Coders' reliability in the study of decision making concepts; replications in time and across topics.

I.N. Gallhofer

Introduction

Political decision making is an important topic in studies of social science. Mathematicians have developed a great variety of models concerning decision processes. Since J. von Neumann's and O. Morgenstern's "Theory of games and economic behavior" (1947) numerous efforts have been undertaken in this direction, e.g.: R.D. Luce and H. Raiffa (1957), P. Fishburn (1964), A. Rapoport (1966,1974). Policy Analysis and Cost and Benefit Analysis have also contributed to the development of this type of inquiry, e.g. G. Kuypers (1973) and the publications of the Dutch Commission for development of policy analysis (1971 etc.).
Apart from these normative approaches there also exist several empirical studies. These explore the extent to which formal theories are capable of explaining political processes, e.g. M. Leiserson (1968, 1970), A. de Swaan (1973), W.E. Saris and I.N.Gallhofer (1975) and various publications in the Journal of Conflict Resolution.
Whenever the student of decision making processes depends on documents as the main source of his data, procedures must be developed for searching for relevant concepts in texts. Since automatic search procedures are not available for this purpose, human coders must be used. Problems consequently arise concerning coding reliability. Because research is limited in this specialized field, it seemed advisable to first investigate problems of coders' reliability before proceeding with our validation of decision theory in the making of Dutch foreign policy.Satisfactory results for the reliability study would encourage further research on the rules which coders implicitly use (1). Such knowledge could facilitate the development of an automatic procedure for the search of decision making concepts in documents (2).
In the following we shall first introduce the concepts the coders used. Subsequently the research design, the coding procedure and the reliability measures will be discussed. The results are then presented and interpreted in the conclusion.

-

We thank R.J.M. Does and F.J.A. Overweel who developed the tree reliability measure and E. Kuypers, R. Nauta, B. de Valk and D. Reyne who assisted in the coding effort.

1. The selected concepts

Considering that the concepts of the normative approach (3) have proved to be useful for empirical studies (4) the theoretical framework of our study was derived from Decision Theory.

Assuming that individual (or groups of) decision makers subject the actual political situation to a thorough analysis before taking measures deemed necessary in order to achieve desirable goals, the following concepts were defined and used to describe the decision making process:

Actual state

Statements in documents concerning the immediate political situation of the decision maker, constitute the "actual state".

Evaluation of the actual state

Political decision makers often evaluate the actual state. Verbal statements which indicate the degree of desirability of the immediate situation belong to this category.

Although Decision Theory makes no use of this concept, it seemed expedient to incorporate it in our content analytic instrument.

Possible actions

a) Actions of the own party

After considering the actual state, a decision maker may examine the means which are available to him in order to obtain the desired goals. He may then review a series of possible alternative actions in such a case.

b) Actions of the other party

Choosing among available actions a decision maker must take into account the actions of the other party: the other party in persuance their goals might take measures which counteract his own. In order to exclude undesirable effects, a decision maker is therefore likely to review the available actions of the other party before selecting his own measures.

Possible new developments

Events may occur which change the entire political situation. They are neither caused by actions of the decision maker himself nor by actions of the opponent(s). Before deciding on his policies a decision maker may also take into account the likely occurences of new developments.

Possible outcomes for the own party

The choice of action(s) is based on the results that they may produce. Since not all consequences of an action are desirable a decision maker should examine the entire set of possible outcomes before selecting.

Values of the possible outcomes

Some outcomes are more desirable than others; the choice of action(s) is based on the degree of desirability of the different outcomes. A decision maker will therefore explicitly assign subjective values or utilities to the different outcomes.

Probabilities of "actions of the other party" of "outcomes" and of "new developments"

Whether "actions of the other party", "new developments" and "outcomes" occur is uncertain. "Which actions will most probably produce the desirable goals ?" To answer this question, it is necessary to estimate subjectively the probabilities of occurence of "actions of the other party", "new developments" and "outcomes".

