LONGITUDINAL UNDERSTANDING OF CONDITIONAL PROBABILITY BY SCHOOL STUDENTS

Jane M. Watson
University of Tasmania
<Jane.Watson@utas.edu.au>

Jonathan B. Moritz
University of Tasmania
<Jonathan.Moritz@utas.edu.au>

Two survey items asking for estimates of probability or frequency of conditional events, (A|B) and (B|A), were completed by 2719 school students in grades 5 to 11. Cross-sectional and longitudinal analyses revealed improvement with grade in expressing probability numerically and in distinguishing conditional events. Conditional events were better distinguished for the frequency item than the probability item. Comparisons with responses to other probability items indicated understanding of conditional probability was related to development of basic chance measurement.

Historically understanding of conditional probability was explored with tertiary students. Not only were tertiary students readily available as participants in the researchers' classes but also there was a lack of emphasis on probability in the school curriculum generating little interest by mathematics education researchers. Over the last decade, however, mathematics curricula of most western countries have recognised the importance of chance and probability generally and included specific reference to conditional probability. In *Mathematics - A Curriculum Profile for Australian Schools* (Australian Education Council, 1994) secondary students are expected to "assign conditional probabilities based on data in two-way tables" (level 7.23, p. 124). In the *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989) in the United States, recommendations for grades 9 to 12 include addressing probabilistic intuitions (Fischbein, 1975; Kahneman & Tversky, 1972).

> Concepts of probability, such as independent and dependent events, and their relationship to compound events and conditional probability should be taught intuitively. Formal definitions and properties should be developed only after a firm conceptual base is established... (National Council of Teachers of Mathematics, 1989, p. 171)

What it means to teach or learn intuitively however may still be open to debate following the research of Fischbein and others in probability generally (Fischbein, 1975; Fischbein & Gazit, 1984; Fischbein & Schnarch, 1997) and more specifically in relation to conditional probability (Bar-Hillel & Falk, 1982; Falk, 1986). Findings show mixed results in terms of the development of intuitive understanding; for example, Fischbein and Schnarch (1997) found diminished performance in relation to some concepts with increasing age. In discussing misconceptions associated with significance testing, Falk (1986) suggested that much of the difficulty is related to confusing conditional probabilities and equating the probability of rejecting a null hypothesis when it is true, $P(R|H_0)$, with the probability of it being true when it is rejected, $P(H_0|R)$.

The contexts in which conditional problems have been set have varied widely. In questioning middle school students' understanding of replacement and non-replacement aspects of conditional probability, Fischbein and Gazit (1984) and Tarr and Jones (1997) employed events of drawing objects from bags or choosing among alternatives in games. In questioning tertiary students who had not undertaken a statistics course, Pollatsek, Well, Konold, Hardiman, and Cobb (1987) employed events from everyday social settings, such as having green eyes, for which some degree of topic knowledge was assumed. No studies were found that used everyday definitions of events in terms of social characteristics with school students.

The format of response for conditional problems has also varied. Pollatsek et al. (1987) found slightly better performance for items requiring probability estimates than for forced choice items concerning the equality or inequality of P(A|B) and P(B|A). They concluded that asking for "numerical estimates may have resulted in a more careful analysis of the problem" (p. 264). Gigerenzer and Hoffrage (1995), using more complex problems with tertiary students, found that frequency items were easier than probability items. Results from other studies of development of chance measurement for students of grades 3 to 11 (Watson, Collis & Moritz, 1997; Watson & Moritz, 1998) have indicated that younger students often use words or non-normative numerical expressions for measuring chance, and that sex differences favouring males are evident at some secondary grade levels. Thus the expression of response to open-ended tasks and sex differences may be relevant factors in the development of conditional probability reasoning for school students.

