Estimating Negotiator Performance Without Preference Information

Rudolf Vetschera
University of Vienna
Austria
Rudolf.Vetschera@univie.ac.at

Abstract

In empirical studies of negotiation support systems, it is often not possible to elicit utility functions from experimental subjects, since this would lead to additional interventions into the behavior of subjects. The present paper develops several methods to evaluate the performance of negotiators in multi-issue negotiations without referring to their utilities. These methods are empirically compared using data from the NSS Inspire. Our results indicate that negotiators quite frequently behave in a way which is inconsistent with their estimated utility functions, thus a performance measure which is not based on utility values could provide a more robust basis for empirical studies of negotiator performance.

1 Introduction

During the last years, a considerable number of Negotiation Support Systems (NSS) have been developed both by academic researchers, e.g. Inspire. (Kersten & Noronha, 1999) or NegoIsst (Schoop, Jertila, & List, 2003) and by industry, e.g. SmartSettle (`www.smartsettle.com`). There are considerable differences in the methodological approaches underlying these systems. Similar to classifications in the field of Group Decision Support Systems (DeSanctis & Gallupe, 1987; Vetschera, 1990), one can distinguish between systems which mainly focus on the communication process and systems which provide analytical support to enable negotiators to find efficient (Pareto optimal) solutions to their problems.

While there is some empirical research on NSS (Moore, Kurtzberg, Thompson, & Morris, 1999; Jain & Solomon, 2000; Kersten, Koeszegi, & Vetschera, 2002; Koeszegi, Vetschera, & Kersten, 2004), most of these studies have analyzed the impact of one specific system. Only recently have researchers begun to empirically study the differences in impact of various systems. One important question in this context is the additional benefit offered by analytical support as compared to systems that only support communication processes.

A difficult problem in conducting empirical studies that compare different NSS is the definition and measurement of benefits of using such a system. Some obvious types of benefits can be measured objectively like reduced time to reach an agreement, or subjectively using well-established instruments like increased user satisfaction with an agreement. These measures are similar to concepts used in the evaluation of information systems (DeLone & McLean, 1992; Farbey, Land, & Targett, 1995), and instruments developed in that field can be used.

But the ultimate goal of an NSS is to assist the parties (or one party) to achieve better outcomes. In negotiations dealing with several issues, a comparison of outcomes achieved with different systems would require knowledge about the multiattribute utility function of negotiators. When the utility functions of all parties involved in a negotiation are known, it is easy to compare compromise solutions across systems, or to determine whether a compromise is Pareto-optimal.

However, in empirical studies which try to compare different types of NSS, utility functions of the parties are not readily available. One important goal of such studies is to evaluate the advantages of analytical decision support. Eliciting a utility function is a method of analytical support in itself. When experimental subjects, who are using an NSS without analytical support, go through an elicitation of their utility functions, their perceptions of the problem and their way of making decisions during the negotiation might change and become more similar to negotiators using a system with analytical support. This effect

could obscure the very behavioral differences which one wants to study.

This argument precludes the elicitation of utility functions before a negotiation experiment. But eliciting preferences after an agreement has been reached also leads to problems. Since subjects already know the outcome of the negotiation, they might attempt to bias the utility function to make the compromise look more favorable. There is an obvious incentive to do so if rewards for subjects are tied to their performance in the experiment. But even if subjects do not consciously distort the preference information they provide, there is a danger of inconscious distortions to avoid cognitive dissonances and to ex post rationalize one's behavior during the experiment.

Thus, for empirical studies comparing NSS with and without analytical support, one needs methods to evaluate the outcomes of multi-issue negotiations without explicitly referring to the utilities of the parties involved. In the present paper, we study different methods for this purpose.

The remainder of this paper is structured as follows: in section two, we introduce an approach for preference-free performance measurement, which is based on the dominance relation. In section three, we discuss several interpretations of this measure, which lead to different extensions and related approaches. These variants are compared in section four using empirical data. Section five discusses the consequences of these empirical results and provides an outlook on future research questions.

2 A Dominance-Based Approach

We consider a negotiation in which two (or more) parties bargain over several issues (attributes). The total number of attributes is denoted by K, and we assume that all K attributes are relevant for all parties. For simplicity, we will present most of our arguments for the case of two parties only, but the approach easily generalizes to more parties.

Within each attribute, the parties must agree on one compromise value, which is to be chosen from a discrete set of possible values. While some attributes (e.g. the price in a buyer-seller negotiation) could be interpreted as continuous decision variables, the restriction to a given set of discrete values is not overly restrictive and is also made in several NSS (e.g. Inspire). We denote the number of discrete values in attribute k by N_k.

We also assume that for each attribute, the direction of improvement is known for each party (e.g. that the seller in a buyer-seller negotiation prefers a higher price over a lower price) or, more generally, that for each attribute and each party, there exists a known ranking of the discrete attribute values. We denote the ranks of values in attribute k by r_{ik}. For a given party, values are sorted in increasing order of preference, i.e. r_{ik} is

preferred to $r_{i-1,k}$ and so on. A decision alternative A_i can thus be characterized by a vector of rank numbers of all attributes:

$$A_i = (r_{i,1}, \ldots, r_{iK}) \tag{1}$$

When (ordinal) preferences within each attribute are known, it is possible to establish a dominance relation between alternatives. Alternative i dominates alternative j, iff the value of alternative i in each attribute is considered to be as good as that of alternative j, and strictly better in at least one attribute, that is:

$$\begin{array}{ll} \forall\, k: & r_{ik} \geq r_{jk} \\ \exists\, k: & r_{ik} > r_{jk} \end{array} \tag{2}$$

The dominance relation provides a partial ordering of alternatives (Pomerol & Barba-Romero, 2000). Thus the position of an alternative in that relation is an indicator for the quality of the alternative. The worst alternative $A_0 = (1, \ldots, 1)$ is dominated by all the other alternatives, and the ideal alternative $A_{Ideal} = (N_1, \ldots, N_K)$, which has the best values in all attributes, dominates all the other alternatives.

We can obtain a measure for the quality of an alternative by looking at the number of other alternatives which it dominates and the number of other alternatives by which it is dominated. For our setting, both values can be computed without explicitly constructing the dominance relation.

When the value of an alternative in attribute k has rank r_k within that attribute, there are $r_k - 1$ values to which that value is preferred, or r_k values which are considered at most as good. The same argument can be applied to all attributes. By forming the combinations of all those values, we obtain the total number of alternatives which are dominated by alternative $A_i = (r_{i1}, \ldots, r_{iK})$ as

$$LO(A_i) = \prod_k r_{ik} - 1 \tag{3}$$

The correction term -1 in (3) is necessary to take into account that an alternative does not dominate itself.