Motivation for the selection or rejection of actions

This category consists of verbal statements which indicate the rationale for the selection or rejection of an action. The category therefore contains information concerning the criteria a decision maker uses, for instance, that he minimizes his risks. Decision Theory does not deal with this notion; it is unique to our approach.

The conceptual scheme thus defined is used for describing theoretically relevant aspects of the decision making process.

2. Research design

Texts were selected from a 1900 -1920 collection of documents in the Archives of the Dutch Council of Ministers (5). They dealt with two separate decision issues.

Syntactical parsing procedures and the content analytic classification of semantic units were taught individually to three coders for two weeks. Subsequently the coders and the author analyzed independently the documents of the first decision issue. Thereafter, the coders were split into two groups which compared together the results for each document. Seeking a common solution, the group discussed their coding disagreements and differences whenever they occured. After a period of two months the analysis and all its procedures were replicated on the same texts. Finally, the two groups came together, compared their results and sought a common solution resolving the remaining differences.

The documents of the second decision issue were coded similarly, but without replication.

We thus gathered individual and group coding results for two decision issues. The selection of two decision issues was based on the assumption that when analyzing for the first time decision issue 1, the coders lack experience and may, therefore, not produce optimal results. In case of a learning process the results of the second decision issue would thus be a more accurate representation of the possible coders' agreement (6).

In order to get an idea concerning the agreement over issue 1 when coders are experienced, the content analysis was replicated after a period of two months under the assumption that effects of memory would be minimal.

Since the pilot study had thaught us to prefer group results instead of those of the individual codings (7), the computations of inter and intra coding reliability are therefore based on the content analysis efforts of groups. However, the agreement between individual coders and the group results will also be computed.

3. The determination of semantic units of sentences and their classification

Document coding using the above mentioned concepts requires two separate procedures:

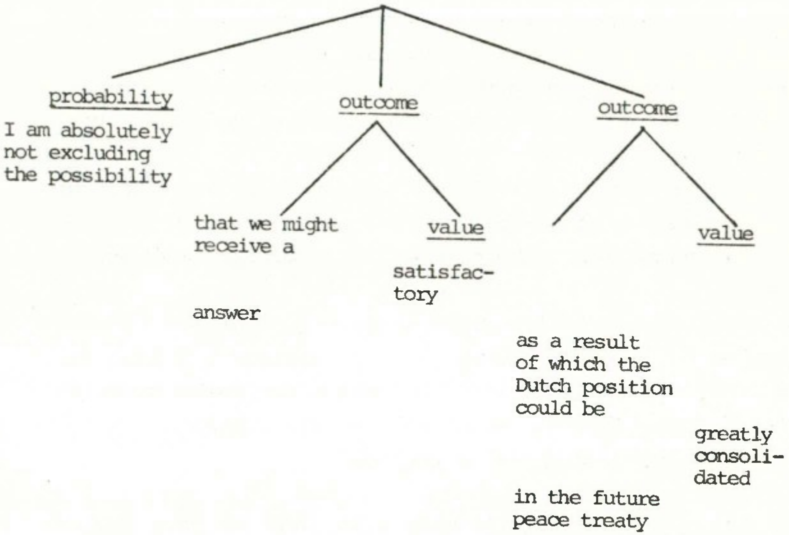
1) the decomposition of the sentences into semantic units

2) the classification of those units using one of the selected concepts.

For the first procedure we established some syntactical guidelines. To enable coders to carry out the second step, i.e. to classify, the residual category "undefined" was added to the content analytic instrument, This residual category contains all statements unrelated to the decision making process (8).

We shall use an example to illustrate the content analysis procedure, namely the sentence: "I am absolutely not excluding the possibility that we might receive a satisfactory answer as a result of which the Dutch position could be greatly consolidated in the future peace treaty." Scheme 1 shows the tree structure for one of the possible codings of this sentence. From the graph it is easy to see that this sentence is split up in three main categories:probability, outcome, outcome. Values are nested in the two outcome categories, indicated by the second level of the tree.