The current study of school students' responses to conditional problems investigated not only students' conditional estimates of probabilities or frequencies, but also the ways that students express likelihood estimates. As the items on conditional probability in this study were part of a larger survey, it was not possible to include as many formats and settings for questions as covered by earlier researchers. Based on the experience of Pollatsek et al. (1987), it was decided to use the open-ended format allowing students to produce their own estimates of likelihood. Rather than using in-school contexts such as drawing objects from a bag with and without replacement, out-of-school contexts similar to those of Pollatsek et al. were used in an effort to explore student reasoning in everyday contexts and to reflect the application goals of the school curriculum. Three research questions were of interest. (1) In everyday contexts, do school students interpret conditional events in an appropriate fashion? What alternative interpretations arise and in what formats are responses offered? (2) Does student performance on these tasks improve with age, or differ between cohorts or sexes? (3) Is there an association between performance on conditional probability items and performance on other chance measurement items?

METHOD

Participants

Survey responses were gathered from 2719 students at 20 government primary schools, secondary schools, and matriculation colleges distributed throughout Tasmania. Responses were collected in 1993, 1995, and 1997, and totalled 3730 responses, including 785 responses from the same students surveyed again after a two-year interval, and a further 113 responses from students surveyed three times, in 1993, 1995, and 1997. The numbers of students surveyed from different schools varied across years and grades (see Table 1) due to availability of students. Approximately equal numbers of males and females were surveyed in each year at each grade level. More details of the cross-sectional and longitudinal aspects of the larger study are found in Watson and Moritz (1998).

Items and Procedure

Two short answer items, shown in Figure 1, were based on questions of Pollatsek et al. (1987; Q5, p. 269, and Q3, p. 260). Both items required subjective likelihood estimates of everyday events, Item 1 in frequency form and Item 2 in probability form. For each item, it was expected that part (b) would be estimated as approximately half of the total, whereas part (a) would be estimated as less than half of the total. These items were questions 14 and 16 of a 20-item chance and data written survey (Watson, 1994). The survey was administered to whole class groups during 45 minutes of class time. Some students who did not respond to these or later items, due to time or inclination, were excluded from this analysis.

Figure 1
Two Items Involving Conditional Events

1. Please estimate:
   (a) Out of 100 men, how many are left-handed.
   (b) Out of 100 left-handed adults, how many are men.
2. Please estimate:
   (a) The probability that a woman is a school teacher.
   (b) The probability that a school teacher is a woman.

Coding and Analysis of Responses

Responses were entered into a spreadsheet in the form students wrote them. Responses to each part were then (1) assigned a numerical value for the probability or frequency, and (2) coded according the type of expression used. Chance words were assigned numerical values where possible: "unlikely" and "low" were assigned 0.25, "maybe", "medium" and "average" were assigned 0.50, and "likely" and "high" were assigned 0.75, with modifiers such as "very" being assigned more extreme values. Chance words were not assigned values if responses to different parts could not be differentiated to determine which was intended to have a higher probability value. Responses to each of the two items were then coded according to the numerical relation between responses to parts (a) and (b), as well as the type of expression. Numerical Relations are represented symbolically by *b>a* if P(part(b))>P(part(a)), *b=a* if the numerical values of the two parts were the same, *b<a* if P(part(b))<P(part(a)), and *b=a/2* if P(part(b))=0.5P(part(a)). The last of these numerical relations appeared to be associated with a failure to distinguish a new conditional statement in part (b), and thus, for example, "a=30, b=15" was interpreted as, "a=30 [males are left-handed], b=15 [of these 30 left-handed people are male]." An *undefined* code was entered for responses where a numerical value was not given or assigned for both parts to the item, for example, "not enough info" or "yes Miss Alan is [a woman]" for Item 2. In coding Expressions, whole number answers between 0 and 100 were coded as *frequency* expressions. The *percentage* category refers to responses with "%" attached. The *fraction* category included common or decimal fractions or odds. Using a *word* such as "likely" or answering *yes/no* was common for Item 2 but not Item 1. The few such responses for Item 1 were included in the *mix/other* category. Examples of this last category included: "a=.5 [men are left-handed], b=80% [of left-handed adults are male]" and "a=likely, b=50%."

