Using data bases to test methods for decisions under uncertainty

Raphael T. Haftka and Raluca I. Rosca
Mechanical and Aerospace Engineering Department, University of Florida, Gainesville, FL 32611-6250

Efstratios Nikolaidis
Mechanical, Industrial and Manufacturing Engineering, University of Toledo, Toledo, OH 43606

Abstract. To address the need for efficient and unbiased experimental testing of methods for decision under uncertainty, we devise an approach for probing weaknesses of these methods by running numerical experiments on readily available or easily obtainable databases of real life data. Since the approach uses real life data, it allows us to study the effect of modeling error on the performance of a method. For illustration, we apply probabilistic and possibilistic approaches to a database of results of a domino tower competition. The experiments yielded several surprising results. First, even though a probabilistic metric of success was used, there was no significant difference between the rates of success of the probabilistic and possibilistic models. Second, the common practice of inflating uncertainty when there is little data about the uncertain variables shifted the decision differently for the probabilistic and possibilistic models, with the latter being counter-intuitive. Finally, inflation of uncertainty proved detrimental even when very little data was available.

1. Introduction

Engineering design decisions commonly involve mathematical models, such as models for calculation of stresses in structural design, to help decision makers predict the outcomes of alternative courses of action. Errors in models are usually investigated experimentally, such as in aircraft certification tests, and occasionally such tests reveal weaknesses in the underlying models. An example of such weakness is sensitivity of failure loads to inevitable small imperfections in geometric shape.

Uncertainty affects the ability of a decision maker to make good decisions. Increasingly, uncertainty is taken into account in design decisions using models, such as probability distributions. Again, design decisions that are sensitive to errors in models of uncertainty may look good on paper but may be very poor in reality. However, there has been little work on using experiments to probe for such sensitivity or other weaknesses in methods or practices for building models of uncertainty.

Examples include the Dartboard Contest, conducted by *The Wall Street Journal* (*WSJ*), (see Greene and Smart, 1999), which compared active and passive investing. In the contest, experts (analysts or fund managers) competed with the *WSJ* staff, which selected stocks by throwing darts at a printout of the *WSJ* stock tables. *WSJ* reported that experts won 61 percent of 140 contests. Baer and Gensler (2002) re-analyzed the study, accounting for additional factors including dividends and risk (experts favored high-risk stocks). With these factors included, passive investing (throwing darts) turned out to be as good as active investing.

Walley (1991, pp. 632-638) conducted an experiment using data from the 1982 Soccer World Cup to compare Bayesian and imprecise (upper and lower) probabilities in making decisions about gambles on games. Of 17 participants, those who used upper and lower probabilities did better than one participant who used Bayesian probability. Participants whose probabilities of the three outcomes (win, lose, tie) of each game were uniform did better than those participants whose probabilities were far apart.

Winkler (1971) and de Finetti (1972) performed experiments to investigate how people assess precise probabilities. Like Walley, Winkler observed that more uniform probabilities tend to improve the degree of success. Walley's and Winkler's studies suggest that if one has low confidence in the probabilities of the outcomes of an uncertain event, one should select a probability distribution with large variance, and consequently large Shannon's entropy. This is consistent with the practice of using the maximum entropy principle (Kapur and Kevasan, 1992) to model uncertainty.

Unlike the game and investing examples, it is difficult to carry certification tests for engineering design decisions to probe models of uncertainty, because products are usually designed for low probabilities of failure, and many thousands of tests may be required to reveal weaknesses. Occasionally, disastrous failures reveal inadequacy of probability of failure estimates, as happened with the space shuttle. Instead of waiting for disasters, we can also use ingenuity to test methods for making decisions under uncertainty. This involves inventing decision problems for data already available in existing databases.

Gigerenzer and Todd (2000, pp. 97-118) pioneered this approach, pitting a complex decision-making method against a simpler, heuristic one. If the simpler method wins or draws, it reveals possible weakness in the more complex method. They used 20 existing available databases to compare methods for making binary decisions (e.g., find which of two professors has a higher salary, given cues such as each professor's rank and gender). They found that a heuristic method that takes into account only a single dominant cue bested the standard (and more complex) regression approach that takes all the cues into account.

We generalize Gigerenzer's testing procedure to compare methods for making decisions under uncertainty that require choice of optimum values of design (decision) variables. We have two objectives. First, we want to demonstrate that it is easy to take a database and invent scenarios calling for a decision (in short, decision scenarios) that lead to meaningful tests of the effectiveness of decision-making methods. Our testing procedure allows us to study the effect of modeling error because it uses real life data. Second, we wish to demonstrate with a simple example that such tests can raise concerns about aspects of methods that may not be readily apparent by examining the theoretical foundations of the methods.

As an example, we use a database (Table 1) of experiments in which one of us (Rosca), as well as a group of students engaged in a competition (Rosca, 2001), stacked domino blocks until they toppled (Fig. 1). We invent a decision scenario for a decision maker to guarantee a height for a domino tower that she will build so as to best a competitor by selecting a guaranteed height that is both attainable and competitive (it is unlikely that Competitor's tower will be taller by a given margin). This scenario is similar to a class of decision problems where a decision maker guarantees a performance level, and wins if she delivers it and the competitor fails to do so. This example allows us to compare the use of probability and possibility for making decisions.

Section 2 presents the approach for probing such methods for design under uncertainty using existing data. Sections 3-4 present the example with the domino towers, the results, and the lessons learned. Section 5 summarizes the conclusions of the study.

2. Testing approach

Figure 2 is an influence diagram of a decision with imperfect information. Elements of the decision are a *decision maker(s)*, *alternative courses of action* (in short*, actions*), *uncertain event(s)* and their *outcomes*, *consequences* of actions and *information* about the likelihood of the outcomes of the uncertain events. The consequence of an action depends on the outcomes of the uncertain events. The decision maker wants to select the action with the most desirable consequence. The decision maker has imperfect information about the likelihood (e.g., the probabilities) of the possible outcomes of the uncertain events. In this paper, we consider uncertain events whose outcomes are characterized by variables (e.g., the uncertain collapse height of a tower of domino blocks is a variable).

Our approach for testing methods for decision making under uncertainty is to select a database with samples of some variables, construct a decision scenario in which these variables represent the uncertainties that the decision maker faces, make repeated decisions and evaluate the consequences of these decisions using the database. The four steps of the testing approach are explained in detail in the following (Fig.3). Note that the variables in the data base do not have to be random; rather, the decision maker is uncertain about the values that they assume.

Table 1. Domino competition database: maximum built height (in domino units)

Height	Number of towers of given height		Height	Number of towers of given height		Height	Number of towers of given height	
	Ro-sca	Competi-tors		Ro-sca	Competi-tors		Ro-sca	Competi-tors
20	1	0	**32**	3	7	**44**	0	1
21	1	0	**33**	4	7	**45**	2	2
22	0	1	**34**	4	4	**46**	2	4
23	2	0	**35**	1	7	**47**	0	3
24	0	0	**36**	3	3	**48**	0	0
25	2	1	**37**	5	9	**49**	0	0
26	1	0	**38**	1	2	**50**	0	0
27	1	9	**39**	2	3	**51**	0	0
28	3	2	**40**	2	1	**52**	0	0
29	3	4	**41**	1	3	**53**	0	0
30	3	6	**42**	0	2	**54**	0	0
31	3	5	**43**	0	3	**55**	0	1

Step A. Select a database. We can start with almost any database with samples of a reasonable size (e.g., greater than or equal to 30).

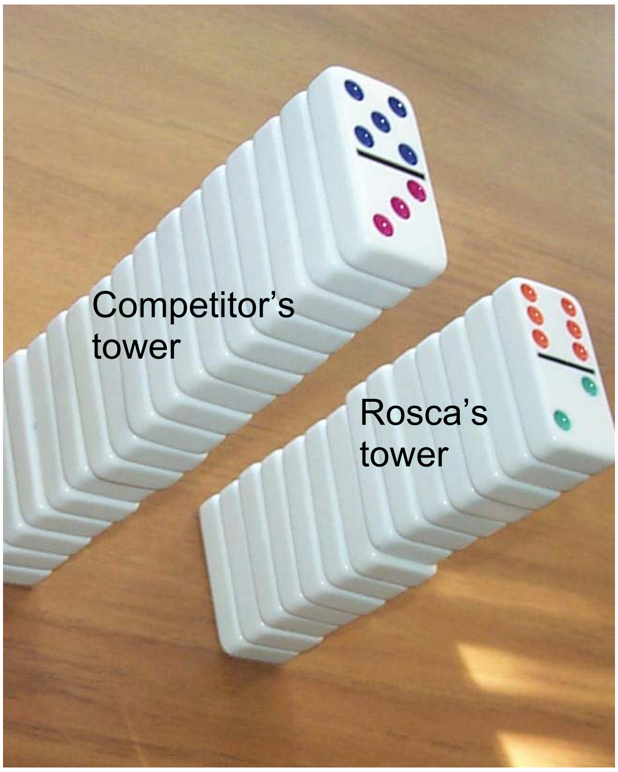

Figure 1: Domino towers in a competition

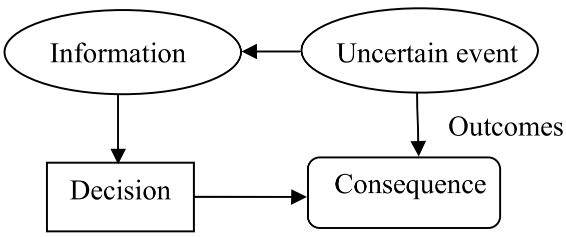

Figure 2. Decision under uncertainty. Arrows show relationships between the elements of a decision.