Scheme 1: Tree structure of a coded sentence


probability
I am absolutely not excluding the possibility
outcome
that we might receive a
value
satisfactory
answer
outcome
as a result of which the Dutch position could be
value
greatly consolidated
in the future peace treaty

The same structure can also be represented in bracket notation:
(probability: I am absolutely not excluding the possibility)
(outcome: that we might receive a (value: satisfactory) answer)
(outcome: as a result of which the Dutch position could be (value:greatly consolidated) in the future peace treaty)

Agreement measures concerning the described structural decomposition (semantic units) and the classification of these units (concepts) can be and have been computed.

Some syntactical guidelines for the search of semantic units

In order to minimize coding errors due to lack of syntactical knowledge,we developed some guidelines for the decomposition of sentences into semantic units. These rules are based on a pilot study: it required a coder to search for the above mentioned concepts in documents. Determination of the boundaries of the semantic units and the handling of embedded phrases were the main problems encountered by the coder during the pilot phase. To solve these problems, the grammatical notion of constituents was introduced (9). It defines a sentence as consisting of two components : a noun phrase (NP) and a verb phrase (VP). NP contains the subject and at times prepositional objects;VP is comprised of the verb and all kinds of objects (10). Semantic considerations lead us to determine the range of grammatical units a concept is composed of, i.e. where it starts and where it ends. This information should therefore allow one to place the unit in the tree structure. The pilot study demonstrated that in our texts the concepts coincided with the following grammatical units:

(1) a combination of more than one (NP+VP)

(2) one (NP+VP)

(3) one VP

(4) one NP

(5) a combination of words within a NP

(6) a combination of words within a VP

(7) one word within a NP

(8) one word within a VP

To determine the position of the unit in the tree structure required some additional rules. They are:

a) Semantic units consisting of one or more combinations of (NP+VP), i.e. (1),(2), immediately following each other, are classified as components of the first level. Scheme 1 illustrates this rule for the concepts "probability", "outcome","outcome".

b) Semantic units composed of a NP (4) are considered to be embedded in the unit of the VP which forms the main category. An example will illustrate this rule:

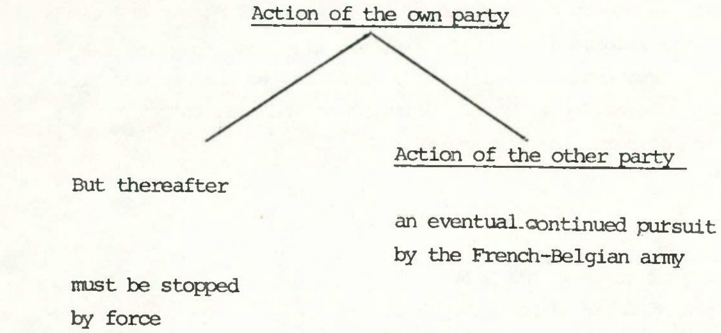
"But thereafter, an eventual continued pursuit by the French-Belgian army must be stopped by force."

The Np and Vp units are represented below in bracket notation. We also used the convention that introductory conjunctions belong to the VP.

(VP:But thereafter (Np: an eventual continued pursuit by the French-Belgian army) must be stopped by force).

Scheme 2 shows the tree structure of the classification of this sentence. By rules a), b), which classify strings of (NP+VP), resp. VP, directly succeeding each other as main components, it follows that the grammatical combinations, as mentioned above under (4) till (8), are nested in the latter. The example of scheme 1 illustrates this for the value concepts "satisfactory" and "greatly consolidated" which are both located within a VP and,therefore,are components at the second level.