Students' responses were analysed in three different ways. *(1) Cohort and cross-sectional analyses* using χ^2 tests involved the independent factors of cohort (1993, 1995, 1997), sex, and grade level. Responses from comparable grades collected from different cohorts were compared to investigate whether recent curriculum reform and implementation had affected students' estimates of conditional probability and their expressions of chance. Responses of students from a cross-section of grades (and both sexes) were compared as one method of investigating conceptual development of students. *(2) Longitudinal analysis* was a second method for exploring conceptual development, in this case analysing differences in responses of 113 individual students gathered longitudinally over two 2-year intervals. *(3) Cross-item analyses* involved comparing students' responses to the probability (Item 2) and frequency (Item 1) forms. Responses were also compared to those of chance measurement tasks reported in previous studies (Watson, et al., 1997; Watson & Moritz, 1998) to explore understanding of other probability concepts that may impact on estimating conditional probabilities.

RESULTS

Cohort and Cross-Sectional Analyses

Tables 1 and 2 show the numerical relations and expressions used in responses to Items 1 and 2 respectively, for all 3730 responses grouped by grade and year of survey. The results reported include both those who were completing it for the first time as well as those who were repeating the items in each year. There were significant differences between non-repeating and repeating students (1) in grade 7 for the expressions used for Item 2, (2) in grade 10 for the expressions used for Item 1, and (3) in grade 11 for the numerical relations and expressions used for Item 1; in each case repeating students performed better. It is difficult to determine if these differences indicate repeat effects or if they are due to non-repeating students being drawn from different feeder schools, particularly for those in grades 7 and 11. Further analysis assumed repeating and non-repeating students could reasonably be grouped together.

For Item 1, χ^2 tests for equivalent grades in the different cohorts (grades 6 and 9 in all three years and grades 5, 8, and 11 in 1995 and 1997) indicated no significant differences across the cohorts. There was improvement in performance with grade from grade 5 but this tended to level off from grade 8. Percentages of responses in the undefined category varied from 2% to 10%. At all grades, most students responded with a frequency. Over all students, 17% responded "a=50, b=50" (using various expressions); a further 2% of students gave responses to parts (a) and (b) which summed to the total frequency but were not equal to 0.50.

Table 1
Percentage Responses to Item 1 Coded by Numerical Relation and Expression

Response	1993 Grade		1995 Grade					1997 Grade						
category	6	9	5	6	8	9	11	5	6	7	8	9	10	11
Numerical Relation														
b>a	16	29	8	10	25	27	39	11	15	15	23	20	32	41
b=a	44	52	38	46	57	45	50	41	46	51	56	63	56	49
b<a	4	2	6	6	4	1	4	4	6	4	3	1	2	0
b=a/2	0	0	0	1	1	0	1	0	0	0	0	1	0	2
Undefined	35	16	48	38	13	27	7	44	33	30	18	15	9	8
Expression														
Frequency	3	3	5	5	4	1	1	8	6	4	2	4	3	2
Percentage	30	42	23	34	54	36	54	17	31	38	50	45	51	45
Fraction	14	28	10	12	17	23	30	16	13	11	17	11	26	43
Word	15	10	15	12	10	11	4	14	16	15	11	20	9	0
Yes/No	18	4	27	18	4	12	1	27	19	17	5	4	3	0
Mix/Other	20	14	20	19	11	17	9	18	15	16	15	16	8	10
N	307	377	435	321	361	345	215	216	215	305	186	104	292	51

For Item 2, there were significant differences between grade 9 cohorts for both numerical relation and expressions; there were more *Yes/No* expressions in 1995, and more *b=a* numerical relations in 1997, as can be seen in Table 2. The level of correct performance increased with grade from 8% for one grade 5 group to 41% for one grade 11 group. Responses for all grades were more likely to be in the *b=a* category than the *b>a* category. Percentages or fractions were the dominant form of response for older students, *Yes/No* responses were common at grades 5, 6 and 7, whereas chance words were used by between 10-20% of students until grade 11. Over all students, 23% responded "a=0.50, b=0.50" (using various expressions), increasing with grade from 12% at grade 5 to 33% at grade 11; this trend was observed even after removing undefined responses from analysis. A further 2% of all students gave responses to parts (a) and (b) which summed to the total frequency but were not equal to 0.50.

