ARE PART WISE COMPARISONS RELIABLE?

DONALD G. SAARI AND KATRI K. SIEBERG

Abstract. It is not unusual for decisions in engineering or customer surveys to compare pairs or subsets of alternatives. Surprisingly, this standard, natural approach can cause valued information to be lost: a loss so severe that it can cause demonstrably incorrect decisions. By understanding why these errors occur, we identify an alternative, closely related decision approach that eliminates these problems. Also, we identify the nature of the lost information, and we show how to compute the likelihood that an incorrect outcome will occur.

1. Introduction

In January, 2002, General Motors announced that it was abandoning portions of its customer survey unit. The fact that GM viewed this unit as expendable raises questions about the effectiveness of current survey approaches. Does this action mean, as suggested by news accounts, that conclusions based on our decision and survey approaches may not be accurate or even particularly useful? Going beyond customer surveys, what about the reliability of the traditional approaches employed to make engineering and design decisions? These are the kinds of basic questions addressed here. After we identify weaknesses of certain standard approaches, we show how to sidestep the difficulties.

A standard way to make decisions is to use pairwise comparisons of the alternatives; i.e., a head-to-head competition is waged over the various criteria. This approach is utilized when comparing design alternatives, making assessments, or conducting market surveys. It may be adopted when selecting a geometric shape or the choice of a material based on how alternatives are ranked over various criteria. An advantage of using pairwise comparisons is that they avoid those complicating side issues associated with simultaneously considering several alternatives, and they eliminate the cost and need to rank all alternatives over each criterion.

While these and other arguments seem to be persuasive, the surprising, counterintuitive fact developed here is that "pairwise comparisons" can generate misleading conclusions by introducing significant errors into the decision process. More generally, even if we use "parts" that are larger than pairs—say, triplets—subtle errors and biases still emerge. To support these unsettling assertions, we show why these decision methodologies can, unwittingly, generate obviously incorrect outcomes with customer surveys and other decision processes. An equally unsettling conclusion is that rather than rare, these problems arise with an alarmingly large likelihood.

The unexpected source of these problems is that these comparisons dismiss valuable information: we outline the amount and kind of lost information. But by knowing what kind of information is lost, we can obtain a positive conclusion. We do this by showing how reliable decisions can be found by "cancelling" the conflicts of pairwise or part-wise decisions.

Saari's research was supported by NSF grant DMI-9971794. Our thanks to two referees for their several excellent suggestions. Some of these results were presented at a 2001 UCI conference on decision making in engineering.

2. Bothersome behavior

It is natural to base the pairwise comparison between alternatives A and B on the number of criteria that rank one better than the other. This includes those settings where criteria are weighted to reflect their relative importance, or market surveys that use the number of individuals who prefer one alternative over the other.

A first surprise is that the order in which alternatives are compared can determine the outcome. For notation, let $< A, B, C, D, E >$ represent where A and B are first compared, the winner is compared with C, that winner is matched with D, and that winner is compared with E. In practice, we usually have no prior information about how the alternatives are ranked over each criterion. But, for purposes of illustration, assume that the unknown rankings for each of three criteria are

$$A \succ B \succ C \succ D \succ E, \quad B \succ C \succ D \succ E \succ A, \quad C \succ D \succ E \succ A \succ B \tag{2.1}$$

where $A \succ B$ means that "A is preferred to B." With the $< A, B, C, D, E >$ ordering

- A is better than B on two of the three criteria, so A is advanced to the second stage,
- C is better than A on two criteria, so C is advanced to the third stage,
- C is better than D or E over all three criteria, so C is the winner.

Compare this conclusion with that obtained by applying the $< D, C, B, A, E >$ ordering to the same data. Here, C advances to the second stage, but B beats C over two criteria so B advances to the third stage, A advances to the fourth, and E is the winning alternative. In other words, by using this different order of comparison, it appears that the "optimal alternative" is E, not C—a particularly bothersome conclusion because E is not "optimal." After all, C and D are superior to E over *all* criteria, and B is superior to E over two of the three criteria: E fares well only when compared with A. Consequently, E's "optimal" status reflects the particular ordering of the pairwise comparisons rather than the data.

While it is arguable that C is the optimal choice, each of these five alternatives is the "winner" with an appropriate ordering; e.g., A is "optimal" with $< E, D, C, B, A >$. As this example demonstrates, the "optimal choice" may more accurately reflect how the alternatives are compared rather than their merits. This means that we must anticipate serious decision errors to arise with pairwise comparisons.

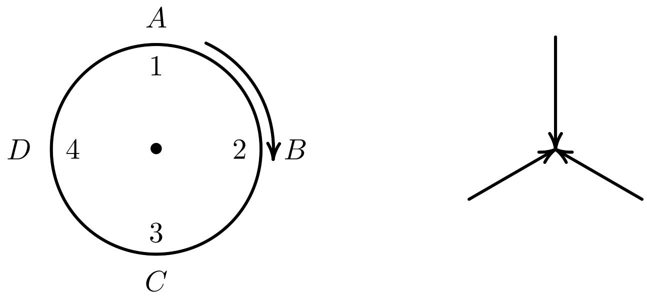
2.1. **Ranking wheel.** To explain the Eq. 2.1 example consider what we call a "ranking wheel." With n alternatives, this freely spinning disk lists the numbers $1, 2, \dots, n$ equally spaced along the inside edge (see Fig. 1a for $n = 4$ alternatives), and the disk is attached to a surface. On the surface, place the names of the alternatives next to a "ranking number;" e.g., in Fig. 1a, A, B, C, D are placed, respectively, next to the numbers 1, 2, 3, 4. The ranking designated by the numbers is $A \succ B \succ C \succ D$. Next, rotate the ranking wheel so that "1" is under the next alternative B, and read off the ranking of $B \succ C \succ D \succ A$. Doing so for all four choices defines the four rankings

$$A \succ B \succ C \succ D, \quad B \succ C \succ D \succ A, \quad C \succ D \succ A \succ B, \quad D \succ A \succ B \succ C. \tag{2.2}$$

Call this set of rankings the "Condorcet four-tuple" in honor of the Marquis de Condorcet (1785) who discovered, seemingly by trial and error, the three-alternative arrangment

$$A \succ B \succ C, \quad B \succ C \succ A, \quad C \succ A \succ B \tag{2.3}$$

in his eighteenth century analysis of voting procedures. For any $n \geq 3$, a Condorcet n-tuple consists of the n rankings similarly constructed by using n ranking numbers on the ranking wheel.



a. Ranking wheel **b.** Force diagram

Fig. 1 Comparisons

Clearly, if each criterion is of equal weight, then no alternative is "better than" any other in the Eq. 2.2 Condorcet four-tuple. This is because by construction each alternative is in first, second, third, and fourth place precisely once. However by sizable 3:1 scores for each pair, the data defines the pairwise cycle

$$A \succ B, \quad B \succ C, \quad C \succ D, \quad D \succ A. \tag{2.4}$$

The source of this pairwise cycle is immediate. Because pairwise decision procedures concentrate solely on the information concerning each *pair,* they cannot recognize, nor react, to these crucial data symmetries that involve all alternatives and indicate a complete tie. The cycles occur, then, because pairwise comparisons inadvertently ignore valuable and available information. By ignoring the "cancelling" information that indicates a tie (based on the full rankings), the cyclic outcomes manifest the cyclic structure of the data.