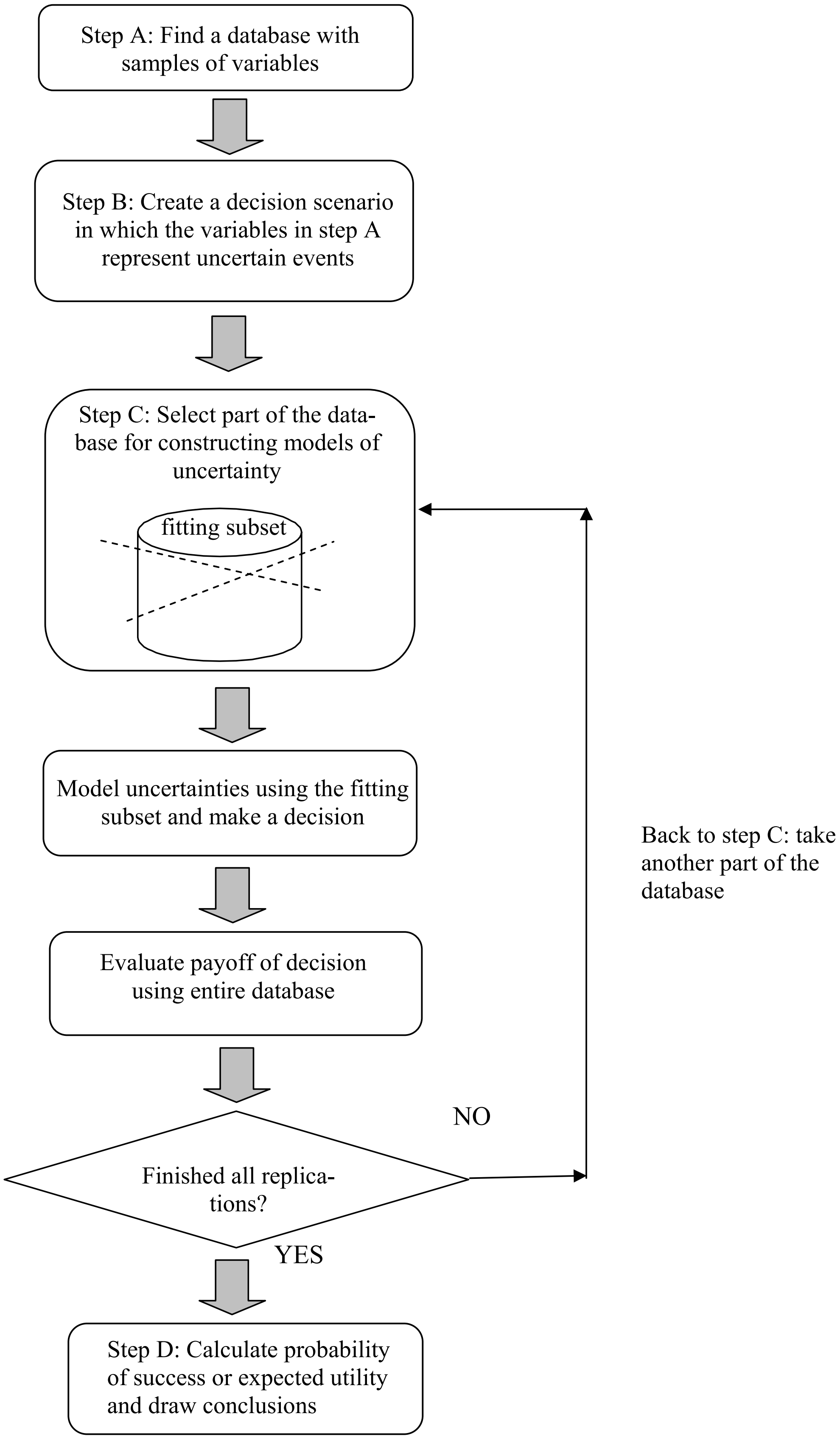

Figure 3: Approach for testing a method for decision under uncertainty

In the example of Table 1, the database contains heights of domino towers (just before they toppled) built by one of the authors (Rosca) in 50 trials and by 16 competitors in 90 trials (Fig. 1). The maximum height of a stable tower built by stacking domino blocks until the tower topples will be called *maximum built height* in this paper. This data gives us some statistical information (summarized in the histograms of Figure 4) on the height of the domino towers that Rosca and the Competitor can build.

Step B. Create a decision scenario given the variables in the database selected in step A. This is an unusual way to construct a decision scenario; instead of identifying the uncertainties in a given decision scenario, we invent a decision in which the variables in the database represent the uncertainties. A decision scenario is defined in terms of the following:

- the decision maker(s)
- the decision maker's objective
- the alternative courses of action (or choices)
- the possible consequences of an action
- the variables that affect the consequences of an action
- an algorithm for determining the consequence of an action given the values of the variables
- the information available for modeling the uncertainty associated with the variables.

For the domino-tower competition, the decision scenario we created is for Rosca (decision maker) to compete with a randomly chosen competitor (called Competitor) and guarantee a minimum height that she would build. Rosca loses if she did not meet her guarantee even if her tower was taller than that of Competitor's. To compensate for this disadvantage of Rosca, the rules of the competition stipulate that Competitor wins only if his tower height exceeds that of Rosca's guarantee plus a handicap. Both Rosca and Competitor are to build towers until they topple, and the maximum built height counts. If Rosca makes a high guarantee she risks not meeting it. If she makes a low guarantee she risks Competitor beating her guarantee plus the handicap. Figure 5 shows the decision tree, while Table 2 shows the elements of the decision scenario.