Scheme 2: Tree structure of a coded sentence



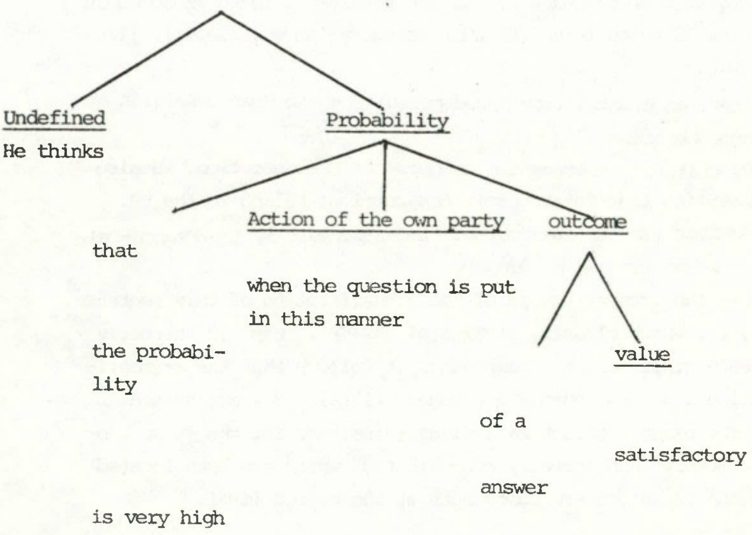
But thereafter

an eventual continued pursuit
by the French-Belgian army

must be stopped
by force

c) Semantic units consisting of one or more combinations of (NP+VP) (1), (2) or VP (3) might be nested in other units, composed of (NP+VP) or VP. This means that the units do not follow each other directly. Scheme 3 gives an illustration of this special case with some other nestings.

Scheme 3: Tree structure and classification of a sentence



He thinks

that

when the question is put
in this manner

the probability

of a

satisfactory

answer

is very high

The first concept "undefined" consists of (NP+VP):(NP:he) (VP:thinks) and is therefore placed at the first level. The probability unit immediately follows after "undefined" and contains (NP+VP): (VP: that (NP: the probability of a satisfactory answer) is very high). This concept is the second component at level 1. Within the probability unit the action phrase, composed by (NP+VP): (VP: when (NP: the question) is put in this manner), is nested. It is therefore placed at the second level of the tree. The outcome unit is also embedded in the probability segment and constitutes another branch at level 2. Since there is a value notion nested in "outcome" there exists a third level of the tree.

These guidelines enabled coders to determine the boundaries of semantic units and to handle the various nestings. The next section describes the reliability measures used for the statistical test.

4. Reliability measures

Since our coding procedure is divided into two parts, i.e.

1) the decomposition of a sentence into semantic units

2) the assignment of a concept to the unit

and since both parts can produce disagreement among and between coders, therefore, two sets of agreement measures are used.

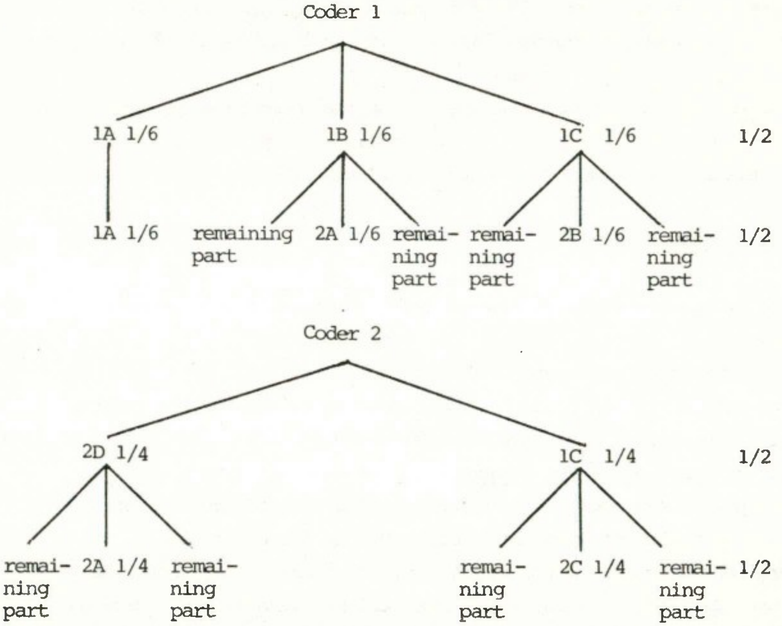
The tree structure agreement measure

When decomposing a sentence into semantic units tree structures can differ in a) the number of components and/or b) the number of levels. We therefore need a measure which is sensitive to these differences. The literature (11) largely deals with structural measures of the degree of agreement among coders combining pairs of words in the same bracket notation. Since we intend to measure the agreement between larger units, a more rigorous measure is preferable. With this in mind a measure was developed (12) which accounts equally for each tree level. An example follows of the tree structure agreement computation based on two alternative decompositions of a sentence:

Coder 1 : (I am absolutely not excluding the possibility)
(that we might receive a (satisfactory) answer)
(as a result of which the Dutch position could be (greatly consolidated) in the future peace treaty)

Coder 2: (I am absolutely not excluding the possibility that we might
receive a (satisfactory) answer)
(as a result of which the Dutch position could be (greatly)
consolidated in the future peace treaty)

Scheme 4: Representation of two alternative tree structures and computation of the agreement measure



measure = 1/2 (1/6 + 1/4 + 1/6 + 1/4) = 0.416

The bracket notation shows that the first coder discerned three main components which are indicated in scheme 4 by 1A, 1B, 1C. There is a unit 2A embedded in 1B and an other one, 2 B, in 1 C. The embeddings are placed at the second level of the tree where also the remaining part of the main categories are indicated but not categorized. Component 1A has no nestings. Since this structure consists partially of two levels a dummy level is added to 1A.

Considering the second coder's results,all units which are bracketed similarly to coder 1 receive the same label. In this case coder 2 distinguished two main categories. Because the first component is not equivalent to 1A it receives a different label (2D).

The structures in scheme 4 have the same number of levels. If structures differ in levels dummy levels are added in order to match the one(s) with the most levels. Subsequently, each level is assigned an equal partition of one (thus all levels summing to 1) which is partitioned proportionally among the units of each level. The measure is computed by adding the weights of all identically coded units and dividing by 2. In case of perfect agreement among two coders the measure yields 1 and in case of no agreement 0.

A comparison of the agreement score with the one produced by the computation of D' (13) provides an impression of the rigour of our structural measure. Because of less stringent requirements D' yields in this case a score of .85 while our measure reaches .416.

The agreement measure for the concept assignment to the units

Different concepts can be assigned to equally bracketed semantic units. To determine the extent of agreement among coders, we needed an association measure for nominal data. Among the variety of existing measures, Scotts π was selected. It is defined as the ratio of the above-chance agreement to the maximal above-chance agreement (14). If P_o is the observed proportion of agreement and P_e the expected proportion of agreement then

$$\pi = \frac{P_o - P_e}{1 - P_e}$$

If there are k different possible categories for two coders and these categories are ordered in the same way then

$$P_o = 1/n \sum_{i=1}^{k} n_{ii}$$

where n = total number of codings.

Assuming that the marginals for the two coders are identical one can use as an estimate of the marginal distribution the average frequencies in the categories and specify the proportion of expected agreement as

$$P_e = 1/n^2 \sum_{i=1}^{k} (n_{i.} + n_{.i} /2)^2 .$$

When the level of agreement equals chance expectancy, its value is zero, if perfect, it is one, and if less thancan be expected by chance, its value becomes negative.

5. Results of the study

Tables 1,2, and 3 summarize the reliability scores of decomposing sentences into semantic units. Because decision issue 1 contained a great many sentences we sampled 81 of them. For the second issue all sentences were scored. Having computed the tree structure measures for all sentences the median was used to describe the central tendency (15).

Table 1: Coding reliability of the decomposition of sentences into semantic units between individual coders and the results achieved by the two groups

median tree structure scores					
documents	coder 1	coder 2	coder 3	coder 4	total number of sentences
decision issue 1 1st coding	.92	.83	.73	.78	81
decision issue 1 2nd coding	.96	.90	.92	.85	81
decision issue 2	1.0	1.0	.92	.88	48

The scores in this table show that individual coders learned from the first to the second trial (decision issue 2). Even though during the first trial the degree of agreement was already high (.92,.83,.73,.78), it increased further when the documents of decision issue 2 were coded (1.0,1.0,.92,.88), probably the result of increased practice.
This question being resolved, table 2 contains the scores of the intra coder reliability.