Table 2
Percentage Responses to Item 2 Coded by Numerical Relation and Expression

Response	1993 Grade		1995 Grade					1997 Grade						
category	6	9	5	6	8	9	11	5	6	7	8	9	10	11
Numerical Relation														
b>a	16	29	8	10	25	27	39	11	15	15	23	20	32	41
b=a	44	52	38	46	57	45	50	41	46	51	56	63	56	49
b<a	4	2	6	6	4	1	4	4	6	4	3	1	2	0
b=a/2	0	0	0	1	1	0	1	0	0	0	0	1	0	2
Undefined	35	16	48	38	13	27	7	44	33	30	18	15	9	8
Expression														
Frequency	3	3	5	5	4	1	1	8	6	4	2	4	3	2
Percentage	30	42	23	34	54	36	54	17	31	38	50	45	51	45
Fraction	14	28	10	12	17	23	30	16	13	11	17	11	26	43
Word	15	10	15	12	10	11	4	14	16	15	11	20	9	0
Yes/No	18	4	27	18	4	12	1	27	19	17	5	4	3	0
Mix/Other	20	14	20	19	11	17	9	18	15	16	15	16	8	10
N	307	377	435	321	361	345	215	216	215	305	186	104	292	51

Table 3 shows the performances of female and male students at each grade level (cohorts combined) in a consolidated form (*b>a*, *other*, *undefined*) for the two items. Comparing females with males at each grade for numerical relations to both items, the only significant difference was at grade 10, with males more likely than females to respond *b>a* to Item 1 (χ^2_2=12.9, $p<0.01$).

Table 3
Percentage Responses to Items 1 and 2 by Grade and by Sex

Response	Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10		Grade 11	
category	f	m	f	m	f	m	f	m	f	m	f	m	f	m
Item 1														
b>a	43	46	48	47	58	55	55	61	61	67	48	64	62	65
Other	52	44	46	46	38	38	43	34	35	28	51	32	33	28
Undefined	5	10	6	6	4	6	2	5	5	5	1	3	5	7
Item 2														
b>a	8	10	12	15	15	16	21	28	24	31	25	39	41	37
Other	43	46	52	50	54	55	64	57	53	52	64	54	51	58
Undefined	50	45	36	35	31	28	15	15	24	17	12	7	8	6
N	317	334	432	411	150	155	273	274	425	401	146	146	143	123

Longitudinal Development

Table 4 presents the percentages of students in each response category for 113 students who were surveyed in each of the three years of the study. This subset of students closely reflects the whole cohort in each year (grade 6 in 1993, grade 8 in 1995, and grade 10 in 1997). For the appropriate response *b>a*, there was improvement from grade 6 to 8 but little after that. To examine within-student longitudinal change, results were consolidated by grouping numerical relations into *b>a*, *other*, and *undefined*. For Item 1, 33 students gave the *b>a* response three times and 12 consistently gave an *other* response. For Item 2, 12 students gave the *b>a* response three times and 28 consistently gave an *other* response. In two 2-year intervals, 1993-1995 and 1995-1997 (the latter in parentheses for the following results) for Item 1, 65 (or 72) students remained in the same grouping, 23 (or 20) students improved their response from *other* to *b>a*, whereas 11 (or 17) reverted from *b>a* to *other*. Similarly, for Item 2, 56 (or 69) students remained in the same grouping, 10 (or 15) students improved their response from *other* to *b>a*, and 10 (or 9) reverted from *b>a* to *other*. Thus overall, the longitudinal study indicates that students from grade 6 to grade 8 developed in appreciating conditional events for Item 1 and in numerically expressing chance for Item 2, but from grade 8 to grade 10 percentages of *b>a* and *other* response were quite stable, with fluctuations of improvement and reversion in similar frequencies.