An analogy that accurately captures the source of the problem is the Fig. 1b force diagram. Here, the three equal forces separated by 120^o cancel to keep the body in equilibrium. If, however, a novice analyzes the forces in pairs, the ignored information (about the third force) generates the false assertion that the forces move the body. Just as the pairwise decision outcomes create a cycle, a pairwise force analysis does the same by incorrectly suggesting that the body is moving in each of the "missing" directions. The power of this analogy is that no one, with any experience, would ever drop a force. The same message applies to decision making: no one, with any experience, whould ever drop an alternative by considering pairs.

2.2. **Other difficulties.** Surprisingly, it takes only parts of a Condorcet n-tuple to cause decision difficulties. Indeed, the Eq. 2.1 data where each alternative is "optimal" uses only three rankings of a Condorcet five-tuple. As experimentation proves, three or more rankings from a Condorcet n-tuple *always generate pairwise cycles.* This explains the introductory example: the Eq. 2.1 data creates the $A \succ B$, $B \succ C$, $C \succ D$, $C \succ E$, $E \succ A$ cycle, so changing the order in which pairs are compared can change the outcome. Similarly, any three rankings of Eq. 2.2 generate the pairwise $B \succ C$, $C \succ D$, $D \succ A$, $A \succ B$ cycle where each alternative is the "optimal" choice with an appropriate ordering of comparisons.

It is, of course, highly unlikely for a "pure" Condorcet n-tuple to arise in practice. But, as we show next, these n-tuples cause difficulties for *almost all decision problems.* The reason is that if the data includes components of this Condorcet type, it can distort the decision outcome. To illustrate, separate a data set into two parts where one subset has the $A \succ B \succ C$ ranking for 5 criteria and $B \succ A \succ C$ for 2 criteria. The decision outcome for this subset is apparent: C always is bottom ranked so the decision is between A and B. As A is superior over more criteria, the final decision outcome should be $A \succ B \succ C$.

Suppose the second data subset over 12 criteria should create a complete tie because it is evenly split over the Condorcet triplet

$$B \succ A \succ C, \quad A \succ C \succ B, \quad C \succ B \succ A.$$

This suggests that when the two subsets of data are combined to create the full data set of

Number	Ranking	Number	Ranking
5	$A \succ B \succ C$	4	$C \succ B \succ A$
4	$A \succ C \succ B$	6	$B \succ A \succ C$

(2.5)

the decision ranking should be $A \succ B \succ C$. (The second data subset creates a complete tie; the tie is broken by information from the first data subset. As an analogy, if forces acting upon a body are broken into two subsets where the second subset is given by Fig. 1b—the equivalent of the Condorcet triplet—the forces in the first subset determine what occurs.) But, this is not the case: because pairwise comparisons cannot recognize the data symmetry of the ranking wheel, the Eq. 2.5 data defines the pairwise rankings of $B \succ A, B \succ C, A \succ C$ by respective tallies of 10:9, 11:8, and 15:4, or the full ranking of $B \succ A \succ C$. The explanation is that the Condorcet data component introduces a bias falsely suggesting that B, not A, is the superior choice. In other words, even portions of the Condorcet n-tuple in a data set can distort the pairwise decision.

Incidentally, not only can Condorcet n-tuples disrupt pairwise outcomes, but we now know that they are the *only* source of problems with pairwise comparisons. (See Saari 1999, 2000 for the proof of this and related results.) This is true for any pairwise comparison method including, for example, Saaty's (1981) AHP approach.

2.3. **Likelihood of distortions.** Rather than rare, distortions caused by the Condorcet n-tuple are almost always present in data. To explain, represent a data set as a point in a vector space where we assume the ranking for each criterion is strict and transitive. (If not, then, after making the obvious modifications in what follows, the same conclusions hold.) With n alternatives, assign each of the $n!$ ways to strictly rank the alternatives to an axis of a $n!$ dimensional Euclidean space. Three alternatives, for instance, define a $3! = 6$ dimensional data space where the $x_1, x_2, x_3, x_4, x_5, x_6$ axes are assigned, respectively, to the

$$A \succ B \succ C,\ A \succ C \succ B,\ C \succ A \succ B,\ C \succ B \succ A,\ B \succ C \succ A,\ B \succ A \succ C$$

ranking. Once the coordinates are given, a data set can be represented by a point in the $n!$-dimensional space where each coordinate specifies the number of weighted criteria with the specified ranking. The data set for Eq. 2.5, for instance, is represented by the point $\mathbf{d} = (5, 4, 0, 4, 0, 6)$. As another example, the choice of the coordinate axes means that the Condorcet triplet of Eq. 2.3 is given by the point $(1, 0, 1, 0, 1, 0)$.

To reduce the number of variables, use *normalized data sets* where integers are replaced by the fractions of the total number of criteria of each type. Thus the above $\mathbf{d}$ with 19 criteria has the normalized representation $(\frac{5}{19}, \frac{4}{19}, 0, \frac{4}{19}, 0, \frac{6}{19})$ that is uniquely determined by specifying five, instead of six, values. This construction forces the normalized data to reside in the simplex

$$S(n!) = \{\mathbf{x} = (x_1, \dots, x_{n!}) \mid \sum_{j=1}^{n!} x_j = 1,\ x_j \geq 0\}.$$

Designate the center point $\mathbf{c}_{n!} = (\frac{1}{n!}, \frac{1}{n!}, \dots, \frac{1}{n!})$ as the origin of this simplex: $\mathbf{c}_{n!}$ is a "neutral" setting because each ranking has the same support as any other ranking. Now, if $\mathbf{d}$ is a data set, then $\mathbf{p} = \mathbf{d} - \mathbf{c}_{n!}$ describes how each component of $\mathbf{d}$ differs from the average:

negative values are below average, while positive values are above. (Notice that the sum of the components of $\mathbf{p}$ equals zero.)