A similar argument can be used to determine the number of other alternatives which dominate a given alternative. The number of values which are better than or equal to r_{ik} in attribute k is $N_k + 1 - r_{ik}$, thus the total number of alternatives dominating A_i is:

$$UP(A_i) = \prod_k (N_k + 1 - r_{ik}) - 1 \tag{4}$$

The values $LO(A_i)$ and $UP(A_i)$ have several interesting properties. First, from their definitions as the numbers of alternatives dominated by or dominating a given alternative,

it is obvious that

$$\prod_k N_k - UP(A_i) \geq LO(A_i) \tag{5}$$

and

$$\prod_k N_k - UP(A_i) - LO(A_i) - 1 \tag{6}$$

is the number of alternatives which neither dominate nor are dominated by A_i.

Furthermore, when the ranks of values are reversed for two parties (for example, ranks of attributes like price for buyer and seller), then the value of $LO(A_i)$ for one party is identical to $UP(A_i)$ for the opponent.

3 Interpretations and Extensions

The two performance measures LO and UP can be interpreted in different ways. Formally, LO can be interpreted as a multiplicative form of a multiattribute utility function (Keeney & Raiffa, 1976), where the r_{ik} represent partial utility values and all attributes have identical weights of one.

This interpretation suggests to compare this measure to other forms of multiattribute utility functions. A common form is the additive function

$$u(A_i) = \sum_k w_k u_k(x_{ik}) \tag{7}$$

where x_{ik} represents the attribute value (rather than the rank of that value among all possible values) of alternative i in attribute k, w_k is the weight for that attribute, and $u_k(x)$ is the marginal utility function for attribute k.

Since we do not have information about weights or marginal utility functions, simplifying assumptions have to be made. A straightforward assumption concerning weights is to assume equal weights for all attributes, which can, without loss of generality, be set equal to one. For the marginal utility functions, we assume linearity. Denote the largest value in attribute k by $\overline{x_k}$ and the smallest value by $\underline{x_k}$. Assuming that all attributes are to be maximized, the marginal utility values are then given by

$$u_k(x_{ik}) = \frac{x_{ik} - \underline{x_k}}{\overline{x_k} - \underline{x_k}} \tag{8}$$

Using these simplifying assumptions, we can define the *simple additive weighting* performance measure as

$$SAW(A_i) = \sum_k \frac{x_{ik} - \underline{x_k}}{\overline{x_k} - \underline{x_k}} \tag{9}$$

It should be noted that this approach uses more information from the decision problem, since it takes into account the attribute values rather than their ranks. However, this information is readily available from the case descriptions given to experimental subjects, and does not require any utility elicitation from the subjects.

Interpreting the measure LO as a utility function leads to other interpretations. Taking the logarithm of LO, we obtain

$$\begin{aligned} LO + 1 &= \prod_k r_{ik} \\ LLO = ln(LO + 1) &= \sum_k ln(r_{ik}) \end{aligned} \tag{10}$$

Thus the logarithm of $LO + 1$ can be interpreted as an additive utility function, where the logarithm is used as the marginal utility function in each attribute and all weights are equal to one. Using the logarithm for marginal utility functions is a plausible way to represent decreasing marginal benefits of the attributes. We therefore will also consider performance measure LLO as an alternative way to determine the performance of negotiators.

The performance measures SAW and LLO are derived from interpreting LO as a multi-attribute utility function. One can also extend the original interpretation of this measure as a count of worse (dominated) or better (dominating) alternatives.

$LO(A_i)$ can be considered as a lower bound of the estimated rank of alternative A_i in the preference order of the decision maker. If the decision maker's preferences obey the axiom of dominance, all alternatives which are dominated by A_i must be ranked behind that alternatives. When rankings are numbered from the worst to the best alternative, then A_i has at least rank $LO(A_i)$.

Let

$$M = \prod_k N_k \tag{11}$$

denote the total number of alternatives. Then by a similar argument as above, $M - UP(A_i)$ is an upper bound on the rank number of alternative A_i. Based on this interpretation, we can consider the midpoint of the interval between those two values,

$$AVG_0 = (LO(A_i) + (M - UP(A_i))) / 2 \tag{12}$$

as an approximation of the rank number of alternative A_i.

As we have already mentioned, when the preferences of two parties are exactly opposed, the value of LO for one party is identical to the value of UP for the other party. Thus, for AVG_0, the sum of evaluations of the two parties is given by:

$$(LO(A_i) + M - UP(A_i))/2 + (UP(A_i) + M - LO(A_i))/2 = M \tag{13}$$

and thus is constant. Consequently, AVG_0 models the negotiation problem as a zero-sum game, in which all alternatives are Pareto-optimal.

LO and UP are based on the dominance relation, which takes into account preferences of the negotiator only in terms of the direction in which an attribute is to be optimized. During the negotiation, a negotiator reveals more information about his or her preferences. The observation that one alternative is preferred to another alternative can be used to extend the dominance relation. Using this extended relation, measures similar to LO, UP, and AVG_0 can be constructed, which provide a more precise estimate of the rank of an alternative in the negotiator's preference ranking.

Specifically, we can make two assumptions about a negotiator's preferences towards various alternatives which are discussed during the negotiation:

1. A negotiator prefers the final compromise to any offer made by the opponent during the negotiation.
2. A negotiator prefers all offers made by himself or herself during the negotiation to the final compromise.

The first assumption is quite plausible. If a negotiator prefers an offer made by the opponent to the final compromise, it should be possible to return to that previous offer, which has already been on the table, and make it the final compromise. It is unlikely that an opponent would refuse this move, since the opponent has already made that offer himself. Assuming that the opponent has made concessions after that offer, such a move would improve the position of both parties and thus increase efficiency of the outcome.

The second assumption could be viewed as more problematic. One could argue that if a negotiator identifies an opportunity to improve his or her position during the negotiation, this opportunity should be exploited. This assumption therefore presumes that negotiators start with a position which is close to their ideal point, and make concessions during the negotiation process, rather than start from a weak position and search for mutual improvements.

Both assumptions can be used independently of each other. Thus we consider three types of extensions to the dominance relation by adding preferences between the opponent's

offers and the final compromise, preferences between the negotiator's own offers and the final compromise, and both types of preferences.