In this comparison the agreement is also very high. Nevertheless, the scores are lower than those yielded for decision issue 2 in table 1. Table 1 clearly revealed that coders learned. It is therefore quite plausible that the degree of agreement, computed by comparing the results of a group at two different points in time, will be relatively low, since the group improved its efforts by producing different results.

Table 2 : Intra coder reliability between groups 1 and 2 concerning the decomposition of sentences into semantic units replicating the analysis of decision issue 1

coders	median tree structure scores	total number of sentences
group 1	.83	81
group 2	.85	81

The inter coder reliability scores are summarized in table 3.

Table 3 : Inter coder reliability between two groups of coders with respect to the decomposition of sentences into semantic units

documents	median tree structure scores	total number of sentences
decision issue 1 1st coding	.88	81
decision issue 1 2nd coding	.94	81
decision issue 2	.91	48

Table 3 shows that the inter coder reliability is already considerably high during the first content analysis of decision issue 1 (.88). During the second round, i.e. decision issue 2, the score hardly improved (.91).

Tables 4,5,6 contain the agreement measures for the classification of the semantic units by different coders. Since the number of identically decomposed units increased with each subsequent coding of the same texts , the total number of units varies from round to round. The agreement measure, however , is size invariant. Table 4 summarizes the agreement measures between individual and group coders with respect to concept assignment. It clearly indicates that the coders were learning, as the reliability scores increase from the content analysis of decision issue 1 (.81,.78,.82,.77) to that of decision issue 2 (.91,.87,.89, .87).

Table 4 : Agreement between the results of individual coders and those achieved by the two groups together concerning the assignment of concepts to semantic units

documents	coder 1		coder 2		coder 3		coder 4	
	total number of units	π	total number of units	π	total number of units	π	total number of units	π
decision issue 1 1st coding	419	.81	375	.78	315	.82	299	.77
decision issue 1 2nd coding	526	.93	453	.93	408	.92	426	.90
decision issue 2	112	.91	103	.87	96	.89	97	.87

Table 5 shows the intragroup coder reliability.

Table 5 : Intragroup coding reliability relating to concept assignment, based on the two codings of issue 1, original and replication

coders	π coefficient	total number of units
group 1	.91	440
group 2	.88	471

The size of these scores is equivalent to the coefficients derived from decision issue 2, table 4. This similarity, however, is probably the result of memory effects. The group results of the first coding were no doubt already sufficiently satisfactory so that the replication did not require much change. This contrasts sharply with the coding experience of individual analysts, who did learn from joint efforts and discussions in their respective groups.

Table 6 : Intergroup coding reliability pertaining to classification of semantic units

documents	π coefficient	total number of sentences
decision issue 1 1st coding	.84	356
decision issue 1 2nd coding	.92	489
decision issue 2	.86	98

The intergroup coding reliability is already high for the first coding of decision issue 1 (.84). The agreement score of issue 2 (.86) hardly varies from the first coding of decision issue 1.

6. Conclusions

The agreement scores of both coding procedures (i.e. decomposition and concept assignment) are very high. We can thus conclude that the suggested syntactical guidelines are realistic while the syntactical and semantic units coincide. Furthermore, different coders identified the same concepts in most instances; this leads us to believe that decision makers may also think in similar terms. In addition, the noted coincidence of semantic and syntactical units makes computerization desirable, since human coders will never equal machines in precision.
Considering the very satisfactory results of this investigation,further research is certainly warranted concerning the rules of classification

which coders implicitly use , with a view towards the development of more automatic procedures.

Notes

(1) See the contributions in this reader of Z. Namenwirth a.o.

(2) This semantic knowledge could subsequently be combined with Boot's parsing procedure in order to extract the concepts.