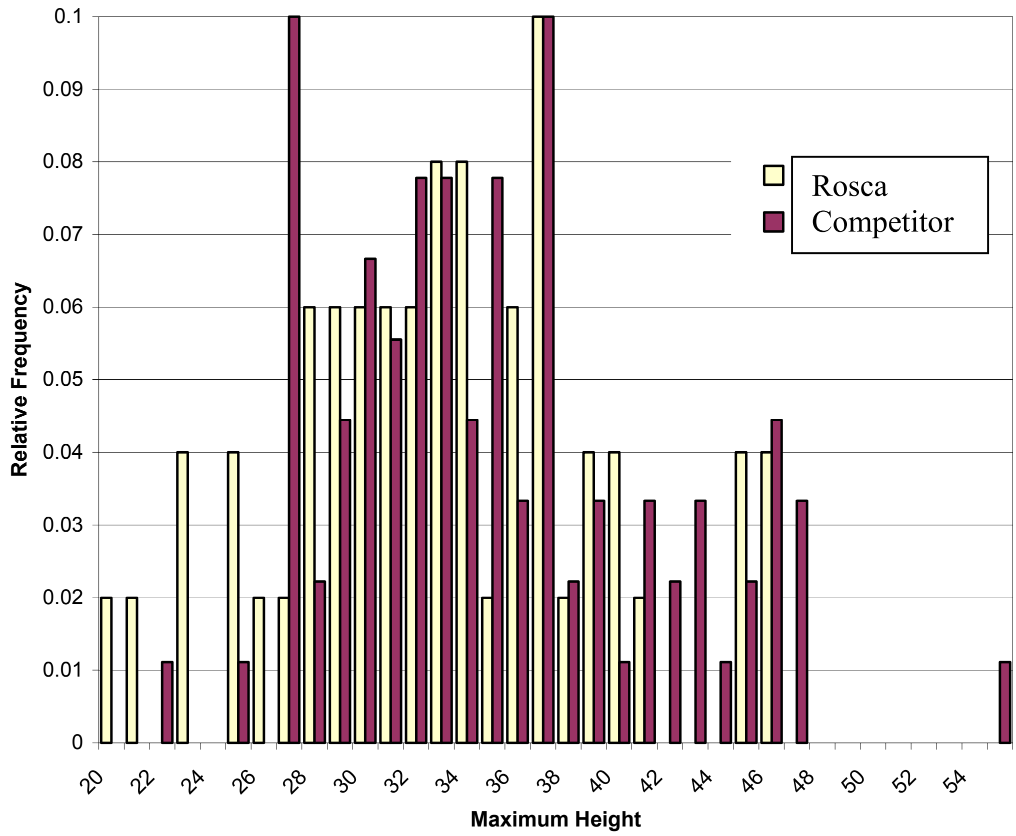

Figure 4. Histograms of maximum built heights of domino towers

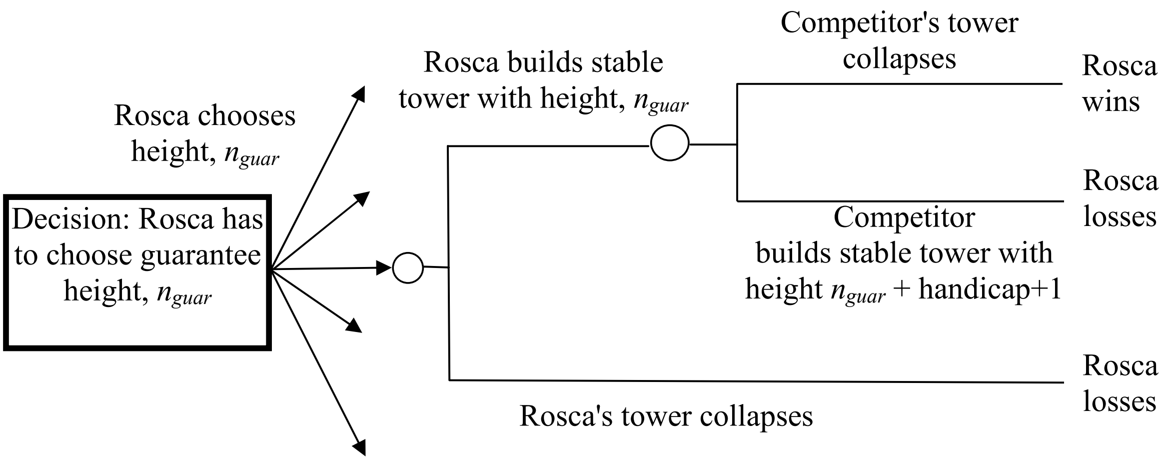

Figure 5: Domino competition: decision/event tree

Table 2. Elements of decision scenario in example

Element	Description
Decision maker	Rosca
Objective	Win contest
Alternative courses of action	Guarantee different tower heights, n_{guar}
Possible consequences of an action	Rosca loses or wins
Variables that affect the consequences of an action	Maximum built heights of Rosca's tower (n_{del}) and Competitor's tower (n_{comp})
Algorithm for determining the consequence of an action given the values of the variables in the database	3 Rosca's tower collapses below guarantee, ($n_{del} < n_{guar}$) (Rosca loses) 3 Rosca builds a stable tower with guaranteed height, ($n_{del} \geq n_{guar}$) and Competitor builds stable tower with height greater than the guaranteed height plus the handicap, $n_{comp} > n_{guar} + n_{hand}$ (Rosca loses) 3 Otherwise Rosca wins
Information for modeling uncertainty	Data on maximum built heights of Rosca's and Competitor's towers in database

Step C. Make decisions using the method(s) we want to test using part of the database. We select a part of the database (called fitting dataset) to construct models of the uncertainties, which are used to make a decision. This adds statistical uncertainty (uncertainty in estimating the statistics of the population in the database from the fitting dataset) to the uncertainty due to variability. It is important to investigate the effect of statistical uncertainty since it is usually present in design decisions.

Using a part of the database to construct models of the variables allows us to test the method on many decisions; each obtained using a different fitting part. We reduce the element of chance in the choice of the fitting part by selecting it randomly, and repeating the process many times. Thus we obtain a large number of decisions whose payoffs can be evaluated. This concept of testing a model using multiple random fitting datasets is commonly used in validating response surface approximations, such as neural networks (e.g., Hush and Horne 1993).

For the domino-tower problem, we employ probabilistic and possibilistic methods to decide what height to guarantee. We provide Rosca with a small random sample (the size is five for the results presented in this paper) of her own past performance as well as a similar sample of Competitor's past performance. We could select $\binom{50}{5}$ different parts of the database with values of the collapse heights of Rosca's towers, where the notation $\binom{50}{5}$ indicates the number of all different 5-tuples taken from a population of 50 objects.

Step D. Evaluate the payoff of a decision by using the database as the entire universe of possible outcomes.

For a binary consequence (success or failure), we measure the payoff of a decision by its probability of success evaluated from the entire universe of all possible outcomes. If success is a matter of degree, then we use the expected utility (Marston and Mistree, 1998, Hazelrigg, 1997, chapter 7) instead.

In the domino-tower problem we have 90×50=4,500 possible combinations of maximum built heights of Rosca and Competitor, and we can readily calculate the ratio of successful decisions out of the total.

Student grade data base

To demonstrate the generality of the testing approach, we present a database with very different characteristics from the domino database. As faculty members we regularly create student grade files, such as the one shown in the table below.

student	quiz1	quiz-2	quiz-3	Exam-1	quiz-4	…	Course average	Grade
1	30	26	0	73	22	…	80.91	B
2	18	25	28	99	11	…	95.05	A
:	:	:	:	:	:	:	:	:
44	25	31	30	62.5	24	…	48.53	F
45	23	21	10	68	11	…	86.13	B+

Using the student-grade database we can create the following decision scenario. A professor wants to identify students who are likely to get D or below (considered failure here) in order to call them for consultation. It is desirable to make the decision process simple and transparent. Therefore, the process is that if a student's course average is below a cutoff value, a_c, at the end of week T of the semester, the student will be called. To aid the professor identify which students to call for consultation, a teaching assistant (decision maker) wants to develop a model predicting if a student, whose course average at the end of week T is known, will fail. The construction of a predictive model can be viewed as a decision in which the teaching assistant decides on the consultation time and cutoff grade (decision variables). The teaching assistant's objectives are to maximize the model accuracy and minimize the waiting time to issue a warning.

The consequences of a choice of the consultation time and the cutoff grade include the number P of false positives (students called for consultation who would have passed), and the number N of false negatives (students who failed but were not called for consultation). It is desirable to minimize P and N, which would call for waiting as long as possible. On the other hand, the longer the professor waits, the smaller is the chance that the student can improve much. We therefore define the following loss function to be minimized by the decision maker:

$$L = k_P P + k_N N + k_T T$$

The coefficient k_P is the weight assigned to wasted time with students who do not need the consultation, k_N is the weight assigned to missing students who need it, and k_T is the weight assigned to the loss of time available to the student for corrective action.

In this problem the decision maker is uncertain if a given student will pass or fail. The grades in the database provide information that the teaching assistant can use to model this uncertainty. Figure 6 explains the decision of the teaching assistant. Using the information in the database the teaching assistant builds a model of the random variables. Then he/she chooses the waiting time and the cutoff grade so as to minimize the loss function of the model for predicting if a student will fail.

This problem may be particularly useful for testing Bayesian approaches against more traditional probabilistic approaches, because data from other courses and previous experience with the same course may be used to create prior probability distributions of failure as functions of the two decision variables. We can update the prior probability distributions using the grades in the database.

The testing procedure involves providing information on the entire semester history for a few students, using it to select the two decision variables, and then testing the consequences on the remaining students. That is, once the week T and cutoff a_c have been chosen, we can use the final grade information to obtain the number of false positive P and number of false negative N, and calculate the loss function. The procedure can be repeated for many random subsets of the students.

The loss function implicitly assumes that a consultation with the student will increase the chances of the student to pass. While this is not obvious, we note that this issue does not lessen the value of this example for testing methods for making decisions, as the loss function involves diagnosis rather than corrective effects.

3. Comparing Possibilistic and Probabilistic Formulations for Domino Problem

A common option for modeling the uncertainty in the maximum built height in the domino decision problem of Table 2 is to fit probability distributions to data on past performance (data from Table 1 or Figure 4). Our previous investigation into the mechanics of the domino problem revealed that the probability distribution of stack heights for a single builder or for a group of builders can be approximated well by a shifted Gamma distribution but is approximated almost as well by a normal distribution (Rosca, 2001).

In order to demonstrate the utility of our testing approach we compare a probabilistic and a possibilistic method for making the decision. The latter is based on a simpler representation of the uncertainty via a triangular possibility distribution function and may be therefore less sensitive to the lack of available data. Possibility theory is presented in several books and papers including Dubois and Prade (1988), Joslyn (1994, 1995), and Nikolaidis et al. (2004).

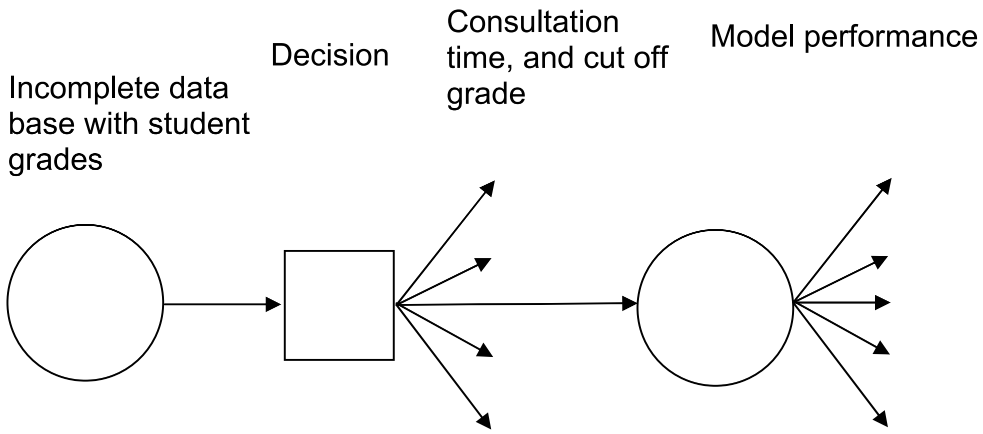

Figure 6. Decision about consultation time and cutoff grade when constructing a model for predicting a student's performance

Problem formulation
In possibility theory, the possibility of an event and the possibility of its complement do not necessarily add up to 1 (as is the case for probability theory). Therefore, we can maximize the possibility of success or minimize the possibility of failure without necessarily obtaining the same guarantee. We assume that the maximum built heights of Rosca's and Competitor's towers are independent. Then the possibility of Rosca winning is equal to the minimum of the possibility of Rosca building a stable tower (*Rosca delivers*) with guaranteed height *n* and the possibility of Competitor failing to build a tower taller than the guaranteed height plus the handicap (*Competitor fails*):

$$Pos\ (winning\ (n_{guar})) = Pos\ (\text{Rosca delivers and Competitor fails}) =$$

$$min\ [Pos\ (n_{del} \geq n_{guar}),\ Pos\ (n_{comp} < n_{guar} + n_{hand} + 1\ )] \qquad (1a)$$

Similarly, the possibility of Rosca losing is:

$$Pos\ (losing\ (n_{guar})) = max\ [Pos\ (n_{del} < n_{guar}),\ Pos\ (n_{com} \geq n_{guar} + n_{hand} + 1)]. \qquad (1b)$$

Both formulations can provide multiple optima. For our data, the sets of optima given by these two possibilistic approaches were not disjoint. We call the intersection of these two sets the *possibilistic optimum*. There might be cases where the intersection contains more than one element.