We now introduce a coordinate system on this simplex, where $\mathbf{c}_{n!}$ is the origin, to describe different components of the data as represented by $\mathbf{p}$. For $n = 3$, let the *Condorcet direction* be defined by the axis $(1, -1, 1, -1, 1, -1)$. Notice, the three components with a positive value correspond to the three rankings in the Condorcet triplet defined by $A \succ B \succ C$ while the three components with a negative value correspond to the three rankings in the triplet defined by the reversed $C \succ B \succ A$. For $n \geq 3$, all ways to use the ranking wheel define $(n-1)!/2$ independent Condorcet n-tuple coordinate directions.

An immediate consequence of this coordinate representation is that for a specified data set $\mathbf{d}$ to be free from all distortion effects of the Condorcet n-tuple, $\mathbf{d}$'s vector representation must be in the lower dimensional plane that is orthogonal to all of the Condorcet n-tuple coordinate directions. But since data residing in lower dimensional settings are "unlikely" for all natural probability measures (e.g., it is unlikely to randomly select points from a (x, y, z) coordinate system that always have $x = 0$), the following assertion follows immediately.

Theorem 1. *For $n \geq 3$ different alternatives, and for any probability measure that assigns a zero probability to a lower dimensional coordinate plane in the normalized data simplex, it is with probability zero that a data set is not influenced by the Condorcet n-tuple data.*

This theorem applies primarily to continuous data, but extensions to discrete data are immediate; e.g., the discrete case allows a positive, but very small likelihood that a data set is not influenced by a Condorcet n-tuple. It is easy to go beyond the Thm. 1 message that the Condorcet effect is highly likely to determine whether a specified data set $\mathbf{d}$ has distorting Condorcet terms—just use the vector dot product to determine whether $\mathbf{d} - \mathbf{c_{n!}}$ has a non-zero component in a Condorcet n-tuple direction. In this manner we demonstrate the omnipresent effect of the distorting Condorcet effect on data by establishing the following:

Corollary 1. *Even unanimity data is adversely influenced by components in the Condorcet cyclic direction.*

To explain, if the $A \succ B \succ C$ ranking holds over all criteria, this data set has the normalized unanimity $(1, 0, 0, 0, 0, 0)$ representation: relative to $\mathbf{c}_{3!}$ it has the $(1, 0, 0, 0, 0, 0) - (\frac{1}{6}, \dots, \frac{1}{6})$ representation. That this data include a distorting Condorcet triplet effect is captured by the non-zero dot product

$$((1,\ 0,\ 0,\ 0,\ 0,\ 0) - (\frac{1}{6}, \dots, \frac{1}{6})) \cdot (1,\ -1,\ 1,\ -1,\ 1,\ -1) = 1.$$

For this assertion to make sense, we need to identify the practical effects this "Condorcet distortion" imposes on the unanimity data. To do so, notice that the pairwise rankings of $A \succ B \succ C$ are $A \succ B$, $A \succ C, B \succ C$ where each has the same 1:0 tally. If we just rely on the pairwise outcomes, this common tally suggests that the $A \succ B$, $A \succ C$ rankings have the same intensity: a conclusion that conflicts with the actual data because A clearly is closer to B than to C. It is this useful intensity information that pairwise comparisons lose by failing to recognize the subtle but essentially omnipresent Condorcet triplet effect.

2.4. **More than pairs.** These decisions deficiencies extend beyond pairwise comparisons to trouble decision procedures based on any k-tuple. To explain by using the Condorcet four-tuple data in Eq. 2.2, suppose triplets are used to make decisions. Start, for instance, by ignoring D; i.e., consider only the data relating $\{A,\ B,\ C\}$. Here the Eq. 2.2 rankings become

$$A \succ B \succ C, \quad B \succ C \succ A, \quad C \succ A \succ B, \quad \text{and} \quad A \succ B \succ C.$$

The first three terms define a Condorcet triplet that causes a complete tie. The tie, however, is broken in A's favor because of the remaining $A \succ B \succ C$ ranking. The source of this phenomenon is revealed by "number counting:" a Condorcet n-tuple has n rankings while a k-tuple has only k rankings. Thus, restricting a Condorcet n-tuple to k alternatives creates a Condorcet k tuple *and* $n - k$ *extra* rankings: these extra rankings cause a decision bias. Indeed, the $k = 2$ decision bias for pairwise rankings is illustrated above.

When varying which alternative is to be dismissed, the data from the Condorcet four-tuple always defines a Condorcet triplet and one left-over ranking. These left-overs

$$A \succ B \succ C, \quad B \succ C \succ D, \quad C \succ D \succ A, \quad D \succ A \succ B$$

always create a cyclic effect. It is this cycle of "left-over rankings" caused by concentrating only on "parts" that introduces a disrupting bias with the decision outcomes.

So, the data from k-tuple comparisons fail to recognize the balanced nature of the data that should result in a complete tie. The consequence is obvious: because the pairwise and k-fold comparison approaches inadvertently ignore valuable information about the data, they introduce decision errors. Using the Fig. 1b force diagram analogy, if n forces are acting on a body, an accurate analysis cannot be expected by considering only $k < n$ of the forces.

Theorem 2. *Whenever analyzing data sets concerning n-tuples in terms of k-tuples where $n > k \geq 2$, the components of the Condorcet n-tuple always introduce a distortion effect into the decision outcomes. With any probability measure that assigns a zero probability to a lower dimensional data set in the normalized simplex, it is with probability zero that a data set is free from the distorting influence of the Condorcet n-tuple data.*

A major point we wish to make is that these these paradoxical outcomes occur because whenever a decision process separates the inputs into disconnected parts, there always exists a concomitant, inadvertent loss of crucial information. On the other hand, this explanation identifies what must be done to resolve these problems: decisions approaches must be found that cancel the disrupting Condorcet n-tuple components from the data. This is done next.

2.5. **Correcting the error.** It follows from Arrow's (1963) seminal impossibility theorem that these errors cannot be eliminated just by using the pairs. But there are ways to resolve this problem. To understand them, notice that the tally for each pair in the $A \succ B, B \succ C, C \succ A$ cycle generated by the Condorcet triplet $A \succ B \succ C, B \succ C \succ A, C \succ A \succ B$ is 2:1. By adding the scores each alternative gets over *all* pairwise comparisons, each alternative receives 3 points: all are tied as they should be. In other words, adding the tallies each alternative receives over *all pairs* cancels the disrupting effects and eliminates the omnipresent decision bias introduced by the Condorcet triplet.