Table 4
Percentage Responses Assessed Longitudinally in Grade 6, 8 and 10 (n=113)

Response	Item 1			Item 2		
category	Grade 6	Grade 8	Grade 10	Grade 6	Grade 8	Grade 10
Numerical Relation						
b>a	46	61	65	22	30	36
b=a	31	25	26	43	58	52
b<a	9	5	4	4	3	2
b=a/2	4	6	5	1	0	0
Undefined	10	3	1	30	10	10
Expression						
Frequency	64	64	77	2	4	3
Percentage	19	30	16	35	59	58
Fraction	4	0	4	16	16	19
Word	-	-	-	13	7	11
Yes/No	-	-	-	12	3	3
Mix/Other	14	6	4	22	11	7

Cross-Item Analyses

Of 3730 responses, *b>a* was more common for Item 1 (2042 responses) than Item 2 (771 responses), *other* was more common for Item 2 (1967 responses) than Item 1 (1492 responses), and *undefined* was again more common for Item 2 (992 responses) than Item 1 (196 responses). Overall only 16% of responses were classified *b>a* to both items. The relationship of answers to the two questions was statistically significant for the combined data set: 29% of those who responded *b>a* for Item 1 replied similarly on Item 2, whereas only 11% of those who gave an other response to Item 1, responded *b>a* on Item 2.

To explore the association between success on conditional estimate items and more general understanding of chance measurement, 3616 responses to Items 1 and 2 were matched to developmental levels determined by Watson and Moritz (1998), scored from 0 to 6, based on responses to three chance measurement items earlier in the survey. These levels represent increasingly complex cognitive functioning evident across responses to three items concerning (1) likelihood of numbers occurring when a 6-sided die is rolled, (2) likelihood for an outcome drawn from a bag, and (3) comparisons of likelihoods of drawing one colour of marbles from boxes with similar ratios of colours. The distribution of responses to Items 1 and 2 by chance measurement developmental level is shown in Table 5. Higher chance measurement developmental levels were associated with more *b>a* responses and fewer *undefined* responses for both Item 1 and Item 2; *other* responses decreased with increasing developmental level for Item 1, but not for Item 2.

Table 5
Percentage Responses to Items 1 and 2 by Chance Measurement Level

Response	Chance Measurement Level							
category	0	1	2	2.5	3	4	5	6
Numeric Relation (Item 1)								
b>a	33	43	47	53	55	58	67	78
Other	53	49	46	41	41	38	28	20
Undefined	14	8	6	5	4	4	5	2
Expression (Item 1)								
Numerical	86	92	93	94	96	96	95	98
Numerical Relation (Item 2)								
b>a	10	6	11	14	19	30	31	47
Other	31	41	49	53	60	56	53	45
Undefined	59	53	39	33	21	14	16	8
Expression (Item 2)								
Numerical	27	35	47	54	65	74	78	85
N	49	177	543	889	653	914	122	269

DISCUSSION

Conditional probability poses a number of difficulties for school students, including expressing chance and identifying the nature of conditional events in different grammatical constructions and contexts. With respect to expressing chance, the use of "Yes", "No" and chance words diminished with increasing grade for the probability item, with older students more likely to express probability numerically. Not surprisingly, numerical expression of chance for this item was also associated with chance measurement level based on other traditional simple probability items. For the frequency item, about 90% of students at all grades responded with numerical expressions.

Identifying the nature of conditional events ($b>a$) for each item improved only marginally with grade, with sex differences favouring males at grade 10. There was, however, clear improvement with increasing chance measurement level. This association is particularly interesting for the frequency item where the association is not confounded by effects of numerical expression for the probability item. Identifying the nature of conditional events is related to responses to probability tasks that involve identifying relevant numbers to measure chance as a fraction, namely the subset as a proportion of the total set.