Since Rosca built her towers without interaction with the other builders and at a different time, Rosca's and Competitor's maximum built heights are assumed statistically independent. Therefore, the probability of Rosca winning a contest when she guaranteed a stack of height $n=n_{guar}$ is equal to the product of probability that Rosca delivers and the Competitor fails:

$$Pro\ (winning\ (n_{guar})) = Pro(n_{del} \geq n_{guar}) \cdot Pro(n_{comp} < n_{guar} + n_{hand} + 1) = [\ 1 - F_{Rosca}(n_{guar})]\ F_{Comp}(n_{guar} + n_{hand} + 1) \qquad (2)$$

where $F_{Rosca}(n)$ and $F_{Comp}\ (n)$ denote the cumulative distribution functions of the maximum built heights of Rosca's and Competitor's towers, respectively. The probability that Rosca delivers decreases with n_{guar}, whereas the probability of Competitor's failure increases with n_{guar}. In the probabilistic formulation, we want to find the guaranteed stack height, n_{guar}, that maximizes the probability of winning.

We will compare the optima obtained by the two formulations when using data from the domino experiments to model uncertainty. We analyze two cases: (1) all data are used to find the optimum, and (2) only a sample of the data is used. When little data is available a designer can use a standard probability distribution that it is known to describe the uncertain variable (in our case the maximum built height) or employ the maximum entropy principle if such a standard distribution is not known. In this study, we consider that both the probabilistic and possibilistic designers know that the Gamma and the normal distributions fit well the maximum built height of a tower. In case (1), there is only uncertainty due to the error in approximating the actual discrete distribution as Gamma or normal probability distributions (called here fitting error). The true values of the parameters of these distributions are computed from the entire database of the maximum built heights, which in this study is considered the universe of the values of the maximum built heights. In case (2), there is additional uncertainty in the values of the parameters of the distributions besides the uncertainty due to fitting error.

In the decision scenario considered in this study, the probabilistic approach has two advantages compared to the possibilistic approach: a) the probabilistic approach seeks to maximize the right objective (the probability of Rosca winning), and b) even when a small sample of data is available, Rosca knows that the Gamma and the normal probability distributions fit well the data for the maximum built heights. The information on the type of probability distribution is one that the possibilistic designer cannot directly utilize, so the probabilistic designer has an advantage.

Definition and evaluation of the likelihood of winning

Generally, for a given sample, the possibilistic and probabilistic formulations yield different optima, because they maximize different objective functions. We compare the two optima in terms of their relative frequency of winning (also called likelihood of winning), considering all possible Rosca-Competitor competitions obtained by combining all the data for the collapse heights of the towers built by Rosca and Competitor. With 50 experiments available for Rosca and 90 experiments available for Competitor, the likelihood is calculated by counting the number of pairs for which Rosca won as a fraction of the universe of possible pairs of Rosca and Competitor data, that is, 4,500 pairs. Consider a competition in which the maximum built height of Rosca is $n_{del}=N_1$ blocks and the maximum built height of the competitor's tower is $n_{comp}=N_2$ blocks. Rosca won if

$$N_1 \geq n_{guar} \text{ and } N_2 < n_{guar} + n_{hand} + 1.$$

The likelihood of winning of n_{guar} is the total number of pairs (N_1, N_2) for which Rosca won, divided by 4,500. This likelihood of winning may be viewed as an approximation to the actual probability based on the limited database. We prefer to view it as an exact calculation for a problem with a limited discrete universe.

Using the likelihood of winning as a metric of the quality of a decision and with all of the data and no fitting errors, the probabilistic formulation should be superior. The possibilistic approach can prevail only if the fitting errors and the errors due to incomplete data overcome the natural advantage of the probabilistic approach.

Splitting the data into fitting and testing sets

If we use all the data for selecting the optimal n_{guar}, we have a single example from which it is difficult to draw conclusions. However, the relatively large amount of data allows us to use subsets for making the decision and evaluating the payoff, and then to repeat the process for different subsets. This reduces the element of chance in the results. Here, we perform the comparison for 80 randomly chosen subsets.

We draw samples of size n_{sample} from both the data sets of Rosca and Competitor. Based on these samples, we fit a shifted Gamma or a normal probability density and a possibility distribution. The fitting processes for the probability and possibility distributions of the collapse heights are described in Appendix 1. Based on the fitted functions and using a probabilistic or a possibilistic formulation, we solve the guaranteed height problem, obtaining one (or more) optimum guarantees.

Step D of our testing approach calls for evaluating the payoffs of the decisions. We compare the guarantees selected by each method in terms of their likelihood of Rosca winning on all possible combinations of the available data.

4. Results

We studied the likelihood of winning of the two methods first when all measurements in the database are known, and then when only five measurements are given. We also investigated the effectiveness of the common practice of inflating the variance of a variable to account for statistical uncertainty (uncertainty in estimating the statistics of a population from those of a sample). This second study revealed an unexpected difference between probability- and possibility-based designs.

All data known

In this case, we do not have multiple samples, and we can make a single decision. However, we vary the handicap through the set of values {2, 5, 8, 11, 15}. Figure 7 shows the likelihood of Rosca winning for the probabilistic (with Gamma distribution) and possibilistic designs versus the handicap. As expected, the likelihood of winning increases with the handicap. One cannot tell which method does better from Fig. 7, as the probabilistic design wins for a handicap of 2, 5 and 15, while the possibilistic design wins for a handicap of 8 or 11. Figure 7 also shows the maximum achievable likelihood of winning in the ideal case where the probability distributions of the populations of the maximum built heights of the two players are known. These are the true probability distributions of the maximum built heights and they are equal to those in the histograms in Fig. 4. The difference between the maximum achievable probability of winning and the likelihood of winning of the probabilistic approach is due to the fitting error of the Gamma distribution to the data. It is observed that the effect of the fitting error is small.

Table 3 shows the optimum guarantee selected by the two probabilistic models and the possibilistic approach. The optimum guarantee decreases with the handicap increasing. This is because the increased handicap makes it harder for Competitor to build a tall enough stable tower and a low guarantee will reduce the risk of Rosca's failure to deliver the guarantee. The average likelihoods of winning of the three methods over the five values of the handicap are: 0.5289 for the probabilistic design method using the Gamma distribution, 0.5258 for the probabilistic design method using the normal distribution and 0.5234 for the possibilistic design approach, which are very close. These results may indicate that when all the data is available to the decision maker, the errors incurred by fitting the data to a probability distribution offset the advantage of the probabilistic approach over the possibilistic one (that it maximizes the same objective as the one used to score the results).

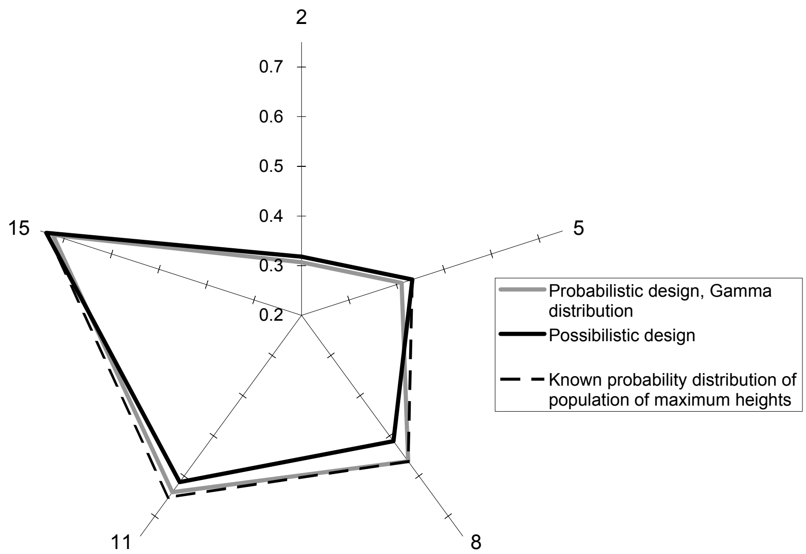

Figure 7. Comparison of the likelihood of success of probabilistic and possibilistic designs versus the handicap for the case where the decision maker has all the data in the database and the case where the decision maker knows the true probability distribution of the population of maximum heights

Table 3: Variation of optimum guarantee and its likelihood of winning with handicap values when all data are known; cases where the optimum guarantee was found are marked in bold. The optimum guarantee for a handicap of 11 is 28, corresponding to a likelihood of winning of 0.6533.