The construction of the Condorcet n-tuple in terms of a ranking wheel makes it clear that this adding approach cancels the Condorcet effect for all values of n. This is because for each alternative, say A, going $k \leq n/2$ steps in one direction along the ranking wheel is an alternative, say B, and k steps in the other direction is another alternative, say C. So, if A beats B, then C will beat A by the same tally. Thus, adding tallies over all pairs always cancels the bias introduced by the Condorcet n-tuple.

To illustrate with the Table 2.5 data, this addition approach leads to the scores of

$$A : 9 + 15 = 24, \quad B : 10 + 11 = 21, \quad C : 8 + 4 = 12$$

generating the desired $A \succ B \succ C$ ranking instead of the misleading $B \succ A \succ C$ conclusion obtained from pairwise rankings. To describe this approach by using a table with the Eq. 2.5 data, give each of the 19 criteria a name: let it be a number. Namely, label the five criteria with $A \succ B \succ C$ as criteria 1, ... , 5, those with rankings $A \succ B \succ C$ as criteria 6, ... , 9, etc. The following table lists which alternative is the better one from each pair for each criterion.

	$1-5$	$6-9$	$10-13$	$14-19$	Tally
$\{A,B\}$	$5A's$	$4A's$	$4B's$	$6B's$	$9A's, 10B's$
$\{B,C\}$	$5B's$	$4C's$	$4C's$	$6B's$	$11B's, 8C's$
$\{A,C\}$	$5A's$	$4A's$	$4C's$	$6A's$	$15A's, 4C's$

The final score just counts the number of A's, B's, and C's; i.e., the 24 A's, 21 B's, and 12 C's reproduce the above tally.

To illustrate this procedure with the unanimity data, the scores

$$A : 1+1 = 2, \quad B : 1+0 = 1, \quad C : 0+0 = 0 \tag{2.6}$$

provide the more accurate information that A is better than B and much better than C.

Another advantage of this "adding approach" is that it automatically cancels cylic data. In the next array, for instance, criterion 1 has cyclic information. But, since each alternative is listed once in this column, distortions introduced by the first criterion cancel.

	1	2	3	4&5	Tallies
$\{A,B\}$	A	A	B	$2B's$	$2A's, 3B's$
$\{B,C\}$	B	C	C	$2B's$	$3B's, 2C's$
$\{A,C\}$	C	A	C	$2A's$	$3A's, 2C's$

Here, the outcome is $B \succ A \succ C$ with the tallies of 6:5:4. Also notice from the table that the first criterion defines a cyclic ranking, the second, third, and {fourth and fifth} criteria have, respectively, the ranking $A \succ C \succ B$, $B \succ C \succ A$, and $B \succ A \succ C$.

Incidentally, it is easy to show that the above method is equivalent to the Borda Count: this is where, when the alternatives can be ranked for each criterion, $n - j$ points are assigned to the jth ranked alternative. For intuition, notice from Eq. 2.6 that for a $A \succ B \succ C$ ranking, A beats B and C, so A receives two points; B beats only C so it receives one point; and C receives zero points. This point assignment agrees with the Borda Count. (Properties of the Borda Count are developed and described in Saari 2001.) In practice, however, the above pairwise comparison approach seems to be more appropriate. This is because the pairwise addition approach does not require ranking the alternatives over each criterion, and this approach allows, and handles, cyclic data.

3. Separation loses information

Engineering decisions often are linked in the sense that the $\{A, B\}$ outcome may be combined with the $\{C, D\}$ conclusion. For instance, a customer survey may have $\{A, B\}$ as the two alternatives for a car's body style while $\{C, D\}$ are alternative choices for engine performance. Or, $\{A, B\}$ may involve a geometric shape to be selected for a design while $\{C, D\}$ concerns the selection of the material. These problems differ from those in the previous section because here the decisions are at least tacitly linked. This issue, which is standard in engineering, also arises in voting (see Nurmi 1999, 2002). To show why the concerns from the previous section persist, we compute the likelihood that crucial information about the data is lost.

3.1. **Representing pairwise comparisons.** With two alternatives, suppose alternative A is ranked above B for 60% of the weighted criteria, or with 60% of the individuals in a survey. To geometrically represent this outcome, associate alternatives A and B, respectively, with the right and left-hand endpoint of a unit interval, and plot a point 0.6 of the way from the B endpoint toward the A endpoint. (See Fig. 2a.) This point means that with certainty, 60% of the criteria support A over B. In general, if $0.x$ of criteria supports A over B, place a point $0.x$ of the way from the B endpoint toward the A endpoint.

The outcomes for the two pairs of alternatives, $\{A, B\}$ and $\{C, D\}$, can be simultaneously represented by plotting the point $\mathbf{p} = (0.60, 0.56)$ in the Fig. 2b square. The vertical and horizontal dashed lines indicate each pair's level of support. So, $p_1 = 0.60$ means that 60% support $A \succ B$ while $p_2 = 0.56$ means that C beats D with 56% of the weighted criteria. Outcomes for N pairs are similarly represented with a point in an N-dimensional cube.

But, what does this Fig. 2b outcome, where A beats B (with 60% of the criteria) and C beats D (with 56%) means about the data? For instance, if measuring potential consumer reaction in a product design where $\{A, B\}$ and $\{C, D\}$ represent different alternatives for different features, it is with certainty that 60% of those being surveyed rank $A \succ B$ and with certainty that 56% rank $C \succ D$. Does this outcome mean that most consumers (or, with a decision problem, most criteria) favor *both* $A \succ B$ and $C \succ D$? If so, then it is reasonable to develop a product with the combined $\{A, C\}$ features. But what if the actual data has a sizeable majority of the criteria disagreeing with this total outcome? In this case, a product is designed with the $\{A, C\}$ features runs the risk of commercial failure. Or, when selecting a geometric design and a material, if only a minority of criteria support *both* A and C, this $\{A, C\}$ combination may be feasible but most surely it is not optimal.