(3) These concepts are described, for instance, in Fishburn (1964), pp.21, or in Leergang besliskunde (1971), vol.5, p.11.

(4) E.g.:Wendt a.o. (1975), Siegel a.o. (1964), Saris and Saris (1975)

(5) Algemeen Rijksarchief, 's-Gravenhage, dossier RA 2^{e}, Archief Ministerraad, 3 october 1914 and Bijlagen tot de Notulen van de Ministerraad, april 1916, Ontwerp vermindering troepenmacht.

(6) Since the documents of the two decision issues were forthcoming from the same department we assumed that they were comparable with respect to their textual characteristics.

(7) While coding individually some aspects were neglected which emerged as evident from group discussion.

(8) In our documents, this category occurs at times in descriptions of the rationale of past actions by the other party or by considerations of Dutch law and treaties.

(9) E.g.: Booij a.o. (1975), p.84

(10) E.g.: de Haan a.o. (1974), pp.30

(11) E.g.:Boorman a.o. in Shephard a.o. (1972) vol.1, pp.233

(12) Drs. R.J.M. Does and drs. F.J.A. Overweel of the Mathematical Centre of the University of Amsterdam developed this measure.

(13) Boorman, a.o. in Shephard a.o. (1972), vol.1, p.235

(14) Krippendorff, in Borgatta a.o. (1970) p.144

(15) Since the distribution of the data is skewed the median is to be prefered above the mean.

References

Beleidsanalyse, driemaandelijks bericht van de commissie voor de ontwikkeling van beleidsanalyse, 1971 etc., Staatsuitgeverij, 's-Gravenhage

C.E. Booij, J.G. Kerstens, H.J. Verkuyl, Lexicon van de taalwetenschap, Het Spectrum, Utrecht/Antwerpen, 1975

S.A. Boorman, P. Arabie, Structural measures and the method of sorting, in Shephard a.o., Multidimensional scaling theory, vol.1, Seminar Press, New York, 1972

P.C. Fishburn, Decision and value theory, Publications in Operations research, nr.10, J.Wiley, New York, 1964

G.J. de Haan, G.A.T. Koefoed, A.L. des Tombe, Basiskursus algemene taalwetenschap, van Gorcum, Assen, 1974

J. Kriens, G. de Leve, Leergang besliskunde, Inleiding tot de mathematische besliskunde, MC syllabus 1.5, Mathematisch Centrum, Amsterdam, 1971

K. Krippendorff, Bivariate agreement coefficients for reliability of data, in Sociological Methodology, E.F. Borgatta, G.W. Bohrnstedt eds., Jossey Bass, San Francisco, 1970

G. Kuypers, Grondbegrippen van politiek, Het Spectrum, Utrecht/Antwerpen, 1973

M. Leiserson, The study of coalition behavior, theoretical perspectives and cases from four continents, New York, Holt, Rinehart, 1970

R.D. Luce, H. Raiffa, Games and decisions, J. Wiley, New York, 1957

J. v. Neumann, O. Morgenstern, Theory of games and economic behavior, Princeton, University Press, New Jersey, 1947

A. Rapoport, Two-person game theory, Ann Arbor, University of Michigan Press, 1966

A. Rapoport, N-person game theory, Ann Arbor, University of Michigan Press, 1970

A. Rapoport, Game theory as a theory of conflict resolution, D. Reidel, Dordrecht-Holland:Boston-U.S.A., 1974

W.E. Saris, I.N. Gallhofer, L'application d'un modèle de décision à des données historiques, in Revue Française de Science Politique, vol XXV, nr. 3, june 1975, pp.473

A. de Swaan, Coalition theories and cabinet formations, a study to formal theories of coalition formation, applied to nine european parliaments after 1918, dissertation, Amsterdam, 1973

S. Siegel, A. Siegel, J. McMichael Andrews, Choice, strategy and utility, McGraw-Hill, 1964

D. Wendt, Ch. Vlek, Utility, probability and human decision making, D. Reidel, Dordrecht, 1975