Review of Psychology,
2007, Vol. 14, No. 2, 109-114 UDC 159.9

Teaching probability: effects of task frames and training on misconceptions

FRANCESCA CHIESI, SILVIA GALLI, GIORGIO GRONCHI and CATERINA PRIMI

The aim of the present work is to improve knowledge about how students think about probability in order to find educational methods intended to eliminate misconceptions. The role of task frames and training on reasoning about sequences of independent outcomes was investigated. Three types of frames have been compared: abstract, concrete, and concrete plus reference to stereotypes. Tasks were administered to two groups of students: one group had followed a brief course on probability, the other had not. Frames influenced the use of normative strategies as well as training, whereas some students, even after the training, continued to hold misconceptions. In particular, the use of normative strategies was more likely in abstract frames. This effect held both for training and non-training group. However, training group was more prone to avoid mistakes in concrete plus reference to stereotypes frames. Results suggest that these tasks can be used to illustrate normative responses and biases in teaching probability.

Key words: statistics education research, probabilistic reasoning, assessment

Many major subjects require familiarity with statistical tools such as collecting, organizing, and presenting data, as well as reasoning about probability and drawing inferences. Unfortunately, many students enter college with very little formal experience with the laws of probability and probabilistic reasoning. Moreover, they may develop their own way of reasoning about uncertain events through informal experiences.

Literature on probabilistic intuitive reasoning has shown that adults as well as children committed several errors when problems were referred to everyday life frames (Davidson, 1995; Jacobs & Potenza, 1991; Kahneman, Slovic, & Tversky, 1982; Kunda & Nisbett, 1986). It was supposed that real life frames could hide relevant information for the correct resolution and that concrete situations (e.g. sports, games) can activate non-relevant cues compared to abstract situations (e.g. tossing a coin). Moreover, as in the previous case, real life frames with references to stereotypes (e.g. gender stereotype) are often associated with non-normative strategies (Bodenhausen & Wyer, 1985; Hamilton & Gifford, 1976; Schaller, 1992), because judgments are based on stereotypical information that can cause misconceptions. A concrete early illustration of the effect of real life frame was provided by Tversky and Kahneman (1983). A trivial rule of probability theory asserts that the probability of a single event (A) is always greater than, or at least equal to the probability of the intersection of the same event with another event (A&B). However, giving a description like "Bill is 34 years old ... an intelligent, but unimaginative, compulsive, and generally lifeless man who was strong in school in mathematics but weak in social studies and the humanities", people were likely to give a higher rating to the probability P(A&B) = "Bill is an accountant who plays jazz for a hobby" than to the probability P(A) = "Bill plays jazz for a hobby" (Tverksy & Kahneman, 1983). The representativeness of the stereotypical information can explain why people neglect the normative principle.

Among the various probabilistic misconceptions, this paper focused on the gambler's fallacy (Boynton, 2003; Kahneman, Slovic, & Tversky, 1982; Loewenstein & Prelec, 1993; Yackulic & Kelly, 1984; Winefield, 1966), a mistake related to probabilistic judgment about sequences of independent outcomes. This bias emerges when probability is estimated on the basis of what happened before, ignoring the independence of events. For example, referring to the Black (B) and Red (R) outcomes of a roulette wheel, we

Francesca Chiesi, University of Florence, Department of Psychology, Florence, Italy. E-mail: f.chiesi@tin.it;

Silvia Galli, University of Florence, Department of Psychology, Florence, Italy. E-mail: silviag79@katamail.com;

Giorgio Gronchi, University of Florence, Department of Psychology, Florence, Italy. E-mail: giorgio.gronchi@tin.it;

Caterina Primi, University of Florence, Department of Psychology, Via S. Niccolò 89/a – 50125, Florence, Italy. E-mail: cprimi@texnet.it (the address for correspondence).

have to judge the more likely event after a sequence as 'R R R R'. According to the normative principle, each spin is independent from the other and the probabilities of R and B remain the same, regardless which sequence of outcome has been produced. According to the non-normative reasoning, B outcome can erroneously be considered more likely than R one, i.e. gambler fallacy. Inside statistics education research, Hirsch & O'Donnel (2001) proposed a valid and reliable test instrument to identify students who hold misconceptions related to this kind of problems. Their test proposed 16 two-part items with abstract scenarios (e.g. tossing a coin or rolling a dice). In the first part of an item, students choose the correct answer to a problem stem from among five options. In the second part, students justified their answer in part one by selecting from a number of explanations. Based on their responses to both parts of each pair of items, they found that about the 60% of the whole sample answered correctly.