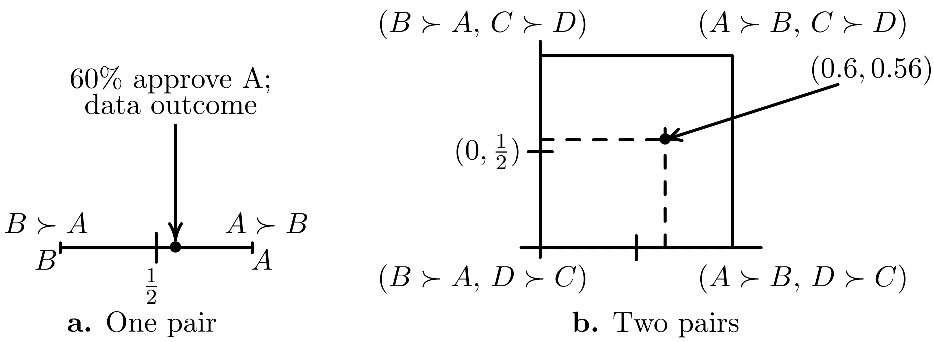


Fig. 2. Data lines and squares

To illustrate with numbers, suppose the Fig. 2b outcomes represent 50 criteria, or 50 people in a survey. With certainty, 30 prefer $A \succ B$ and 28 prefer $C \succ D$. But,

(1) does this outcome mean that over half, 28 of the criteria, support *both* the $\{A \succ B$ *and* the $C \succ D\}$ rankings while 2 disagree with the $C \succ D$ outcome by having the rankings $\{A \succ B$ but $D \succ C\}$ while the remaining 20 criteria reject both outcomes,
(2) or, does it mean that *only* 8 criteria support both outcomes, while the rankings for the sizeable number of 42 criteria, 84% of them, disagree with parts of the final outcome because 22 prefer $\{A \succ B$ but $D \succ C\}$ and the remaining 20 prefer $\{C \succ D$ but $B \succ A\}$?

Both choices generate precisely the same decision outcome for the two pairs, but they have very different interpretations and serious consequences for design decisions. The first scenario provides support for the joint outcome, but the second one, where 84% of the criteria disagree with parts of the outcome, creates significant doubt about the appropriateness of selecting the $\{A, C\}$ combination. Namely, if the second scenario holds, then product design decisions made on the basis of the joint $A \succ B, C \succ D$ outcome could be inferior or even disastrous.[1]

In order to address these issues we need to find all possible data sets that generate the same associated pairwise outcomes. We do this by showing how to find all possible disaggregations of the outcome by geometrically representing the data sets and the pairwise decision outcomes on the same diagram.

3.2. **Data line.** The two pairs $\{A, B\}$ and $\{C, D\}$, divide the the criteria into four types according to how each pair is ranked over each criterion. For notation, let $\nu(A, C)$ be the fraction of all criteria registering $\{A \succ B \text{ and } C \succ D\}$; similar definitions hold for $\nu(A, D), \nu(B, C), \nu(B, D)$. Thus a normalized data set $\mathbf{d}$ becomes the four-vector

$$\mathbf{d} = (\nu(A, C), \nu(A, D), \nu(B, C), \nu(B, D)).$$

With the car body style and engine example, $\nu(A, C)$ represents the fraction of respondents in a survey that prefer body style A and engine type C. The above two 50-criteria sets are, respectively, $\mathbf{d}_1 = (\frac{28}{50}, \frac{2}{50}, 0, \frac{20}{50})$ and $\mathbf{d}_2 = (\frac{8}{50}, \frac{22}{50}, \frac{20}{50}, 0)$.

Assume the pairwise outcome $\mathbf{p} = (p_1, p_2)$ from $\mathbf{d}$ satisfies $p_1 \geq p_2 \geq \frac{1}{2}$. If false, then it can be made true by interchanging the $\{A, B, C, D\}$ labels. Since p_1 is the fraction of all criteria supporting $A \succ B$, $p_1 = \nu(A, C) + \nu(A, D)$. Thus

$$\mathbf{p} = (p_1, p_2) = (\nu(A, C) + \nu(A, D), \nu(A, C) + \nu(B, C)). \tag{3.1}$$

Divide the data into the subset of criteria with the $B \succ A$ ranking (represented by the left edge of the square), and the subset with $A \succ B$ rankings (represented by the right edge). Next, divide the criteria with a $B \succ A$ ranking according to how they rank C and D. Namely, compute the $\{C, D\}$ pairwise decision by using only information from this data set and represent the outcome with a point $\mathbf{p}_L$ on the left vertical line interval as in Fig. 3a. Similarly, determine how the criteria supporting the $A \succ B$ ranking split between C and D. Again, represent this pairwise decision outcome as a point $\mathbf{p}_R$ on the right edge of the box.

Stated in words using the car body style and engine example, the $\mathbf{p}_L$ point indicates the fraction of those respondents who prefer body style B that also like engine type C. Similarly, $\mathbf{p}_R$ indicates the fraction of those respondents who prefer body style A that also like engine type C.

Expressing $\mathbf{p}_L$, $\mathbf{p}_R$ in terms of the $\mathbf{d}$ information is a simple computation; e.g., as $\mathbf{p}_L$ is the fraction of all criteria with a $B \succ A$ ranking that also has a $C \succ D$ ranking, it is given by $\nu(A, C)/[\nu(A, D) + \nu(A, C)]$. The $\mathbf{p}_L$, $\mathbf{p}_R$ coordinates are

$$\mathbf{p}_L = (0, \frac{\nu(B, C)}{\nu(B, C) + \nu(B, D)}), \quad \mathbf{p}_R = (1, \frac{\nu(A, C)}{\nu(A, C) + \nu(A, D)}), \tag{3.2}$$

or, equivalently,

$$\mathbf{p}_L = (0, \frac{\nu(B, C)}{1 - p_1}), \quad \mathbf{p}_R = (1, \frac{\nu(A, C)}{p_1}). \tag{3.3}$$

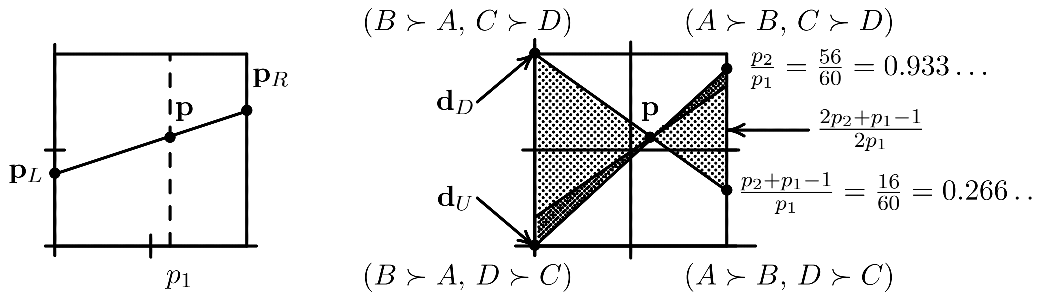
Intuitively, the joint pairwise outcomes $\mathbf{p} = (p_1, p_2)$ should depend upon the relative number of criteria in the two data sets. This happens because

$$\mathbf{p} = (1 - p_1)\mathbf{p}_L + p_1\mathbf{p}_R. \tag{3.4}$$

But Eq. 3.4 is the equation of a straight line, so this means that the combined decision outcome $\mathbf{p}$ is found by first plotting $\mathbf{p}_L$ and $\mathbf{p}_R$ and then connecting them with a straight line. The