To formalize this approach, we represent a relation between alternatives by a binary matrix B, where $b_{ij} = 1$ indicates that alternative A_i is preferred to alternative A_j. Initially, matrix B is set to represent the dominance relation, i.e. a one is entered in b_{ij} if A_i dominates A_j.

Denote the set of offers made by the opponent by $O^P = \{o_1^P, \ldots, o_n^P\}$ and the set of offers made by the negotiator himself by $O^S = \{o_1^S, \ldots, o_n^S\}$. The values o_i^P and o_i^S represent the indices of alternatives. Furthermore, we denote the final compromise by index i^*.

Preferences towards the opponent's offers are added to the extended dominance relation by setting

$$b_{i^*,o_j^P} = 1 \ \ \forall\, o_j^P \in O^P \tag{14}$$

and preferences concerning the negotiator's own offers by setting

$$b_{o_j^S,i^*} = 1 \ \ \forall\, o_j^S \in O^P \tag{15}$$

For the third type of extension, both (14) and (15) are used. We denote the extensions of matrix B by B^{opp}, which is obtained by applying (14), B^{own} from applying (15) and B^{both} from both extensions. In general, we refer to an extended relation as B^x, where $x \in \{own, opp, both\}$.

These changes will add only a few elements to the dominance relation. Under the assumption that the negotiator's preferences obey the transitivity axiom, the relation can further be extended by forming its transitive closure, i.e. by setting all those elements $b_{ij}^x = 1$ for which there exists a sequence of indices $k_1, k_2, \ldots k_n$ so that

$$b_{i,k_1}^x = b_{k_1,k_2}^x = \ldots = b_{k_{n-1},k_n}^x = 1 \tag{16}$$

The transitive closure of the boolean matrix B^x can be computed using Warshall's algorithm (Warshall, 1962).

Similar to the measures LO and UP defined above, we obtain the number of alternatives to which a given alternative is preferred according to relation B^x as

$$ELO^x(A_i) = \sum_j b_{ij}^x \tag{17}$$

and the number of alternatives which are preferred to alternative A_i as

$$EUP^x(A_i) = \sum_j b_{ji}^x \tag{18}$$

where ELO^x and EUP^x indicate the extended versions of the measures LO and UP, respectively. Similar to (12), we can define a measure

$$EAV^x = \left(ELO^x(A_i) + (M - EUP^x(A_i))\right)/2 \tag{19}$$

as another approximation of the rank of alternative A_i in the negotiator's preference order.

The extensions to the dominance relation are different for the two sides of a negotiation. Therefore, (13) no longer holds and the problem is not modeled as a zero sum game. Using these measures, some solutions may be considered as inefficient.

4 Empirical Results

In the following section, we present the application of the measures defined above to empirical data from two sets of experiments. The data of the first set is taken from a negotiation experiment performed in the academic year 2002/2003 during a joint course on International Negotiations held at Concordia University, Montreal and the University of Vienna, Austria (Koeszegi & Kersten, 2003). During these experiments, students negotiated both via a simple Internet-based negotiation platform, which only provided communication support, and the negotiation support system Inspire (Kersten & Noronha, 1999), which also provides analytical support based on the elicitation of the negotiator's utility function via conjoint measurement. For the analysis in this paper, we will only use the data on negotiations performed via Inspire, since this data allows us to compare the different measures to the "true" utilities as measured in Inspire. In total, 15 two-party negotiations were performed using this system, out of which 14 led to a compromise. Thus we can use data points of 28 individual negotiators.

The second data set consists of negotiations performed using the Inspire system between 1996 and 2000. In total, 1606 negotiations were carried out in Inspire in this time frame. After deleting those negotiations which did not lead to an agreement, for which no offers or utility evaluations from either side were available, or in which a negotiator performed several estimations of the utility functions (which is permitted in Inspire), a total of 651 negotiations (or 1302 data points from individual negotiators) remained for this analysis,

In both data sets, the same case was used. It involves a buyer-seller negotiation, in which the parties negotiate about four issues: price, delivery time, payment and the return of defective parts. The case specifies 5 possible values for price, 4 levels for delivery time and 3 levels each for payment and return of defective parts. Thus, there are $5 \times 4 \times 3 \times 3 = 180$ possible alternatives. In all attributes, the rankings of buyers and sellers are strictly reversed.

4.1 Results for data set 1

Figure 1 shows the values of measures LO and UP for all 180 possible alternatives. The filled circles represent the alternatives which were actually chosen as compromise solutions in the experiments. The measures LO and UP here refer to the position of the Seller. Since in this problem, the preferences of the two parties are strictly opposite, the LO values for the buyer correspond to the UP values of the seller and vice versa. Thus figure 1 can also be interpreted as a representation of the problem in the utility space of the two parties.