Handicap	Probabilistic optimum (shifted Gamma fit)		Probabilistic optimum (normal fit)		Possibilistic optimum (triangular fit)	
	Optimum	Likelihood of winning	Optimum	Likelihood of winning	Optimum	Likelihood of winning
2	32	0.3067	**33**	0.3180	**33**	0.3180
5	31	0.4107	**32**	0.4333	**32**	0.4333
8	**29**	0.5633	30	0.5360	31	0.5133
11	27	0.6402	29	0.6153	29	0.6153
15	26	0.7236	27	0.7262	**28**	0.7373

When only few experimental data are available to fit a probability distribution, a standard practice (Fox and Safie, 1992) is to inflate the variance of a distribution, keeping the mean value the same. Considering the parameters of a probability distribution to be random variables in order to account for statistical uncertainty also increases the variance of the distribution. We inflate the variance by adding to it an inflation factor multiplied by the standard deviation of the variance (see Appendix 2). When all the data is known, the effect of inflation is small because the standard deviation of the variance is small (see for example Table A1). Therefore, in order to understand the effect of inflation, we consider also the extreme case of an inflation factor of 15.

For a possibility distribution, there is no standard way to inflate the uncertainty. We use the simple approach of keeping fixed the mode of the distribution, which is equal to the sample mean, and inflating the support by the inflation factor. That is, if the mean is 32, and the support of the possibility distribution function is the interval (30, 35), then an inflation factor of 1 will inflate the interval to (28, 38), and an inflation factor of 2 to (26, 41). Here we use an inflation factor of 2, which corresponds to extreme inflation, similar in magnitude to an inflation factor of 15 for the probabilistic data.

From Table 4 we observe that when the uncertainty in Competitor's performance increases (inflation of 15), the probabilistic optimum guarantee decreases. On the other hand, when the uncertainty in Rosca's performance increases, the probabilistic optimum guarantee increases. The possibilistic optimum guarantee exhibits the opposite trend.

Table 4: Effect of inflating the uncertainty Rosca's and the Competitor's performance on the optimum guarantee and its likelihood of winning; handicap value, n_{hand} *is 5*, all-data case. The true optimum height is 32. The probabilistic optimum decreases when the variability in Competitor's performance increases, and increases when the variability in Rosca's performance increases; the possibilistic optimum exhibits the opposite trend.

Rosca	Competitor	Probabilistic optimum (shifted Gamma fit)		Probabilistic optimum (normal fit)		Possibilistic optimum (triangular fit)	
Inflation factor		Optimum	Likelihood of winning	Optimum	Likelihood of winning	Optimum	Likelihood of winning
0	0	31	0.4107	32	0.4333	32	0.4333
0	15	28	0.3920	29	0.3987	33	0.4020
15	0	33	0.4020	34	0.3578	31	0.4107
15	15	30	0.4240	32	0.4333	32	0.4333

When the variance of the Competitor's performance is inflated by a very large amount then the probabilistic optimum guaranteed height always decreases. The reason is that when the variance becomes very large, the probability of the Competitors' failure becomes insensitive to the guaranteed height. Therefore, for increasing the probability of winning given by (Eq. 2) *Pro (winning (n)) = [1-*F_{Rosca}*(n)]* F_{Comp}*(n+*n_{hand}*+1)* it is more important to increase the probability of Rosca's delivering the guarantee than to increase the probability of Competitor's failure.

The effect of inflating uncertainty on the optimum can be understood by examining the condition that the optimum must satisfy. At the optimum, the derivative of the logarithm of the probability of success is zero,

$$\frac{\partial \log(P(winning))}{\partial n} = \frac{-f_{Rosca}(n)}{(1-F_{Rosca}(n))} + \frac{f_{Comp}(n+n_{hand}+1)}{F_{Comp}(n+n_{hand}+1)} = 0 \qquad (3)$$

The two terms on the right hand side of the above equation are the sensitivities of the logarithms of the probabilities that Rosca delivers and Competitor fails. Extreme inflation of the uncertainty in the Competitor's performance makes the sensitivity of the derivative of the probability that the Competitor fails almost zero. The optimum guaranteed height decreases in order to maintain equality of the two terms in Eq. (3) (Fig. 8). Rosca (2001) provided a similar explanation as to why the optimum guarantee increases when the probability density function of the decision maker (Rosca) is inflated.

So in this probabilistic guarantee-setting problem, when the decision maker is highly uncertain about the capability of the competition, she should set conservative goals. On the other hand, when the decision maker is very uncertain in her own capability, she should set aggressive goals because it is more important to prevent the Competition from succeeding than to help the decision maker deliver the guaranteed performance. This fits the common sense notion that given two dangers, one should pay more attention to the danger which one can manage more easily.

In possibility, we can minimize the possibility of Rosca losing the contest or maximize the possibility of her winning. We minimize the possibility of losing because the possibility of winning is equal to one for heights between 30 and 33. The height for which the possibilities that Rosca delivers and the Competitor fails become equal (n_{opt} in Fig. 9), minimizes the possibility of losing. Indeed, any deviation from n_{opt} increases the possibility of Rosca losing. Smaller heights than n_{opt} have higher possibility of Competitor success, while larger heights have higher possibility of Rosca failure. In both cases, the possibility of Rosca losing the contest (Eq. 1b) is higher than that for n_{opt}. In Table 4, the possibilistic optimum displays the opposite trend than the probabilistic optimum, increasing when we inflate Competitor's possibility distribution and decreasing when we inflate Rosca's possibility distribution. This can be explained by observing Fig. 9; inflating the Competitor's possibility distribution will increase the optimum (which is the intersection of the possibility distributions of the two players). Thus, in contrast to probabilistic design, inflation increases the importance of a failure mode in the possibilistic approach.

The philosophies of the probability and possibility can be further understood by examining a scenario that accentuates the difference of the optima of the two approaches. Consider the extreme case where the uncertainty in the Competitor's performance is very large and the uncertainty in Rosca's performance very small (that is, Rosca predicts quite accurately the maximum height of a tower that she can build but she does not know much about Competitor's performance). Figure 10 shows the probability densities and possibility distributions of the maximum built heights of the two players. The optimum guarantee that maximizes the probability of success is on the left tail of Rosca's probability density function where there is a small probability that she will not deliver. Then, the probability of winning is approximately equal to the probability of Competitor's failure for this height, which is 0.5. On the other hand, the possibilistic optimum is very close to the mean value of the Rosca's distribution and has a probability of success 0.256, which is approximately equal to one half of the probability of success of the probabilistic optimum. It is also interesting that the possibilistic optimum is less robust to errors in the mean value of Rosca's maximum built height than the probabilistic optimum. For example, even a small reduction in the true mean value of Rosca's probability distribution will reduce greatly the probability of success of the possibilistic optimum. *Even though we have been comparing probability and possibility for the past few years, we needed this experimental result to discern the important differences between probability and possibility identified in this study.*

It is not difficult to check that the effect of inflation of the possibilistic optimum depends on the relative positions of the peaks for the two possibility distributions. That is, when the two are reversed, probability and possibility will behave in the same way. However, for probability distributions, the relative positions of the peaks do not affect the result that the inflated mode loses importance.

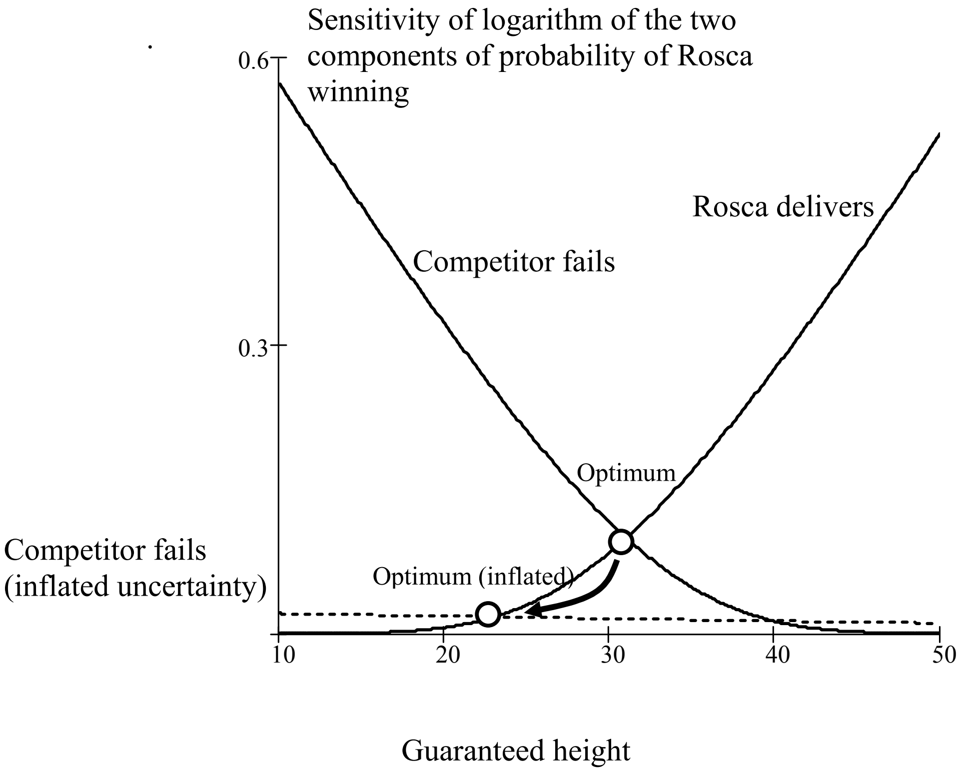

Figure 8. Extreme inflation of the uncertainty in Competitor's performance reduces the sensitivity of the probability of Competitor's failure to the guaranteed height, thereby reducing the optimum guaranteed height.