Misconception may be due to a lack of experience with mathematical laws, for instance, some studies (Fong & Nisbett, 1991; Fong, Krantz, & Nisbett, 1986; Lehman, Lempert, & Nisbett, 1988) reported that experts in statistics and probability theory are more likely to adopt normative strategies of resolution, but there is an evidence that misconceptions about the laws of chance are still common among people who received a formal training (Kahneman & Tversky, 2000). Hirsch & O'Donnel (2001), administering the test described above to assess the effects of instructional interventions, reported that although formal instruction appears to reduce the proportion of students who gave incorrect answers, a substantial number of students with formal training continues to have misconceptions.

In his reflections on teaching statistics and probability, Shaughnessy (1992) stressed the need to develop consistent methods of assessment that more accurately reflect student's conceptual understanding. As a matter of fact, the more accurately students' conceptual understanding of probability is assessed, the more effective instructional methods intended to eliminate misconception can be proposed (Garfield, 2002; Konold, 1991, 1995). However, few studies in the educational assessment field examined carefully the scenarios and their role in inducing different strategies of resolution, i.e. normative vs non-normative.

Following these assumptions, the aim of present research was to examine how students think about probability, to ascertain when misconceptions occur, and to explore the effect of formal training on probabilistic reasoning. In particular, it was hypothesized that problems with different frames can be helpful to assess both the normative and non-normative reasoning, and that correct answers and fallacies could be related to the different task scenarios. We are expecting that participants will use more normative strategy in tasks with abstract scenario and vice versa as the literature suggests (Bodenhausen & Wyer, 1985; Hamilton & Gifford, 1976; Schaller, 1992). In addition, formal training could be effective in reducing errors depending on the scenarios. For instance, it can be thought that the instruction they received was effective and that they use it independently from the problem frames. Alternatively, students could maintain an own way of reasoning about uncertain events characterized by real life scenarios, whereas they might apply correctly the normative rule in abstract scenarios.

We employ as a starting point the test created by Hirsch & O'Donnel (2001). They proposed 16 multiple choice items with abstract scenarios, e.g. classical frame employed in teaching probability (rolling a dice, tossing a coin, picking an object from a bag without describing who rolls the dice, why, and his/her motivation). Based on the above, tests formally similar but with two different scenarios (concrete and stereotypical) were developed. Concrete scenario presents situations in which a random experiment was held, for example, the owner of a factory that wants to select randomly an employee. Concrete with stereotypical information was very similar to the previous frame, for example, the owner of a factory that wants to select an employee among Italian and immigrant workers.

In sum, we aimed to investigate the occurrence of normative and non-normative responses in students that had received vs. did not have received a brief training in probability.

METHOD

Participants

Undergraduate students in Psychology major (N = 473, 83.9% female, average age 20.8 ± 4.2) attending introductory statistics courses at the University of Florence participated in the study. Participation was voluntary. Students were offered extra credit to complete the test.

Measure

We prepared a test in three different versions, each one containing six items. The Abstract version was derived directly from the instrument created by Hirsch & O'Donnel (2001) choosing 6 items from the original version. The other two versions (Concrete and Stereotypical) were developed keeping constant the formal content of the problems of the Abstract version but changing the scenario. An index of readability derived from Gunning Fog Index (Gunning, 1952) and adapted for the Italian language (total number of words divided by all independent clauses, and added the number of difficult words defined as those with five or more syllables) was calculated to evaluate the formal equivalence among the three tests. A small sample (n = 9) was administered the three versions and requested to estimate the clearness of the different contents that resulted equally simple to understand. At last, the three versions were characterized by the same reading and understanding difficulties.

Each test item was composed of two parts. The first part asked students for an assessment of probability. The second part asked students to identify specific justification for their answer to the first part. Both answer and justification were given in a multiple-choice format (one correct among five choices).

The Abstract version employed problems that present classical situations used in teaching probability. For example:

If a fair coin is tossed ten times, which of the following ordered sequences of heads (H) and tails (T), if any, is most likely to occur?

H H H T H T H H T H

H H H H H H H H H H

H T H T H T H T H T

T T H T H T T T H H

All sequences are equally likely.