[1] It is interesting how when these results are presented, almost always members of the audience suggest the illustrating examples of the design of the Edsel, or the QFD decision method.

outcome $\mathbf{p} = (p_1, p_2)$ is the intersection of this line with the vertical line $x = p_1$ (the Fig. 3a dashed line) that identifies the fraction of all criteria (people in a survey, etc.) preferring $A \succ B$.



a. Data line **b.** All supporting data sets

Fig. 3. Data set representations

So, a specified data set $\mathbf{d}$ uniquely determines $\mathbf{p}_L = (0, y_L), \mathbf{p}_R = (1, y_R)$, the joint outcome $\mathbf{p}$, and the line. Of particular importance is that the converse is true. Namely, given a line segment and a particular point on the line, it just takes algebra to recover the associated normalized data set. Namely, specified $\mathbf{p}_L, \mathbf{p}_R$, and $\mathbf{p}$ values uniquely determine a normalized data set $\mathbf{d}$. This observation leads to the following definition.

Definition 1. *A* data line with distinguished point $\mathbf{p}$ *in a representation square is given by a straight line connecting points on the vertical edges and a distinguished point,* $\mathbf{p}$*, on the line. The points on the vertical edges represent how these two sets of criteria supporting* $A \succ B$ *or* $B \succ A$ *rank the alternatives* $\{C, D\}$*. The distinguished point* $\mathbf{p}$ *is determined by the fraction of all criteria supporting* $A \succ B$ *as represented by the horizontal axis;* $\mathbf{p}$ *represents the outcomes of separate pairwise comparisons.*

To summarize, a data set $\mathbf{d}$ defines a data line with distinguished point $\mathbf{p}$. Conversely, *each line that connects opposite vertical edges of the square and passes through* $\mathbf{p}$ *is a data line that identifies* a unique normalized data set $\mathbf{d}$ *with the pairwise outcomes* $\mathbf{p}$. Thus we can *identify all possible data sets that support a specified outcome* $\mathbf{p}$. Call this set the *data cone*: this two-dimensional cone, or triangle, is the set of all possible lines that connect the vertical edges and pass through $\mathbf{p}$.

The data cone represented by $\mathbf{p}$ is always defined by the extreme data sets. For Fig. 3b, the two extreme data lines (denoted by $\mathbf{d}_U$ and $\mathbf{d}_D$ for "upward" and "downward" sloping) start from a vertex on the left edge and pass through $\mathbf{p} = (p_1, p_2)$. (By naming the alternatives so that $p_1 \geq p_2 \geq \frac{1}{2}$, this vertex condition always holds.) The equations for these bounding data lines are

$$y = \frac{p_2 - 1}{p_1}x + 1 \text{ for } \mathbf{d}_D \text{ and } y = \frac{p_2}{p_1}x \text{ for } \mathbf{d}_U. \tag{3.5}$$

The extreme y values of $\frac{16}{60}$ and $\frac{56}{60}$ on the right edge of Fig. 3b follow from Eq. 3.5 with $\mathbf{p} = (0.60, 0.56)$ and $x = 1$.

By construction, the $\mathbf{d}_D$ and $\mathbf{d}_U$ lines pass through *vertices* of the square. This means that their associated data sets must include a pair that is not supported by any criteria; e.g., the normalized set for $\mathbf{d}_D$ has $\nu(B, D) = 0$ and the one for $\mathbf{d}_U$ has $\nu(B, C) = 0$. Illustrating with the earlier 50 voter illustration, $\mathbf{d}_U$ and $\mathbf{d}_D$ are, respectively, the first and second example.

For a given line (i.e., for specified $\mathbf{p}_L, \mathbf{p}_L$ and $\mathbf{p}$), the normalized data set is computed by using Eqs. 3.1, 3.3. For instance, to find $\nu(A, C)$ for line $\mathbf{d}_D$, we have that $\frac{\nu(A,C)}{p_1} = \frac{16}{60}$ where $p_1 = 0.60$, or $\nu(A, C) = 0.16$. In this manner, the extreme data sets for $\mathbf{p} = (0.60, 0.56)$ are:

Profile	**approve all**	**approve only** $A \succ B$	**approve only** $C \succ D$	**approve none**
$\mathbf{d}_U$	56%	4%	0	40%
$\mathbf{d}_D$	16%	44%	40%	0

(3.6)

As Table 3.6 displays, the data sets supporting $\mathbf{p}$ vary from where support for both majority winners A and C (the $\nu(A, C)$ term) has only 16% of all criteria (with $\mathbf{d}_D$) to 56% (with $\mathbf{d}_U$). Disagreement with at least one of A or C ranges from 44% (with $\mathbf{d}_U$) to an surprisingly large 84% of the criteria. Complete disapproval of both outcomes (the $\nu(B, D)$ term) ranges from 0% to 40%. When viewed in terms of a design decision involving, say, a product design, rather than selecting A and C, this variance in interpreting the outcome should cause pause. For instance, although $\mathbf{d}_D$, $\mathbf{d}_U$ have the same pairwise outcomes, if $\mathbf{d}_D$ represents the data, the largest portion (44%) of the criteria prefer the combination of A and D while only 16% like the A and C combination. Consequently, the A C combination could be a costly error.

This listing, particularly the entries indicating an extreme lack of support for the final combined conclusion, strongly indicates that considering separate, disjoint parts can cause a serious loss of information about any relationships among the combined parts. For instance, with the $\mathbf{d}_D$ data, 73% of those surveyed who prefer body style A actually prefer engine style D, not C. It is this loss of relational information that can cause costly design decisions in terms of economics, quality, etc. depending on the decision variables—errors that can be prevented. More precisely, instead of knowing $\mathbf{d} = (\nu(A, C), \nu(A, D), \nu(B, C), \nu(B, D))$, we merely know the outcome $\mathbf{p}$ that could result from a large range of data sets with surprisingly contrary interpretations. By ignoring information about how the criteria connect the pairs, we create the danger that a majority of the criteria need not embrace the combined outcomes.