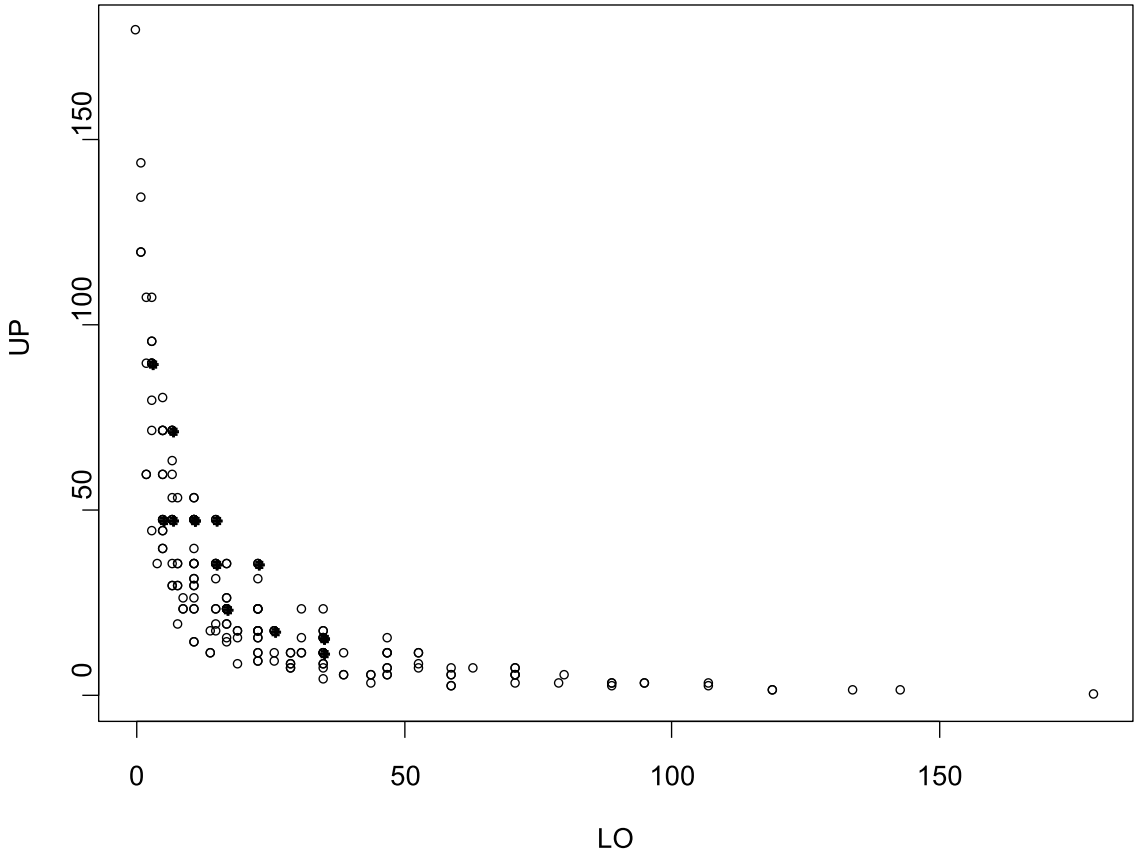

Figure 1: Distribution of all alternatives and compromise solutions selected in LO/UP-space, data set 1

Most negotiations lead to a rather balanced compromise giving both parties results which are evaluated between 0 and 50. Only in a few instances, the buyer side was able to achieve considerably better outcomes.

In some instances, the solution seems to be dominated. However, dominance in LO/UP-space needs to be interpreted with caution when, as is the case in our experiments, there is a different number of levels in different attributes. An exchange of ranks between two attributes will leave the value of LO unchanged, but changes the value of UP and thus leads to dominance in LO/UP-space.

Consider for example two alternatives, which are characterized by the rank vectors $A_1 = (4, 3, 2, 3)$ and $A_2 = (4, 2, 3, 3)$, thus alternative A_1 is somewhat better in the second attribute, while A_2 is better in the third attribute. The value of LO is the same for both alternatives. But if the number of levels in the second and third attribute are different, then the value of UP for the two alternatives will be different. So one alternative appears to be dominated in LO/UP space.

Figures 2 and 3 show the relationship between the utility values as determined by Inspire and the scalar measures introduced in sections 2 and 3. These figures indicate that the performance measures all generate a rather similar pattern.

Table 1 shows the correlation coefficients between all the measures tested. The measures defined above are highly correlated with each other. Their correlation with the utility values elicited by Inspire are a bit lower, but still highly significant.

Table 2 compares the correlation coefficients (ρ) and their significance levels (p) between the different measures and the utility values estimated by Inspire for buyers and sellers. For all measures, the fit is considerably better for buyers, while for sellers, the correlation coefficients in some instances fail to reach the 5 % significance level.

Interestingly, this phenomenon occurs mostly for the more elaborate measures based on the extended dominance relation. This result is quite surprising, since we expected that taking into account the actual behavior of subjects should improve the fit of our measures to their utilities. In fact, the extension of the preference relation in some instances had reduced the difference between upper and lower bound considerably, as can be seen from figure 4, where we plot the bounds for both the dominance relation and the extended relation across all experiments. The figure also shows that these bounds in most instances do move in the same direction as the utility values calculated by Inspire, which are plotted as unconnected points in the middle of the graph.

4.2 Results for data set 2

Table 3 shows the correlation coefficients of the different measures to the true utility values as determined by Inspire for this data set. While the correlation between most measures based on the (extended) dominance relation remains high, correlation coefficients with the utility values measured by Inspire are much lower than in the first data set.