Which of the following best describes the reason for your answer to the preceding question?

Since tossing a coin is random you should not get a long string of heads or tails.

Every sequence of ten tosses has exactly the same probability of occurring.

There ought to be roughly the same number of tails as heads.

Since tossing a coin is random, the coin should not alternate between heads and tails.

Other.

The normative criterion solution claims that for any new toss we have one favorable event and two possible events (p=1/2), independently of what happened before. Non-normative strategies conduct to estimate the second event more or less likely to occur on the basis of the fist one, producing the gambler fallacy bias.

In the same way the Concrete and Stereotypical version presented items formally equivalent in which independence was implicitly assumed. For example:

Piero has taken a history test (composed by 10 true/false questions) responding at chance. Which of the following ordered sequences of correct or wrong responses, if any, is most likely to occur? (Concrete).

Medina, a girl belonging to the gipsy community, has taken a history test (composed by 10 true/false questions) responding at chance. Which of the following ordered sequences of correct or wrong responses, if any, is most likely to occur? (Stereotypical).

As in the Abstract version, items were followed by answer and justification in a multiple-choice format. For example:

W W W C W C W W C W

W W W W W W W W W W

W C W C W C W C W C

C C W C W C C C W W

All sequences are equally likely.

Which of the following best describes the reason for your answer to the preceding question?

Since responding at chance you should not get a long string of correct or wrong responses.

Every sequence of ten responses has exactly the same probability of occurring.

There ought to be roughly the same number of correct or wrong responses.

Responding at chance, the responses should not alternate between correct or wrong.

Other.

Finally, each version was composed of six items that were presented in a random order.

Procedure

Each participant was administered one of the three versions of the test during course activity: 282 students (Non-training group) received the test before following lessons on probability (in detail, 94 students received the Abstract, 91 the Concrete, and 97 the Stereotypical version), while 191 students (Training group) received the test after following lessons on probability (in detail, 66 received the Abstract, 65 the Concrete, and 60 the Stereotypical version). Students were randomly assigned to the two groups and, within each groups, they received randomly different test version. Training group attended four lessons (each one 2 hours long). Lessons focused on basic rules of probability calculus (random experiments, probability of simple events, probability of mutually exclusive and mutually non-exclusive events, probability of independent and dependent events) and examples made by the teacher were different from the experiment tasks. Moreover, no special attention was given to dispelling students' reliance on their intuitive reasoning and to compare intuitive versus normative reasoning.

Students were explicitly told that the study focused on probabilistic reasoning and that the test was not administered for assessment purpose. They were requested to complete the task individually. There was not time limit and the test took 15 minutes on the average.

Successively, results were reported to the students in order to explain the rationale of the research.

RESULTS

Examining the two parts of each item, a score based on the combined response to both parts was obtained. The item was scored as correct if the justification was consistent with the probability assessment (e.g. answer: “All sequences are equally likely”, and justification: “Every sequence of ten tosses has exactly the same probability of occurring”). A total score for the test was obtained by summing the number of correct responses (ranged from 0 to 6 points). The standardized values of skewness (z = -13.93, *p*< .001) and kurtosis (z = 7.46, *p*< .001) indicated a strong departure from normal distribution (Tabachnick & Fidell, 1989). Therefore, the scores were transformed into a dichotomous variable (labeled Probabilistic Reasoning), i.e. correct responses for the whole test vs. at least one wrong response.

Probabilistic Reasoning percentages were calculated for Non-training group and Training group in the Abstract version (74% and 78% respectively), in the Concrete (57% and 61%), and the Stereotypical version (46% and 63%) (Figure 1).

A hierarchical log-linear analyses was performed considering three variables: Test Version [V] (Abstract, Concrete, and Stereotypical), Training [T] (Non-training, Training), and Probabilistic Reasoning [R] (Correct, Wrong). Table 1 presents Likelihood-Ratio chi-square values with the relative degrees of freedom for the models with a good data fit (*p*> .10) (Knoke & Burke, 1980). Since it was found that more than one model provided an adequate fit to the data, differences between Likelihood-Ratio chi-square values were used to compare the nested models. The $\chi 2$ comparison statistics indicated that the Model 3 was the most informative model that described the association in the data.

