shaughnessy, Mike (2000) "Student Perceptions of Variation in a Sampling Situation" Pp. 4-89–4-96 in Nakahara, Tadao; Koyama, Masataka (eds) *Proceedings of the*

24th Conference of the International Group for the Psychology of Mathematics Education Hiroshima: Japan: Hiroshima University

Truran John M. (1994) “Diagnosing Children’s Probabilistic Understanding”. Pp. 623–630 in Bell et al. (1994)

Truran John M.; Truran Kath (1999) “Using a Handbook Model to Interpret Findings about Children’s Comparisons of Random Generators” Pp. 4-281–4-288 in Zaslavsky, Orit (ed.) *Proceedings of the 23rd Conference of the International Group for the Psychology of Mathematics Education* Haifa, Israel: Technion - Israel Institute of Technology

Watson, Jane M.; Collis, Kevin F.; Moritz, Jonathan B. (1994) “Assessing Statistical Understanding in Grades 3, 6 and 9 Using a Short-Answer Questionnaire” Pp. 675–682 in Bell et al. (1994)

Zaleska, Maryla (1974) “Influence du mode de présentation de séries aléatoires sur le choix d’une stratégie: I.ʃSituations hypothétiques” *L’Année psychologique* 74: 125–144