3.3. **More pairs.** With two pairs, all data sets that yield $\mathbf{p}$ have some criteria that approve of both outcomes; e.g., if A and C are majority winners, then $\nu(A, C)$ may be small, but it is not zero. Because $\mathbf{p}$ is in the upper right quarter region, no supporting data line can meet the $(1, 0)$ vertex so $y_R > 0$ for all data lines. This means that all supporting data sets have some criteria that approve of both pairwise outcomes.

With three or more pairs of alternatives, however, the final outcome need not be supported by any of the criteria, or potential customers, or ... To explain, the problem is again caused by an effect similar to the Condorcet n-tuple. To illustrate, notice that when describing the Condorcet triplet in terms of pairs, we obtain

Ranking	$\{A, B\}$	$\{B, C\}$	$\{A, C\}$
$A \succ B \succ C$	$A \succ B$	$B \succ C$	$A \succ C$
$B \succ C \succ A$	$B \succ A$	$B \succ C$	$C \succ A$
$C \succ A \succ B$	$A \succ B$	$C \succ B$	$C \succ A$
Outcome	$A \succ B$	$B \succ C$	$C \succ A$

(3.7)

Altering the names of the alternatives generates the table over three separate pairs

Criteria	$\{A, B\}$	$\{C, D\}$	$\{E, F\}$
1	$A \succ B$	$C \succ D$	$E \succ F$
2	$B \succ A$	$C \succ D$	$F \succ E$
3	$A \succ B$	$D \succ C$	$F \succ E$
Outcome	$A \succ B$	$C \succ D$	$F \succ E$

(3.8)

where *none of the three criteria agrees with the three combined outcomes!* The argument of the last section extends to assert that with three or more pairs, this Condorcet effect almost always affects decision outcomes.

3.4. **Likelihood estimates.** The $\mathbf{d}_D$ data line proves that a substantial percentage of criteria can disapprove of parts of a $\mathbf{p}$ outcome even though each pair's victory is by a substantial margin. But do examples, such as $\mathbf{d}_D$, represent isolated oddities, or do they identify a troubling behavior that must be addressed? To answer this concern, recall that with a single $\{A, B\}$ comparison, it is with certainty that the winning alternative is supported by at least 50% of the criteria. A natural question with two pairs is to determine the likelihood that at least 50% of all criteria agree with both $\mathbf{p}$ outcomes.

The first step is find all data lines that satisfy the 50% approval behavior. Since 0.60 of the criteria support the $A \succ B$ comparison, the only way at least half of all criteria approve both outcomes is if $\nu(A, C) = y_R \times 0.60 \geq \frac{1}{2}$, or if $y_R \geq \frac{5}{6}$ of the criteria from the data set on the right prefer $C \succ D$. To find all data lines satisfying this condition, draw the line passing through the $y_R = \frac{5}{6}$ value on the right edge and $\mathbf{p}$. As all data lines with $y_R \geq \frac{5}{6}$ have the desired 50% approval behavior, this set of data lines is the slender heavier shaded region of Fig. 3b. Notice the relatively small size of the heavily shaded region; this size already indicates that it is reasonably rare for data sets to have a 50% approval of the combined outcome.

The likelihood estimates of consumer (or criteria) satisfaction, or dissatisfaction, with the joint outcome can be extracted from the figure. As these values require imposing probability assumptions about the distribution of the data, we consider two commonly used assumptions; results from other distributions are computed in a similar manner.

3.5. **Uniform distribution.** Since the Fig. 3b geometry captures all data lines supporting $\mathbf{p}$, a first estimate is to assume that the likelihood of any property is represented by the ratio of the number of endpoints of data lines with the desired property divided by number of endpoints of data lines with the $\mathbf{p}$ outcome. A rough estimate of this likelihood is the ratio of the *length* of the heavily shaded region on one edge to the *length* of the fully shaded region on that edge. If each data line is equally likely, the likelihood of a specified behavior is the edge length of the segment defining the behavior divided by the edge length of all possible outcomes.

To illustrate with the Fig. 3b example, the segment of endpoints on the right edge satisfying this 50% agreement property is $\frac{5}{6} \leq y \leq \frac{56}{60}$ with length $\frac{56}{60} - \frac{5}{6} = \frac{1}{10}$. The segment of right endpoints with data lines with this $\mathbf{p}$ outcome is $\frac{16}{60} \leq y \leq \frac{56}{60}$ with length $\frac{2}{3}$. The ratio of these lengths, only 0.15, defines the proportion of data lines that support at least a 50% approval of both outcomes. Consequently, with a uniform distribution over the data sets, there is only a 15% likelihood that 50% of the criteria support the joint conclusion. Conversely, we reach the surprising conclusion that although each of the Fig. 3b pairwise comparisons receives strong support, *it is with a* 0.85 *likelihood that over half of the criteria are dissatisfied with at least part of the outcome!* When this conclusion is interpreted in terms of product design decisions that involve considerable investments, the odds of having the wrong interpretation are intimidatingly large and suggest a considerable risk.

To further illustrate this geometry, we determine the likelihood that at least a quarter of the criteria disagree with both outcomes. These criteria are represented by the $\nu(B, D)$ value, or the vertex in the lower left hand edge. Thus, we are interested in those data lines where the left point is below height y_L where $y_L \times \frac{28}{50} \leq \frac{1}{4}$, or $y_L \leq \frac{25}{56} \approx 0.446$. In other words, by assuming that all data lines are equally likely, it follows from Fig. 3b that *with probability of about* 0.446 *at least a quarter of all criteria disagree with both pairwise outcomes.*

So, what level of support—what $\mathbf{p}$ values—ensure that at least 50% of the respondents support both winning outcomes? This is a special case of the following more general assertion.

Theorem 3. *Assume a uniform distribution of normalized data sets. To ensure that with probability P at least a fraction x of all criteria support the joint outcome, the pairwise outcomes $\mathbf{p} = (p_1, p_2)$ must satisfy the inequality*

$$\frac{p_2 - x}{1 - p_1} \geq P. \tag{3.9}$$

To have a 50% likelihood ($P = \frac{1}{2}$) that half of the criteria support the joint outcome ($x = \frac{1}{2}$), it follows from Eq. 3.9 that $2p_2 + p_1 \geq 2$. Since $p_1 \geq p_2$, this means that only if $p_1 \geq \frac{2}{3}$ and p_2 has nearly this value can we hope for a 50% chance that half the data supports both outcomes. Namely, both pairs need around a 67% victory to allow only a 50% chance that the data supports both victorious outcomes! To ensure (so $P = 1$) that half of the data ($x = \frac{1}{2}$) supports both outcomes, we need that $p_1 + p_2 \geq \frac{3}{2}$. Consequently, a necessary (but not sufficient) requirement for such assurances is the surprisingly high value of $p_1 \geq 0.75$.