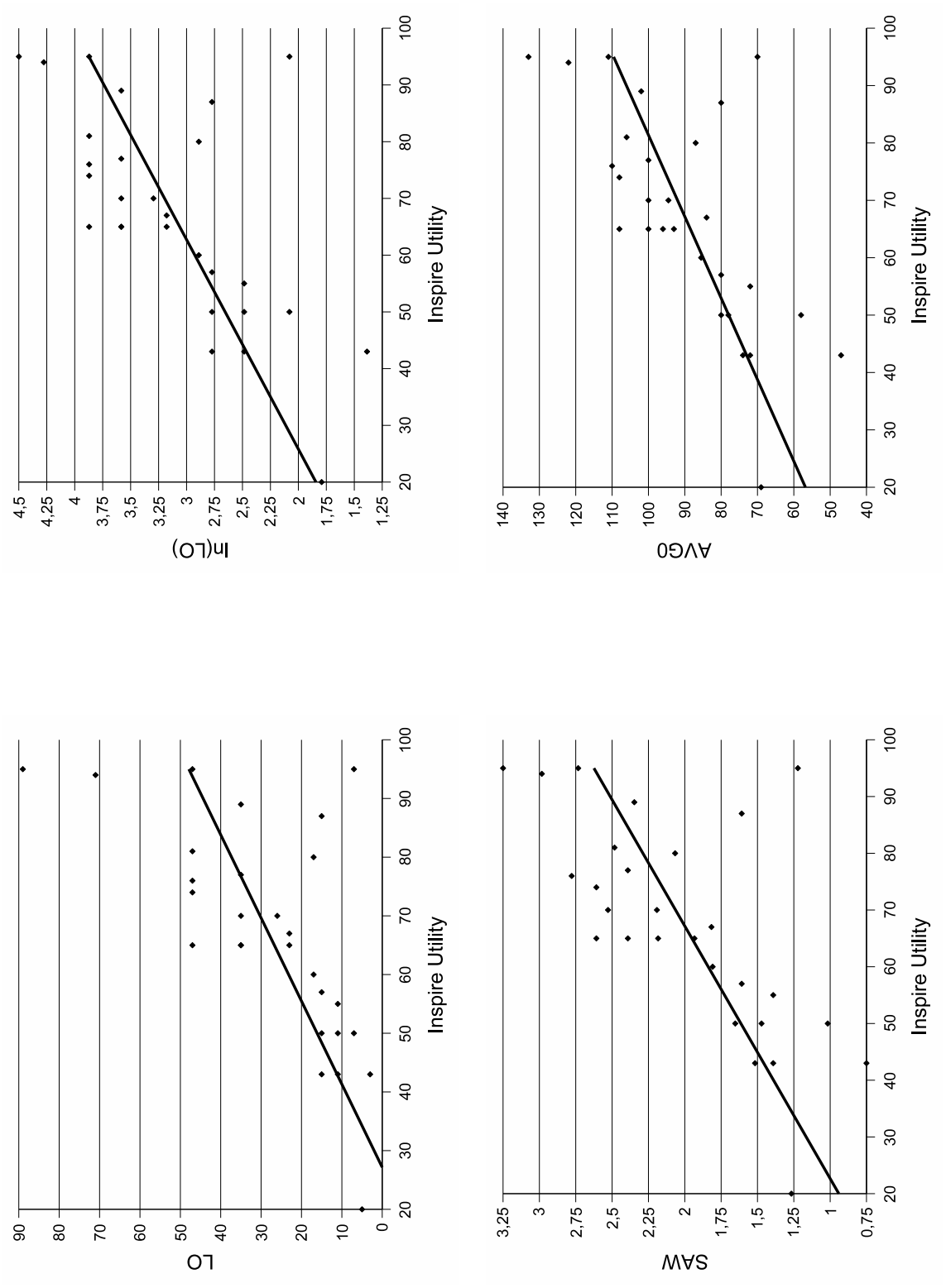

Figure 2: Correlation with Inspire utility, basic measures

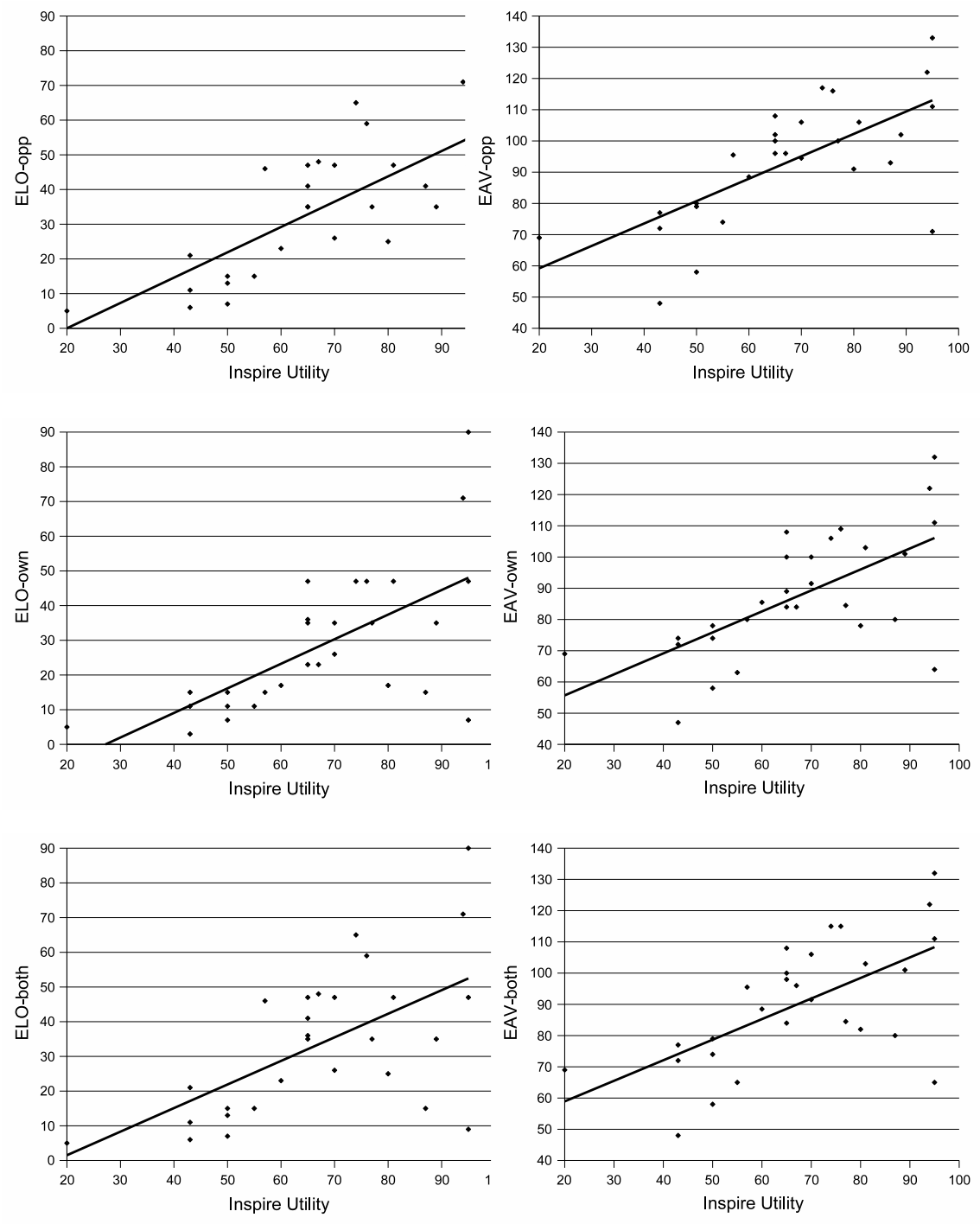

Figure 3: Correlation with Inspire utility, extended dominance measures

	LO	ln((LO)	AVG0	SAW	ELO-opp	ELO-own	ELO-both	EAV-opp	EAV-own	EAV-both
Inspire	0.64821	0.66654	0.68427	0.67067	0.65063	0.64680	0.59528	0.68975	0.64129	0.61372
	0.0002	0.0001	<.0001	<.0001	0.0002	0.0002	0.0008	<.0001	0.0002	0.0005
LO		0.90957	0.94457	0.93327	0.90340	0.99994	0.92223	0.89961	0.93862	0.89782
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
ln(LO)			0.97188	0.96732	0.87429	0.90839	0.88236	0.95905	0.94364	0.92898
			<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
AVG0				0.98804	0.88162	0.94349	0.89378	0.97270	0.97760	0.95050
				<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
SAW					0.86958	0.93179	0.88606	0.96004	0.96753	0.94229
					<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
ELO-opp						0.90270	0.97366	0.93930	0.89474	0.93124
						<.0001	<.0001	<.0001	<.0001	<.0001
ELO-own							0.92165	0.89819	0.93678	0.89568
							<.0001	<.0001	<.0001	<.0001
ELO-both								0.92717	0.89969	0.94032
								<.0001	<.0001	<.0001
EAV-opp									0.96044	0.97428
									<.0001	<.0001
EAV-own										0.97865
										<.0001