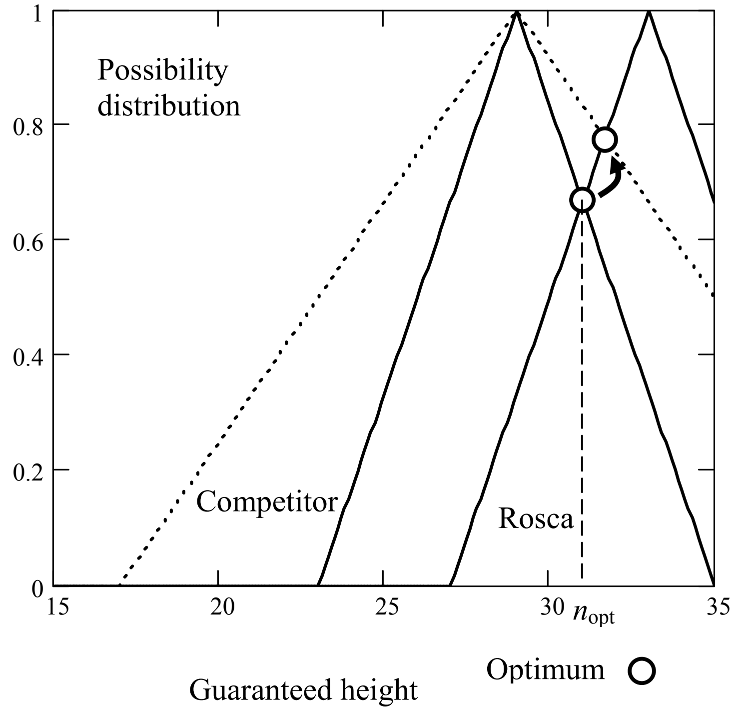

Figure 9. Inflation of the uncertainty in Competitor's performance increases importance of competitor's failure, thereby increasing the optimum guaranteed height.

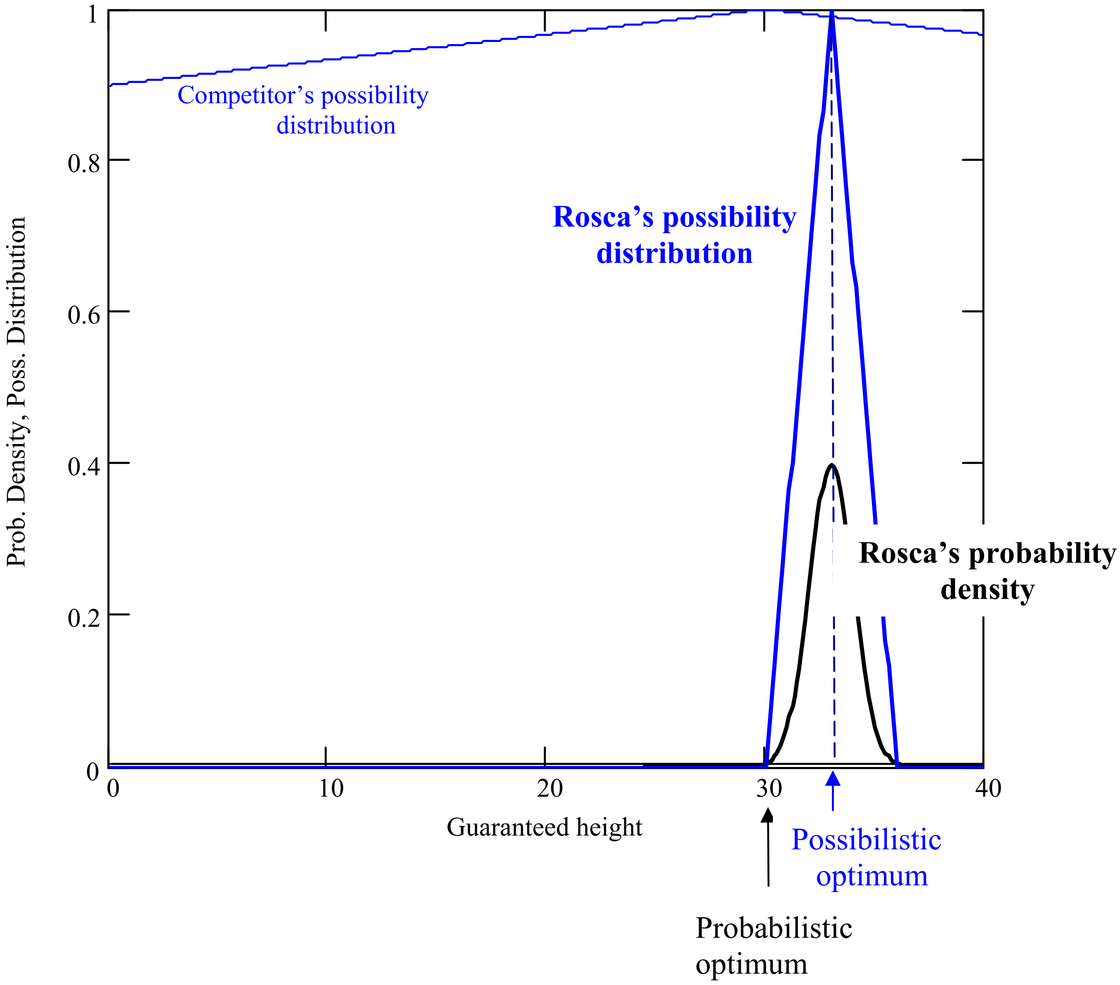

Figure 10. Probabilistic and possibilistic optima when there is high uncertainty in Competitor's performance and little uncertainty in Rosca's performance. The difference between the ways probabilistic and possibilistic design find the optimum guaranteed height is accentuated in this case. The probability density of Competitor's maximum built height is almost zero over the entire range of heights in the figure.

The contrast between the effect of inflation on probability and possibility is clearly due to the non-additivity of possibilities. Consider in Fig. 8 the optimum guarantee when the probability distributions of the players are not inflated. Reducing the optimum guarantee by 1 increases the chance that Rosca delivers more than reducing the chance that the Competitor fails. In probability, inflating the probability distribution of Competitor automatically reduces the probability of any one outcome. Thus, it allows us to reduce the guarantee by 1 with a smaller reduction of the chance of Competitor's failure. In the possibility model we can increase the possibility of one event without reducing the possibility of another. Thus inflating Competitor's possibility distribution increases the possibility of all the outcomes.

Finally, Table 4 shows that inflating only one of the uncertain variables reduces the likelihood of success of both methods. The reason is that increasing the uncertainty in the performance of a player introduces bias in the probabilistic model, which reduces the quality of a decision. Also, inflation affects probabilistic design more that the possibilistic design.

Scarce data – small sample size

For the scarce data case, we use only a randomly selected small subset of the data for fitting a distribution and selecting a guarantee. The process is repeated 80 times to average out the effect of chance in the selection of the sample. Rather than presenting all 80 examples of optima, we present their average (over the 80 samples) likelihood of success. We tested the models of uncertainty for five handicap values. Thus, we were able to test the models on 5×80 = 400 different decisions using the same pair of datasets for Rosca's and Competitor's collapse heights. Each decision was evaluated using 4,500 pairs of maximum built heights.

Figure 11 shows the average likelihoods of success of the probabilistic method that uses the Gamma distribution to model uncertainty in the maximum built heights of the towers and the possibilistic design for different handicaps. Since only five data points are available to the decision maker instead of 50 or 90, the likelihood of success of the optimum guarantee deteriorates compared to the all-data case. The likelihood of success of the probabilistic design is slightly higher than that of the possibilistic design but the difference is small. Table 5 presents the average and standard deviation of the likelihood of success when a sample of five values is used. When a Gamma probability distribution is fitted to the data, the reduction of the likelihood of success ranges from 2% to 6%, compared to the all-data case. The reduction in the likelihood of success ranges between 2% to 4% when a normal distribution is fitted to the data. Finally, when a possibilistic approach is used, the reduction in the likelihood of success ranges between 2% and 4%.

For both possibilistic and probabilistic methods, increasing the handicap value increases the mean of the likelihood of success. The average likelihoods of winning of the three methods over the 400 cases are 0.4947 for the probabilistic design method using the Gamma distribution, 0.4968 for the probabilistic design method using the normal distribution and 0.4909 for the possibilistic design method, which are very close, with a small advantage to the probabilistic models. This result surprised us, because, generally, we expected the possibilistic approach to do better relative to the probabilistic approach for the scarce data than for the full data case. But in the decision problem considered, probabilistic design has the advantage over the possibilistic design that the type of the probability distribution of the maximum built height is known even in the scarce data-case. Possibility does not permit the designer to account directly for this information even if she knows the type of the possibility distribution.

The poorer results of the possibilistic approach could also be due to the way we constructed a possibility distribution function based on the available data or the inability of the approach to properly account for the independence of the built heights of the towers of the two players. The possibility of the intersection of two events is equal to the minimum of the possibilities of these events (Eq. 1). This yields counterintuitive results when the events are known to be statistically independent. For example, the possibility of Rosca building a tower with height that has high possibility (e.g., 30) and the competitor building a very tall tower (e.g. 55) is equal to the possibility of both players building very tall towers as long as Rosca's tower has higher possibility than Competitor's tower. This is clearly wrong because it is very unlikely that both players will build high towers simultaneously (Eq. 2). A hybrid probabilistic/possibilistic approach, that characterizes uncertainty in the maximum built height using probability distributions and uncertainty in the distribution parameters using a possibility distribution would avoid the above pitfalls and could do better than the approach that we considered in this study.