The proof of Thm. 3 uses simple algebra. To achieve the total approval fraction $x = \nu(A, C)$, we have from Eq. 3.3 that the lower edge of the data cone has a y_R component satisfying $x = \nu(A, C) = y_R/p_1$, or $y_R = xp_1$. The likelihood data satisfies this condition (with the uniform distribution) is the ratio of the edge lengths of what would be the darker cone and the full data cone. So, just solve the inequality

$$[\frac{p_2}{p_1} - \frac{x}{p_1}]/[\frac{p_2}{p_1} - \frac{p_1 + p_2 - 1}{p_1}] \geq P.$$

3.6. **Binomial and Normal distributions.** Rather than a uniform probability distribution, with $m = p_1 n$ criteria that support $A \succ B$, it is more realistic to use the binomial distribution to determine the likelihood that certain numbers of criteria have one $\{C, D\}$ ranking or the other. For purposes of exposition, use the neutral assumption that $p = \frac{1}{2}$ to compute the likelihood that at least half of the criteria support the joint decision. Namely, for Fig. 3, compute $P(y_R \geq \frac{5}{6})$ with the binomial distribution. Converting into numbers, the goal is to find the likelihood that 25, 26, 27, or the maximum of 28 of the 30 criteria with a $A \succ B$ ranking also have $C \succ D$. Equivalently, find the likelihood that only 2, 3, 4, or 5 of these criteria have $D \succ C$. According to the binomial distribution, the likelihood[2] is

$$(\frac{1}{2})^{30} \sum_{j=2}^{5} \binom{30}{j} = (\frac{1}{2})^{30} \sum_{j=25}^{28} \binom{30}{j} = 1.62 \times 10^{-4}.$$

Consequently, even with $p_1 = 0.60, p_2 = 0.56$, it is highly improbable (with 10^{-4} or essentially a zero likelihood) that at least half of the criteria support the joint outcomes! The amount of information lost by using pairwise comparisons is significant: design decisions based on these procedures run a serious risk of being in error.

Even for reasonable choices of $\mathbf{p}$, computations provide discouraging information about the reliability of decisions based on pairwise comparisons. As a personal aside, while we anticipated discovering problems with pairwise comparisons, we were surprised about how unlikely it is to have support for a joint outcome. With reflection, the answer comes from the central limit theorem: with any reasonable numbers of criteria, the bulk of the probability centers near the mean.

[2] For technical combinatoric reasons, the actual likelihood is even smaller

This observation suggests other new conclusions. For instance, with a given $\mathbf{p}$, how can we estimate the fraction of all criteria or participants who approve of both outcomes? To find this, notice that the y component of the midpoint of the shaded region on the right edge of Fig. 3b is $\nu(A,C)/p_1 \approx [2p_2 + p_1 - 1]/2p_1$, or $\nu(A,C) \approx [2p_2 + p_1 - 1]/2$. Using the Fig. 3b example and the binomial distribution, there is a strong likelihood that only about 36% of the criteria support both outcomes. These assertions are easy to extend (to obtain, for instance, versions of Thm 3 for different probability distributions) by using standard tools from probability.

All of these conclusions extend to decisions involving more than pairs. For instance, with $\{A_1, A_2, \dots, A_n\}$ representing n geometric designs and $\{C_1, \dots, C_m\}$ representing m choices of material, or consumer surveys over different aspects of a potential product, the geometry becomes much the same. The square becomes an object with an n-dimensional base and m-dimensional sides. While the figure cannot be drawn, it is clear that the same loss of information must occur.

4. Concluding comments

Surprisingly often, decision procedures introduce a bias because the procedure loses information. On the other hand, once it is understood what kind of information is being dismissed, alternative decision approaches can be designed. With rankings of alternatives over criteria, either adding the victories each alternative scores in competitions over all criteria and all pairs eliminates the central problem. (For an expository description of the role of a related procedure in elections, see (Saari, 2001).) In other words, a way to avoid problems is to use the Borda count. This approach avoids the problems of lost information, and it provides an outcome that can be justified. For an analytic analysis of this method, see Saari 1999, 2000, 2001. However, rather than the actual Borda Count, we recommend the adding approach described earlier.

If the comparison is between, say, a geometric design and the choice of a material, and if the combinations of the outcomes matter, then the decisions cannot be disconnected. Rather than comparing $\{A, B\}$ and $\{C, D\}$ separately, the comparison *must* be among the four alternatives of $(A,C), (A,D)$, (B,C), (B,D). As proved above, it is surprisingly likely for decoupling to introduce serious errors. Again, we recommend ranking these four alternatives by using the Borda Count in its alternative form given earlier.

References

[1] Arrow, K.J. (1963) *Social Choice and Individual Values* 2nd. ed., Wiley, New York.

[2] Condorcet, M. (1785) *Éssai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Paris.

[3] Haunsperger, D., and D. G. Saari (1991) The lack of consistency for statistical decision procedures, *The American Statistician* **45**, 252-255.

[4] Nurmi, H. (1999) *Voting Paradoxes and How to Deal with Them*, Springer-Verlag, NY.

[5] Nurmi, H. (2002) *Voting Procedures undeer Uncertainty*, Springer-Verlag, NY.

[6] Saari, D. G (1999) Explaining all three alternative voting outcomes, *Jour. Econ. Theory* **87** 313-355.

[7] Saari, D., G. (2000) Mathematical structure of voting paradoxes I and II, *Economic Theory* **15**, 1 - 101.

[8] Saari, D. G. (2001) *Chaotic elections! A mathematician looks at voting*, American Mathematical Society, Providence, RI.

[9] Saari, D., G., K. Sieberg (2001) The sum of the parts can violate the whole, *Amer. Pol. Science Review* **95**, (Number 2), 415-433.

[10] Saaty, T. (1980) *Analytic Heirachy Process,* McGraw Hill, New York.

Institute for Mathematical Behavioral Sciences, University of California, Irvine, Irvine, CA 92697-5100
E-mail address: dsaari@uci.edu

Fulbright Professor, University of Tampere,, Tampere, Finland and, Political Science, Binghamton University, Binghamton, NY 13902
E-mail address: katri.sieberg@binghamton.edu