Table 1: Correlation coefficients between performance measures, first data set

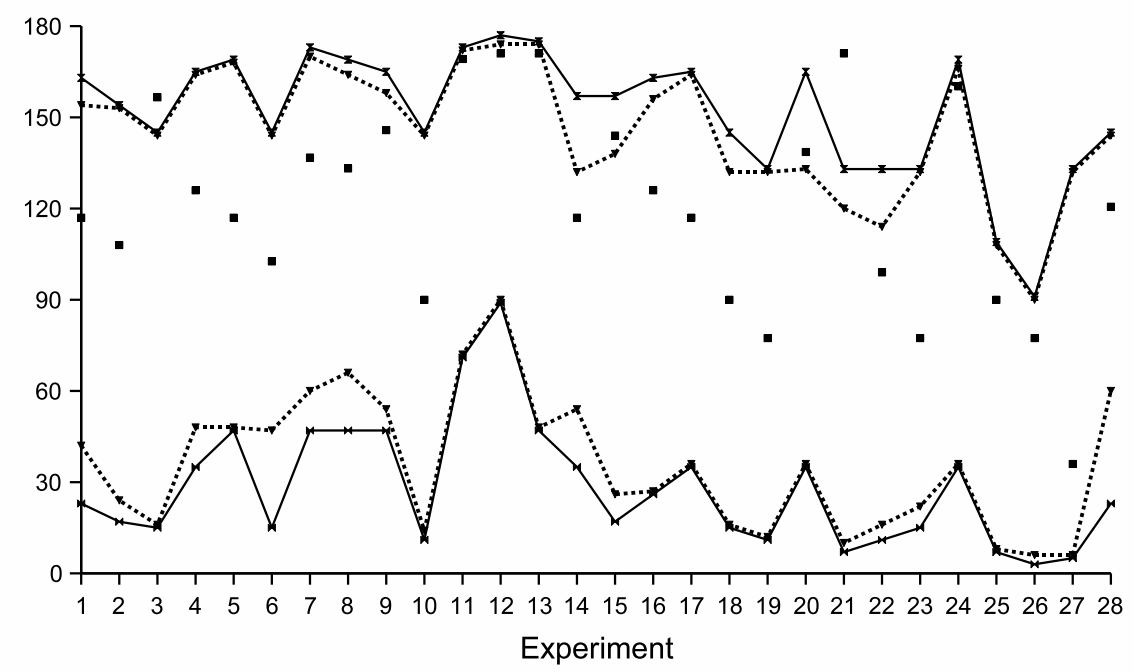

Figure 4: Effect of extending the dominance relation

		Buyers	Sellers
LO	ρ	0.72769	0.54939
	p	0.0032	0.0419
$ln(LO)$	ρ	0.69877	0.54753
	p	0.0054	0.0427
AVG_0	ρ	0.71968	0.57325
	p	0.0037	0.0321
SAW	ρ	0.68336	0.56785
	p	0.0071	0.0342
ELO^{opp}	ρ	0.72147	0.52517
	p	0.0036	0.0538
ELO^{own}	ρ	0.72538	0.54939
	p	0.0033	0.0419
ELO^{both}	ρ	0.57124	0.52517
	p	0.0329	0.0538
EAV^{opp}	ρ	0.73783	0.58396
	p	0.0026	0.0283
EAV^{own}	ρ	0.73289	0.46792
	p	0.0029	0.0915
EAV^{both}	ρ	0.63357	0.47808
	p	0.0150	0.0838

Table 2: Correlation coefficients for performance measures to Inspire utilities, first data set, separated for buyers and sellers

Like in the first data set, taking the average between lower and upper bounds of the rank interval increases the fit to the utility values calculated by Inspire. But the effect of adding different elements to the dominance relation was different. In the first data set, adding either the preferences concerning the negotiator's own or the opponent's offers had the same (small) effect on the correlation to Inspire utility values, and adding both preferences decreased the fit. In the second data set, adding the preferences concerning the negotiator's own offers increases the fit considerably. This is rather surprising, as the theoretical argument for this type of preferences was much weaker than the argument concerning the opponent's offers.

As table 4 shows, this effect holds for both buyers and sellers. Similar as in the first data set, the fit of most measures to the Inspire utility values is much better for buyers than for sellers.

The inconclusive results concerning the effects of taking into account observed behavior require some additional analysis. One explanation could be a violation of the assumptions.

	LO	ln((LO)	AVG0	SAW	ELO-opp	ELO-own	ELO-both	EAV-opp	EAV-own	EAV-both
Inspire	0.25451	0.28253	0.28009	0.02440	0.22785	0.26741	0.22750	0.27603	0.31201	0.29325
	<.0001	<.0001	<.0001	0.3791	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
LO		0.87679	0.90574	0.04377	0.83735	0.97063	0.83693	0.82801	0.86159	0.81493
		<.0001	<.0001	0.1144	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
ln(LO)			0.94427	0.07526	0.75336	0.85167	0.75298	0.87012	0.89368	0.85514
			<.0001	0.0066	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
AVG0				0.04852	0.79480	0.88503	0.79444	0.93250	0.93360	0.90039
				0.0801	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
SAW					0.04699	0.04443	0.04697	0.05127	0.04982	0.05201
					0.0901	0.1091	0.0902	0.0644	0.0723	0.0606
ELO-opp						0.85184	0.99997	0.91242	0.77681	0.88452
						<.0001	<.0001	<.0001	<.0001	<.0001
ELO-own							0.85144	0.84462	0.87950	0.82581
							<.0001	<.0001	<.0001	<.0001
ELO-both								0.91234	0.77568	0.88365
								<.0001	<.0001	<.0001
EAV-opp									0.90175	0.95646
									<.0001	<.0001
EAV-own										0.95527
										<.0001

Table 3: Correlation coefficients between performance measures, second data set

		Buyers	Sellers
LO	ρ	0.35992	0.16840
	p	$< .0001$	$< .0001$
$ln(LO)$	ρ	0.38049	0.23303
	p	$< .0001$	$< .0001$
AVG_0	ρ	0.40070	0.20584
	p	$< .0001$	$< .0001$
SAW	ρ	-0.06758	0.11852
	p	0.0851	0.0024
ELO^{opp}	ρ	0.33566	0.13693
	p	$< .0001$	0.0005
ELO^{own}	ρ	0.38253	0.17356
	p	$< .0001$	$< .0001$
ELO^{both}	ρ	0.33529	0.13658
	p	$< .0001$	0.0005
EAV^{opp}	ρ	0.39625	0.20437
	p	$< .0001$	$< .0001$
EAV^{own}	ρ	0.43586	0.24761
	p	$< .0001$	$< .0001$
EAV^{both}	ρ	0.41570	0.22942
	p	$< .0001$	$< .0001$

Table 4: Correlation coefficients for performance measures to Inspire utilities, second data set, separated for buyers and sellers

Using the utility values elicited by Inspire, we can check whether, according to these utility values, the negotiators should actually have preferred their own offers to the final compromise and the compromise to the opponent's offers or not.