Identifying the nature of conditional events also depended on the context and grammatical structure of the items. Even those students who could readily express probabilities numerically where more likely to respond $b=a$ for Item 2 than for Item 1. These results differ markedly from the results of Pollatsek et al. (1987). When asked to give estimates for the same question as Item 2, Pollatsek et al. found 75% of college students responded $b>a$ and 17% responded $b=a$, whereas in the current study, only 39% of grade 11 students responded $b>a$ and 50% responded $b=a$. It may be that in the probability form of the questions, younger students fail to distinguish the language indicating the two conditional events as being distinctive. The sentences are also more succinct for Item 2 than for Item 1. Item 1 differed from Pollatsek et al. in dealing with left-handedness rather than having green eyes, and in asking for a frequency rather than a probability or percentage. In the green-eyed setting, college students estimated $b>a$ in 45% of cases. Over 50% of students grade 7 or higher in the current study responded $b>a$, reaching a maximum of 63% in grade 11. It is impossible to speculate on the effect of the left-handed/ green-eyed distinction, but it appears more likely that the better performance was associated with the frequency form of Item 1 (cf. Gigerenzer & Hoffrage, 1995) that provides the cue "out of" to make clear the subset conditional relation.

Pollatsek et al. (1987) also noted a consistently large percentage of students (from 18% to 44% on seven different items) giving responses in the open format which summed to 100%, possibly indicating a belief that the events were complementary. It is not clear whether they distinguished the half-half responses from others as done in this study. The combined data from the current study produce values consistent with Pollatsek et al. but further research is required to determine whether these students are responding in the belief that the events are the same, or that the events are complementary.

Classical probability problems, such as involving balls in urns, have long been used in classrooms, and these contexts cue students to elements that are countable. A strong relationship has been shown to exist for reasoning on such problems (Watson & Moritz, 1998) and success on conditional problems. Questions in social contexts, however, ask students to use their contextual knowledge of the environment rather than numbers provided. It is not surprising that many responses about conditional probabilities in a social context were expressed in words and did not appropriately distinguish conditional events. By asking for frequencies in a social context, however, student responses improved markedly. Educational programs in this area should include exposure to social contexts if useful learning is to occur. Overall it appears likely that without a concentrated effort in the school curriculum to assist students to transfer their probabilistic understanding from countable situations to more social settings involving estimation, reasoning or appropriate intuition (Fischbein, 1975) it will not continue to develop.

REFERENCES

Australian Education Council. (1994). *Mathematics - A curriculum profile for Australian schools*. Carlton, Vic.: Curriculum Corporation.

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition*, *11*, 109-122.

Falk, R. (1986). Misconceptions of statistical inference. *Journal of Structural Learning*, *9*, 83-96.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Dordrecht: D. Reidel.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? An exploratory research study. *Educational Studies in Mathematics*, *15*, 1-24.

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

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684-704.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430-454.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Pollatsek, A., Well, A., Konold, C., Hardiman, P., & Cobb, G. (1987). Understanding conditional probabilities. *Organizational Behaviour and Human Decision Processes*, *40*, 255-269.

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.

Watson, J. M. (1994). Instruments to assess statistical concepts in the school curriculum. In National Organizing Committee (Ed.), *Proceedings of the Fourth International Conference on Teaching Statistics. Volume 1* (pp. 73-80). Rabat, Morocco: National Institute of Statistics and Applied Economics.

Watson, J. M., Collis, K. F., & Moritz, J. B. (1997). The development of chance measurement. *Mathematics Education Research Journal, 9*, 60-82.

Watson, J. M., & Moritz, J. B. (1998). Longitudinal development of chance measurement. *Mathematics Education Research Journal*, *10*(2), 103-127.

Acknowledgement

This research was funded by the Australian Research Council, Grant Numbers A79231392 and A79800950.