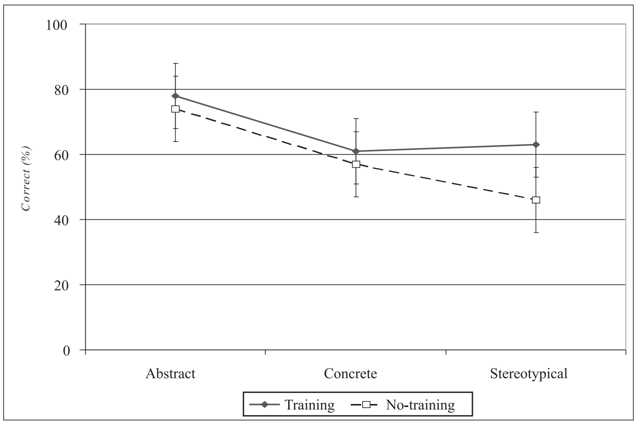

Figure 1. Percentages of correct responses (with 95% confidence intervals) in Probabilistic Reasoning related to task frames and training.

Table 1
Goodness-of-fit and comparisons among models

Model	χ^2	*df*	*p*	$\Delta\,\chi^2$	Δ*df*	*p*
1. [T] [V,R]	5.48	5	.36	-	-	-
2. [T,V] [V,R]	5.01	3	.17	$\Delta_{(1-2)} = 0.47$	2	n.s.
3. [V,R] [T,R]	1.64	4	.80	$\Delta_{(1-3)} = 3.84$	1	.05
4. [T,V] [V,R] [T,R]	1.35	2	.51	$\Delta_{(3-4)} = 4.13$	2	n.s.

Table 2
The effect parameter estimates for the interactions between Probabilistic Reasoning and test Version

		Abstract	Concrete	Stereotypical
Probalistic Reasoning	Correct	4.26**	-1.36	-3.22*
	Wrong	-4.26**	1.36	3.22*

Note. *p< .01 for a monodirectional hypothesis (*df* = 2); ** *p*< .001 for a monodirectional hypothesis (*df* = 2).

Table 3
The effect parameter estimates for the interactions between Probabilistic Reasoning and Training

		Non-training	Training
Probabilistic Reasoning	Correct	-1.95*	1.95*
	Wrong	1.95*	-1.95*

Note. *p< .05 for a monodirectional hypothesis (*df* = 1).

It order to further investigate the effects included in the selected model, parameter estimates were calculated. Among principal effects, Probabilistic Reasoning (z = 5.82, *p*< .001 for Correct) was significant. The interaction effect between Version and Probabilistic Reasoning (Table 2) indicated that Abstract version was associated to a higher probability of Correct, while Stereotypical Version was associated to a majority of wrong responses. Indeed, no significant effect was observed for Concrete version.

The interaction effect between Training and Probalistic Reasonig (Table 3) indicated that Training group obtained an overall majority of correct responses in Probabilistic Reasoning.

DISCUSSION

The aim of the present study was to assess students' conceptual understanding of probability using different frame tasks that was hypothesized to be helpful to assess both the normative and non-normative reasoning, and exploring the effects of formal training on probability.

In line with Hirsch & O'Donnel (2001), it was found an overall majority of correct responses. Participants were able to indicate the right answers and they correctly provided a reason for their answers. Nevertheless, there was an effect of task frames that affected students' reasoning. Several errors occurred when problems were referred to everyday life frames with references to stereotypes in which solutions were more frequently associated with non-normative strategies as well documented by literature on probabilistic intuitive reasoning (Bodenhausen & Wyer, 1985; Davidson, 1995; Jacobs & Potenza, 1991; Kahneman et al., 1982; Kunda & Nisbett, 1986). These results obtained with college students suggested that the instruction they received was not effective in terms of conceptual understanding of probability, while they appeared to be able to use correctly the laws of probability with abstract task, they continue to hold misconception when the formal content of the task was maintained (i.e. the application of the same rule of probability was requested) but the scenarios was modified.

As documented before (Fong & Nisbett, 1991; Fong et al., 1986; Lehman et al., 1992), formal training appeared to reduce the proportion of errors. Students with formal training were more likely to adopt normative strategies of resolution compared to students without formal training. Although a substantial number of students with formal training continued to hold misconceptions (Hirsch & O'Donnel, 2001; Kahneman & Tversky, 2000) especially in task with social or stereotypical information that still inhibited the normative resolution. In other words, some students maintained a personal way of reasoning about uncertain events characterized by real life scenarios.