Assumptions fulfilled		Data set 1		Data set 2	
Own	Opponent	N	%	N	%
no	no	0	0.00	52	3.96
no	yes	5	17.86	124	9.45
yes	no	3	10.71	474	36.13
yes	yes	20	71.43	662	50.46

Table 5: Compatibility of observed behavior with assumptions

Table 5 gives an overview of the number of violations of assumptions concerning the ratings of alternatives when using the utility functions elicited by Inspire. This table lists result for both data sets. However, due to the small sample size, data set 1 cannot be used for further statistical analysis. A negotiator was classified as inconsistent with respect to the negotiator's own offers, if for at least one offer of the negotiator, the utility of that offer was lower than the utility of the final compromise. Similarly, a negotiator was classified as inconsistent with respect to the opponent's offers if the utility of at least one of those offers was higher than the utility value of the final compromise.

Only about half of the negotiators behaved in a way compatible with the assumptions all the time. In interpreting table 5, one should keep in mind that a negotiator is considered to be inconsistent if he or she violated the assumptions with respect to only one out of possibly many offers made during the negotiation. When individual offers are counted, the utility evaluations of 88.3 percent of all offers (from both sides) were consistent with the assumptions. The corresponding number for the negotiators' own offers is 95.2% and for offers from the opponent 82.4%.

Several explanations are possible for this apparent lack of compatibility with the assumptions. One possible explanation is that negotiators indeed behaved inconsistently, and for some reasons refused to revert to previous offers made by their opponent, even if that would have improved their position in the final compromise.

Another explanation could be that the utility values elicited by Inspire provide only an imperfect representation of the true preferences of negotiators. Our data does not allow us to distinguish between those two explanations. The fact that users in the first data set performed more consistent with the assumptions is also compatible with both explanations, since the course from which those subjects were recruited covered both negotiation theory and multiattribute utility theory. Thus subjects in the first data set should have performed

the elicitation tasks better than average subject.

Compatible	Own	no	no	yes	yes
with	Opponent	no	yes	no	yes
Measure	N	52	124	465	661
LO	ρ	-0.05085	0.24756	0.26018	0.30958
	p	0.7204	0.0056	$< .0001$	$< .0001$
$ln(LO)$	ρ	-0.08061	0.18383	0.34611	0.31311
	p	0.5700	0.0410	$< .0001$	$< .0001$
AVG_0	ρ	-0.07641	0.19402	0.33284	0.31742
	p	0.5903	0.0308	$< .0001$	$< .0001$
SAW	ρ	-0.23839	0.04118	0.08776	0.00612
	p	0.0888	0.6497	0.0586	0.8752
ELO^{opp}	ρ	0.06037	0.19412	0.25786	0.31940
	p	0.6707	0.0308	$< .0001$	$< .0001$
ELO^{own}	ρ	0.23324	0.24711	0.26331	0.30990
	p	0.0961	0.0057	$< .0001$	$< .0001$
ELO^{both}	ρ	0.06091	0.19325	0.25755	0.31960
	p	0.6679	0.0315	$< .0001$	$< .0001$
EAV^{opp}	ρ	0.08649	0.16283	0.33482	0.32245
	p	0.5421	0.0708	$< .0001$	$< .0001$
EAV^{own}	ρ	0.09677	0.24271	0.35576	0.32602
	p	0.4950	0.0066	$< .0001$	$< .0001$
EAV^{both}	ρ	0.05349	0.21180	0.34844	0.33077
	p	0.7064	0.0182	$< .0001$	$< .0001$

Table 6: Correlation coefficients with Inspire utilities for various levels of compatibility with assumptions

As table 6 shows, correlation between the utilities measured by Inspire and the performance measures developed here is worst for those subjects who violated both assumptions. This is not surprising for the measures based on the extended dominance relation, since the extensions explicitly are based on preferences which for those subjects differ from the preferences implied by Inspire utilities. But this low (and in some instances even – insignificantly – negative) level of correlation also holds for measures in which no extensions of the dominance relation were performed, like LO or AVG_0.

While one would expect that correlation between the measures and Inspire utilities is mostly influenced by violations of the assumptions which are actually used in the different measures, this is not the case. The correlation of measures ELO^{opp} and EAV^{opp} to Inspire utilities is higher for those subjects who, according to Inspire utilities, should have

preferred an offer from the opponent over the compromise than for those subjects in the second column, who violated the assumption about the negotiator's own offers, although this assumption is not used for those measures.

For measures ELO^{own} and EAV^{own}, which use only the assumption about the negotiator's own offers, there is some improvement in fit when this assumption is fulfilled, but it is not very high.

Compatible with	Own Opponent	no no	no yes	yes no	yes yes	All
*Inspire	Mean	62.75	72.24	57.66	67.82	64.41
	Std	21.53	14.70	19.08	18.41	19.20
LO	Mean	27.94	26.86	27.02	26.57	26.82
	Std	21.14	19.32	19.54	18.91	19.25
$ln(LO)$	Mean	3.04	3.04	3.05	3.05	3.05
	Std	0.84	0.73	0.76	0.71	0.73
AVG_0	Mean	90.00	90.00	90.19	90.00	90.06
	Std	19.49	17.63	17.65	17.11	17.43
SAW	Mean	2.00	2.00	2.01	2.00	2.00
	Std	0.61	0.56	0.55	0.52	0.54
$*ELO^{opp}$	Mean	43.62	33.61	34.25	29.59	32.19
	Std	39.24	19.95	22.79	19.60	22.10
ELO^{own}	Mean	26.58	26.89	27.40	26.60	26.91
	Std	15.56	19.33	19.37	18.91	18.98
$*ELO^{both}$	Mean	43.75	33.65	34.30	29.59	32.23
	Std	39.25	19.95	22.81	19.60	22.11
EAV^{opp}	Mean	98.52	93.36	93.61	91.49	92.70
	Std	23.57	18.14	19.03	17.47	18.42
EAV^{own}	Mean	81.48	86.64	86.88	88.50	87.46
	Std	23.57	18.14	18.95	17.47	18.38
EAV^{both}	Mean	90.00	90.00	90.30	90.00	90.10
	Std	30.55	18.58	20.46	17.83	19.49

* indicates significant differences

Table 7: Values of performance measures for different levels of compatibility with assumptions

Table 7 shows the performance levels indicated by the various levels for the different levels of compatibility with the assumption. Only for three measures (Inspire utilities, ELO^{opp}, and ELO^{both}, marked by an asterisks in table 7) a significant difference between the indicated performance levels was found using a nonparametric Kruskal-Wallis test. Even

for these measures, the pattern is rather puzzling. Especially the measures based on the extended dominance relation seem to indicate that negotiators performed better the more they violated the assumptions.