Table 5: Mean and standard deviation (computed over the 80 cases) of the likelihood of success for probabilistic optimum (shifted Gamma and normal fit) and possibilistic optimum (triangular fit); sample size of 5.

sample size=5	Likelihood of success for probabilistic optimum (shifted Gamma fit)		Likelihood of success for probabilistic optimum (normal fit)		Likelihood of success for possibilistic optimum (triangular fit)	
n_{hand}	Mean (of 80 runs)	Standard deviation	Mean (of 80 runs)	Standard deviation	Mean (of 80 runs)	Standard deviation
2	0.2850	0.0361	0.2822	0.0398	0.2896	0.0290
5	0.3924	0.0441	0.3924	0.0451	0.3917	0.0478
8	0.4995	0.0552	0.5031	0.0496	0.4921	0.0576
11	0.5967	0.0622	0.5993	0.0513	0.5875	0.0656
15	0.6997	0.0559	0.7069	0.0412	0.6937	0.0586

We repeated the fitting and optimization procedure for the case of the sample size of 5, but this time we inflated the standard deviation of the maximum built height and the support of the possibility distribution of this height. We present in Table 6 only the results for symmetric inflation.

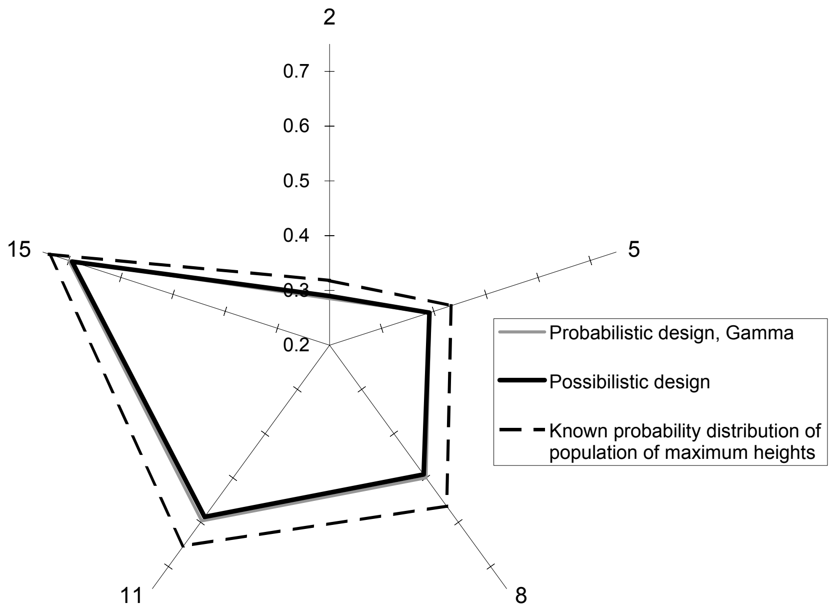

Figure 11. Comparison of the likelihood of success of probabilistic and possibilistic designs versus the handicap for the case where the decision maker has five data points and the case where the decision maker knows the true probability distribution of the population of maximum heights

Table 6: Mean and standard deviation (computed over the 80 cases) of the likelihood of success for probabilistic optimum (shifted Gamma and normal fit) and possibilistic optimum (triangular fit); sample size is 5. Both Rosca's and Competitor's inflation factors are 2.

sample size=5	Likelihood of success for probabilistic optimum (shifted Gamma fit)		Likelihood of success for probabilistic optimum (normal fit)		Likelihood of success for possibilistic optimum (triangular fit)	
n_{hand}	Mean (of 80 runs)	Standard deviation	Mean (of 80 runs)	Standard deviation	Mean (of 80 runs)	Standard deviation
2	0.2687	0.0520	0.2797	0.0466	0.2896	0.0290
5	0.3662	0.0591	0.3899	0.0496	0.3917	0.0478
8	0.4732	0.0758	0.4980	0.0554	0.4917	0.0575
11	0.5717	0.0849	0.6017	0.0481	0.5879	0.0645
15	0.6922	0.0637	0.7075	0.0360	0.6920	0.0606

Comparing Tables 5 and 6, we see that inflation had a detrimental effect on the probabilistic optimum. Indeed, for all but the handicap value of 2, the mean likelihood of success of the optimum given by the inflated shifted Gamma distribution is smaller than the corresponding non-inflated one. The same effect is observed for the normal distribution for all but the handicap of 11. For symmetric inflation, little or no effect is observed on the likelihood of success of the possibilistic optimum, because the possibilistic optimum does not change with symmetrical inflation. We also repeated the study with sample sizes of 3 and 10 and obtained similar results (Rosca, 2001).

The observation that inflation of uncertainty is counterproductive is at odds with Walley's observation in his World Cup experiment where those participants whose probabilities of the outcomes of a game were uniform made more money than those whose probabilities differed a lot. We think that Walley's experiment does not necessarily show that inflating uncertainty is an effective practice; it possibly shows that those participants who estimated uniform probabilities because they were aware of their ignorance did better than overconfident participants whose probabilities were asymmetric.

5. Conclusions

An approach for using existing data for probing weaknesses in models for making decisions under uncertainty has been developed. The approach may expose problems associated with errors in predictive models or in models of uncertainty because it uses real-life data. The approach requires two sets of data on one property (here, domino tower height) for two groups. It then creates a decision problem that involves finding an optimum in terms of one or more decision variables. The same dataset can be used to test methods on hundreds or thousands of different decisions within a short period at low cost. An example employing data on a domino tower competition was used for demonstration.

The utility of the experimental testing of methodologies for decisions under uncertainty was evidenced by several results that surprised us, even though we have been exploring the methods we evaluated for several years. These include the following:

1. Small fitting errors in the probability distributions were sufficient to give an advantage to possibilistic decision-making, even though the metric of success was probabilistic. This may indicate that these fitting errors deserve further study.
2. In contrast, the probabilistic approach suffered less than the possibilistic approach from small sample size. This may indicate that a better way of selecting possibility distribution

functions based on small samples may be needed, or that a hybrid probabilistic/possibilistic approach should be used in which possibility is only used only for those uncertainties for which it is difficult to estimate probabilistic models.

3. The process of magnification of the standard deviation of a probability distribution, which is commonly used when data are scarce, proved to be counterproductive.
4. The effect of magnifying uncertainty had an opposite effect on probability and possibility. Inflating uncertainty reduced the effect of a failure mode on the probabilistic decision, and it increased the effect of the mode on the possibilistic decision. This result was shown to be due to the fact that probability, unlike possibility, is additive. In the extreme case where uncertainty in one player's performance is much greater than in the other player's, the difference between the optimum decisions of the two methods is very large. In this case, possibility yielded a decision with much poorer performance than probability.

We note that there is a wealth of other data readily available for testing methods using the proposed approach, including records of student projects, insurance claims, stock market prices, and medical tests. As educators, we can readily see that universities have useful student databases. For example, records of the performance characteristics (e.g. the stroke and time ratio) of slider-crank mechanisms constructed in a class on design and analysis of mechanical systems can be used instead of domino heights.

References

Baer, G., and Gesnsler, G., *The great mutual fund trap*, Broadway Books, Chapter 10, pp. 148-154, (2002).

de Finetti, B., *Probability, induction and statistics,* Wiley, London Chapters 1 and 2, (1972),.

Dubois, D., and Prade, H., *Possibility theory*, Plenum Press, New York (1988).

Fox, E. P., Safie F., Statistical characterization of life drivers for a probabilistic analy, *AIAA/SAE/ASME/ASEE, 28th Joint Propulsion Conference and Exhibit, Nashville, Tennessee*, AIAA-92-3414, (1992).

Freund, J.E., and Williams, F.J., *Dictionary/outline of basic statistics*, McGraw-Hill, New York, pp. 151, (1966).

Gigerenzer, G., and Todd, P. M., *Simple heuristics that make us smart*, Oxford University Press, New York, (1999).

Greene, J. and Smart, S., Liquidity provision and noise trading: Evidence from the 'investment dartboard' column. *Journal of Finance,* **54,** October, (1999).

R.T. Haftka, E.P. Scott and J.R. Cruz, Optimization and experiments: A survey. *Applied Mechanics Reviews,* **51,** 7, pp. 435-448, (1998).

G.A. Hazelrigg, *Systems engineering: An approach to information-based design*, Prentice Hall (1996).

Hush, D. R. and Horne, B. G., Progress in supervised neural networks. *IEEE Signal Processing Magazine,* pp. 8-38, (1993).

Joslyn, C., Possibilistic processes for complex systems modeling, Ph.D. dissertation, *Department of Systems Science, SUNY Binghamton,* Binghamton, New York, 1994.

Joslyn, C., In support of an independent possibility theory, in: de Cooman, Ruan, Kerre,, editors, *Foundations and applications of possibility theory*, World Scientific, Singapore, pp. 152-164, (1995).

Kapur, J.N., and Kevasan, H.K., *Entropy optimization principles with applications*, Academic Press, New York, (1992).

Marston M, Mistree F., An implementation of expected utility theory in decision based design. Proceedings of the *1998 DETC, 10th International Conference on Design Theory and Methodology, ASME* (1998).

Nikolaidis, E., Chen, Q., Cudney, H., Haftka, R. T., and Rosca, R., Comparison of Probability and Possibility for Design Against Catastrophic Failure Under Uncertainty, *Journal of Mechanical Design*, *ASME,* Vo. 126, Issue 3, pp. 386-394, (2004).