Some suggestions to improve statistics and probability teaching could be derived from these preliminary results. First, tests with different frames can be helpful for teachers to assess real understanding and comprehension of statistics and probability principles. In other words, these methods can help inform instructors about the actual level of students' probabilistic reasoning. Employing only the common abstract scenario problems (widely used to illustrate probability theory in classroom) to assess students' knowledge can be misleading. It is possible that they apply the rule in some routine way without an actual understanding of the structure of the problem. This limitation can be avoided by employing scenarios in which the problem structure had to be found beyond the contest in order to apply the principles taught correctly.

Second, starting from the result showing that violations of laws of probability continued to occur even after formal training, these tasks could be used for instructional interventions designed specifically to eliminate students' misconceptions. During typical classroom activities, tasks with the same formal contest but with different frames can be proposed and compared to stress biases in probabilistic reasoning. Students can simultaneously give the right answer to explicit probability rules application and show wrong responses due to very strong beliefs about non-relevant facets of the problem. Underlying this paradoxical behavior through ad hoc examples could help distinguish relevant vs. non-relevant information. Further investigations are needed to define and validate appropriate teaching strategies to promote probability learning.

REFERENCES

Bodenhausen, G. V., & Wyer, R. S. (1985). Effects of stereotypes in decision making and information-processing strategies. *Journal of Personality & Social Psychology, 48(2),* 267-282.

Boynton, D. M. (2003). Superstitious responding and frequency matching in the positive bias and gambler's fallacy effects. *Organizational Behavior & Human Decision Processes, 91(2),* 119-127.

Davidson, D. (1995). The representativeness heuristics and the conjunction fallacy in children decision making. *Merrill Palmer Quarterly, 41(3),* 328-346.

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120(1*), 34-45.

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.

Gunning, R. (1952). *The Technique of Clear Writing*. New York: McGraw-Hill.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistical Education* [Online journal], *10(3).* Retrieved May 5, 2005 from http:// www.amstat.org/ publications/jse/ v10n3/garfield.html.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypical judgments. *Journal of Experimental and Social Psychology, 12*, 392-407.

Hirsch, L. S., & O'Donnel, A. M. (2001). Representativeness in statistical reasoning: Identifying and assessing misconceptions. *Journal of Statistics Education* [Online journal], *2(9).* Retrieved September 5, 2003 from www.amstat.org/publications/jse/v9n2/hirsch.html.

Jacobs, J. E., & Potenza, M. (1991). The use of judgment heuristics to make social and object decision: A developmental perspective. *Child Development, 62(1),* 166-178.

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

Kahneman, D., & Tversky, A. (Eds.) (2000). *Choices, values and frames*. Cambridge: Cambridge University Press.

Knoke, D. & Burke, P.J. (1980). *Log-linear Models*. Newberry Park, California: Sage Publications.

Konold, C. (1991). Understanding students' beliefs about probability. In E. Von Glasersfeld (Eds.), *Radical Constructivism in Mathematics Education* (pp. 139-156). Amsterdam: Kluwer.

Konold, C. (1995). Issues in Assessing Conceptual Understanding in Probability and Statistics. *Journal of Statistics Education* [Online journal], *3(1)*. Retrieved September 5, 2003 from http://www.amstat.org/publications/jse/v3n1/konold.html.

Kunda, Z., & Nisbett, R. E. (1986). The psychometrics of everyday life. *Cognitive Psychology, 18*, 195-224.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training in reasoning: Formal discipline and thinking about everyday life events. *American Psychologist, 43*, 431-442.

Loewenstein, G. F., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review, 100(1),* 91-108.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. Grouws (Ed.), *Handbook of research for mathematics education* (pp. 115-147). New York: Macmillan.

Schaller, M. (1992). In-group favoritism and statistical reasoning in social inference: Implications for formation and maintenance of group stereotypes. *Journal of Personality and Social Psychology, 63(1)*, 61-74.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics*. New York: Harper Collins.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review, 90*, 293-315.

Winefield, A. H. (1966). Negative recency and event dependence. *Quarterly Journal of Experimental Psychology, 18*, 47-54.

Yackulic, R. A. & Kelly, I. W. (1984). The psychology of the "gambler's fallacy" in probabilistic reasoning. *Psychology: a Journal of Human Behavior, 21(3-4),* 55-58.