5 Discussion and Conclusions

These empirical results lead to several conclusions concerning the usefulness of the different measures introduced in this paper. The first conclusion concerns simple additive weighting. This method is only very weakly correlated with the other measures, especially with the utility values as determined by Inspire. This lack of correlation with other measures is surprising, given that linear models have been shown to be quite robust in other contexts (Stewart, 1996). However, in view of these results, we can conclude that simple additive weighting is probably not an adequate method to evaluate the performance of negotiators in our setting.

For all the other measures, there is a clear difference between situations in which the behavior of negotiators was compatible with the assumptions and when it was not. When the assumptions are fulfilled, most measures, even those which do not rely on the assumptions, correlate quite well with the utility values as calculated by Inspire and are also highly correlated with each other. Thus the choice of a particular performance measure in this case will probably not have a strong influence on the results of empirical studies which need to operationalize negotiator performance.

But when the assumptions are violated, there is a dramatic decrease in correlation. This phenomenon not only occurs for those measures which are explicitly based on assumptions on the evaluation of offers, but also in the measures LO and AVG_0, which only use the values of the final compromise and make no assumptions about behavior during the negotiation process.

Obviously, Inspire's utility evaluation on one hand and the measures introduced in this paper on the other hand measure something different when negotiators do not behave as assumed. The question now arises which construct is closer to "true performance" of the negotiators. One can argue that when subjects behave in a way which is inconsistent with their utility functions as measured by Inspire, these functions have only limited value as a description of their preferences. Inspire not only elicits utility functions, but also displays the utility value for each offer made by a negotiator or the opponent. Thus, negotiators do know when they have rejected a previous offer from their opponent which had a higher utility value to them than the final compromise. Still, more than 40% of the subjects included in this analysis behaved in this way and thus knowingly overruled their utility evaluations.

When there are such serious concerns about the validity of utility values as an indicator of preferences, one can argue that it is safer to base a measurement of negotiator performance solely on objective information. This argument applies to the measures LO, $ln(LO)$ and AVG_0. Among those measures, AVG_0 performed best in the consistent cases.

On the other hand, the extensions of the dominance relation, on which measures ELO^x and EAV^x are based, use actual, observed behavior rather than elicited utility values. If we consider actual behavior as an adequate indicator of true preferences, then this behavior should be taken into account. The results for consistent subjects also show the value of this information.

There is also a technical argument in favor of using EAV^x instead of AVG_0. As we have shown, AVG_0 models the negotiation as a zero sum game, which leaves no room for Pareto improvement. When we assume that trade-offs between the issues are sufficiently different between the two parties to make Pareto improvements possible, this effect cannot be captured using AVG_0.

Several issues raised in this paper require further research. The high number of inconsistencies in the negotiation experiments needs to be analyzed further. Maybe a more detailed analysis of characteristics of the subjects or the negotiation process could lead to an explanation of their causes. On the other hand, one needs to study whether there are differences in the outcomes of such negotiations. According to table 7, Inspire's utility values seem to indicate that there are such differences, while most of the other measures do not indicate them. A more detailed analysis of the compromise solutions actually chosen could help to resolve this puzzle.

Finally, the generalization of our empirical results remains an open question. While the empirical database used here covers a large number of negotiation experiments, it is based on only one case with a fixed decision problem, and only one NSS with a single method of utility elicitation. Different structures of negotiation problems, for example different numbers of attributes or levels within each attribute, or different methods of preference elicitation, might lead to different results.

References

DeLone, W. H., & McLean, E. R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research*, *3*(1), 60-95.

DeSanctis, G., & Gallupe, R. B. (1987). A foundation for the study of group decision support systems. *Management Science*, *33*, 589-609.

Farbey, B., Land, F. F., & Targett, D. (1995). A taxonomy of information systems appli-

cations: the benefits' evaluation ladder. *European Journal of Information Systems*, *4*, 41-50.

Jain, B., & Solomon, J. (2000). The effect of task complexity and conflict handling styles on computer-supported negotiations. *Information and Management*, *37*(4), 161-168.

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs.* New York: J. Wiley & Sons.

Kersten, G., Koeszegi, S., & Vetschera, R. (2002). The effects of culture in anonymous negotiations. Experiments in four countries. In J. Nunamaker & R. Sprague (Eds.), *Proceedings of the 35th Hawaii International Conference on Systems Sciences.* Los Alamitos: CA: IEEE Computer Society Press.

Kersten, G., & Noronha, S. (1999). WWW-based negotiation support: Design, implementation, and use. *Decision Support Systems*, *25*(2), 135-154.

Koeszegi, S., & Kersten, G. E. (2003). On-line/off-line: Joint negotiation teaching in Montreal and Vienna. *Group Decision and Negotiation*, *12*(4), 337-345.

Koeszegi, S., Vetschera, R., & Kersten, G. E. (2004). National cultural differences in the use and perception of internet-based NSS – Does high or low context matter? *International Negotiation*, *9*, in print.

Moore, D. A., Kurtzberg, T. R., Thompson, L. L., & Morris, M. W. (1999). Long and short routes to success in electronically mediated negotiations: Group affiliations and good vibrations. *Organizational Behavior and Human Decision Processes*, *77*(1), 22-43.

Pomerol, J.-C., & Barba-Romero, S. (2000). *Multicriterion decision in management: Principles and practice.* Dordrecht: Kluwer.

Schoop, M., Jertila, A., & List, T. (2003). Negoisst: A negotiation support system for electronic business-to-business negotiations in e-commerce. *Data and Knowledge Engineering*, in print.

Stewart, T. (1996). Robustness of additive value function methods in MCDM. *Journal of Multi-Criteria Decision Analysis*, *5*(4), 301-309.

Vetschera, R. (1990). Group decision and negotiation support - a methodological survey. *OR Spektrum*, *12*, 67-77.

Warshall, S. (1962). A theorem on boolean matrices. *Journal of the ACM*, *9*(1), 11-12.