Rosca, R., Use of experimental data in testing methods for design against uncertainty. Ph.D. Dissertation, *Aerospace Engineering, Mechanics and Engineering Science Department*, University of Florida, Gainesville, (2001).

P. Walley *Statistical reasoning with imprecise probabilities*, Chapman and Hall, (1991).

R.L. Winkler, Probabilistic prediction: Some experimental results. *Journal of American Statistical Association,* pp. 675-685, (1966).

Appendix 1: Description of the fitting process (fit of probability/possibility distribution functions)

In the possibilistic formulation, to each sample we fit an asymmetric triangular membership function, such that the mean of the sample corresponds to the peak of the membership function. The minimum and the maximum values in the sample are the minimum and the maximum values of the support of the triangular membership function. An example is shown in Fig. A1.

In the probabilistic formulation, we fit a probability density function (PDF, rather than CDF) to each sample. When all data are available, the best PDF fit is given by normal and shifted Gamma density functions. Therefore, even for small data samples (3, 5), we use the normal PDF and the shifted Gamma PDF to fit the data.

To find the shifted Gamma function, we choose the scale and shape parameters such that the mean and standard deviation are the same for the sample and the fitted PDF. We choose the third parameter (shift) as an integer that minimizes the sum of the squares of the differences between sample points and fit at the points of the sample. We choose the two parameters (mean and standard deviation) of the fitted normal PDF to be the mean and standard deviation for the sample.

Figure A2 shows the CDF of experimental data and of the fit for the same Competitor sample as in Fig. A1. Like for scarce data, the comparison of CDFs is more meaningful than the comparison of PDFs.

Appendix 2: Definition of inflation factor

Consider $\{x_1,\ldots x_n\}$, a sample of values of a random variable X. Use of small sample sizes (say 5) for estimating the variance of X, may lead to large statistical errors. It is important to estimate the error in the variance and adjust the variance to account for the error.

<u>If the mean value of the population is unknown,</u> then an unbiased estimator of the variance of the variable is:

$$s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2 \qquad (1)$$

where $\bar{x}$ is the sample mean $= \frac{1}{n}\sum_{1}^{n} x_i$.

The variance of the above estimator is (see Freund and Williams, 1966, pp. 151 (F.7a)):

$$\sigma_{s^2}^2 = \frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)}$$

where μ_4 is the fourth moment of the population about the mean $= \frac{1}{n}\sum_{1}^{n}(x_i - \mu)^4$, and σ^4 is the square of the variance of the population.

The following equation is used to inflate the unbiased estimate of the variance obtained from equation (1):

$$s'^2 = s^2 + r \cdot \sigma_{s^2} = s^2 + r \cdot \sqrt{\frac{\mu_4}{n} - \frac{(n-3)\sigma^4}{n(n-1)}} \qquad (2)$$

where r is called the *inflation factor*.

When both the mean and standard deviation of the population are unknown, we use the corresponding estimates of these values in Eq. (2). Then the variance of the estimated variance becomes:

$$\bar{\sigma}_{s^2}^2 = \frac{1}{n^2}\sum_{1}^{n}(x_i - \bar{x})^4 - \frac{(n-3)}{n(n-1)^3}\left(\sum_{1}^{n}(x_i - \bar{x})^2\right)^2 .$$

The inflated estimate of the variance becomes:

$$s'^2 = s^2 + r \cdot \bar{\sigma}_{s^2} .$$

Table A1 presents the standard deviation of the data to be fitted, before and after we inflate the standard deviation. An inflation factor of 0 corresponds to no inflation. In Table A1, the increase in inflated standard deviation does not vary linearly with the inflation factor, but the increase in inflated variance does.

Table A1: Inflated standard deviation for Rosca's and Competitor's data; the mean of Rosca's data is 33.10 while the mean for Competitor's data is 35.08.

Inflation factor	Inflated standard deviation	
	Rosca Data	Competitor Data
0	6.21	6.30
1	6.76	6.75
2	7.26	7.17
15	12.02	11.30

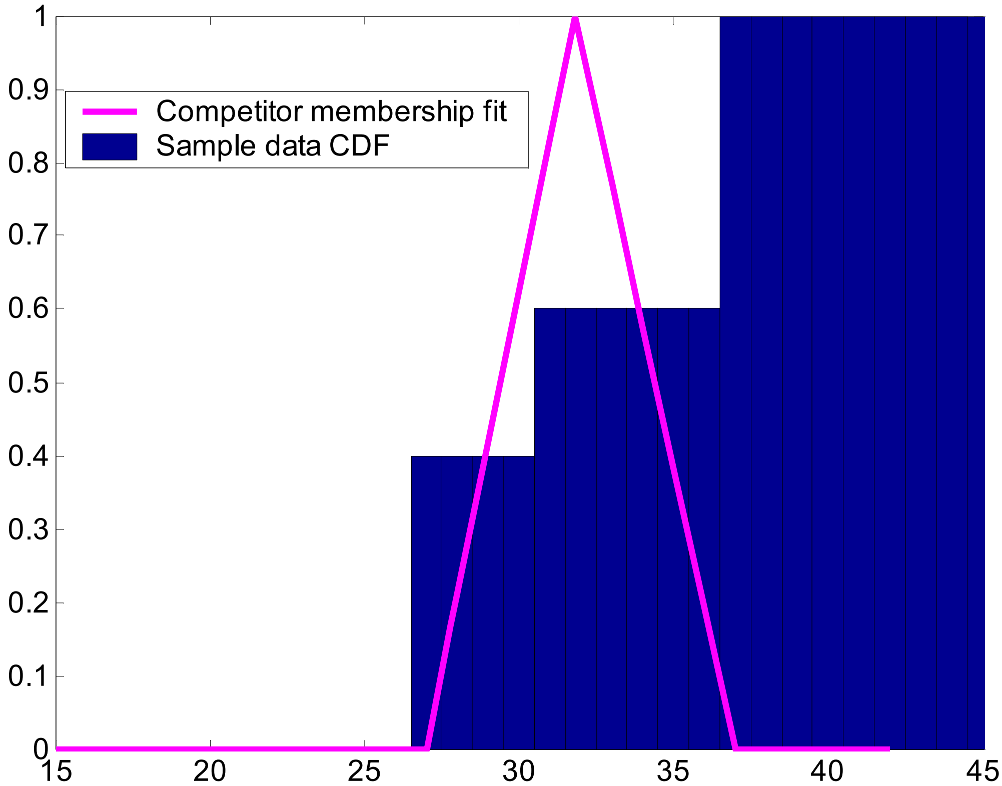

Figure A1: Triangular membership function (solid line) fitted to the sample of 5 from the Competitor's experiments [27 37 37 27 31] and sample cumulative histogram.

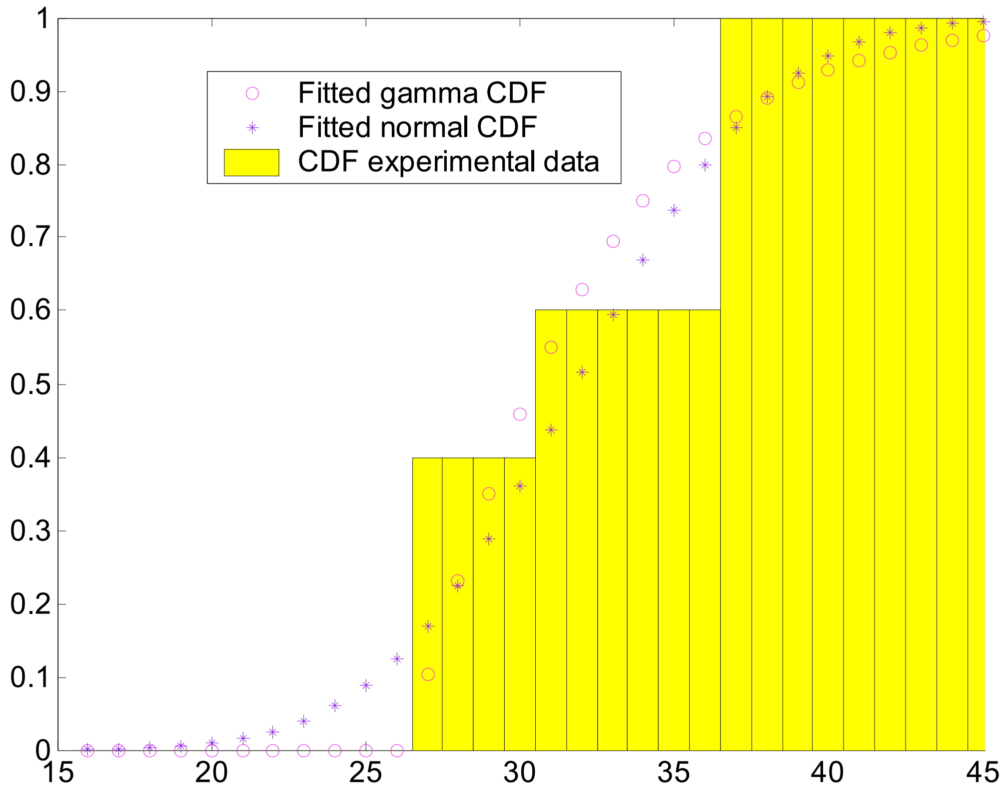

Figure A2: Experimental data CDF (bars), fitted shifted gamma CDF (circles) and fitted normal CDF (asterisks) for the same data as in Fig. A1.