Studies in Educational Evaluation. Vol. 16, pp. 529–550, 1990
Printed in Great Britain. All rights reserved.

0191–491X/90 $0.00 + .50

RELIABILITY-TESTING OF TWO ANALYSIS INSTRUMENTS FOR DECISION-MAKING IN CURRICULUM CONFERENCES

Martin Mulder and Jacqueline te Brake

University of Twente, Department of Education, Enschede, The Netherlands

An important condition for analysis of the deliberation process during curriculum conferences is the availability of reliable instruments. In this study, carried out at the Department of Education of the University of Twente in The Netherlands, the design and reliability-testing of two of these instruments is reported. The instrument for argumentation analysis (MARS) is reliable for two components: marking opinions, and marking arguments. Specifying the skeleton of the argumentation structure is an unreliable component of the instrument. The instrument for process analysis (CODOM) is reliable for three components: sender (of a message), message/interruption and nature of the message. The 'message/receiver' is an unreliable component of the instrument. These results can be used for analysis of the deliberation process during curriculum conferences and for the improvement of the quality of the practice of curriculum conferences.

1. Introduction and Theoretical Framework

Decision-making in curriculum groups is a complex problem in curriculum development, as it is the process in which various stakeholders, who in most cases hold conflicting philosophies of education and different views on society, subject matter and the learner, collectively approve certain characteristics of a series of planned educational experiences. The intercontextual diversity of the division of responsibilities for curriculum development among, for instance, teachers, curriculum supervisors, subject matter experts, publishers, curriculum bodies, committees and institutes (Van Bruggen, 1987, p. 232), or a mix of these persons and organizations, even increases the complexity of this problem. Furthermore, within existing differentiated curriculum development systems and practices, several strategies are being used to orchestrate decision-making processes to accomplish a common basis for a new or revised curriculum.

In a series of recent studies, a relatively new approach of curriculum development was studied, in which the multi-dimensionality of curriculum decision-making is taken into account, emphasizing its argumentative and interactive components. This approach is called the curriculum conference, and can be conceived of as a set of coherent heuristics or a strategy to reach consensus on certain curriculum design specifications in heterogeneous nominal curriculum groups. Argumentation and interactivity of curriculum decision-making processes are emphasized in this approach, because precise justification of curriculum decisions on sound theories of education on the one hand, and careful open dialogue on competing or controversial curriculum proposals on the other hand, are seen as necessary conditions for valid curriculum decision-making in pluralistic western democracies. Space limits elaboration of these ideas, but in short we can say that a curriculum conference is a workshop-approach for designing curricula in groups (Frey, 1982). It is a pre-structured group-oriented legitimizing deliberation in which goals, content and instructional strategies are discussed and approved (Mulder, 1988). The term curriculum is conceived of as a broad concept for varying documents (Beauchamp, 1981). A curriculum conference can be used for instance for designing a course syllabus, course materials or a course scheme. The organizers of a curriculum conference prepare the sessions by writing a kind of consultation document. In our studies the information in this document was the result of curriculum analyses, alumni surveys, teacher surveys, employer demands and the like, by which the participants could get a full and common understanding of the problem and its context. The results of a preparatory empirical study are analyzed and evaluated during a deliberation process in a curriculum conference. This deliberation process (in this paper called the decision-making process) takes place under certain conditions and is aimed at reaching decisions on practical curriculum problems. An example of an issue in curriculum decision-making is the question whether the curriculum should provide students with specific competencies for their future position in the world of work, or with general knowledge and skills that are applicable in a variety of problem situations (Wilms, 1984). Discussions about such issues can be the core of deliberation during curriculum conferences. The criteria used in analyzing and evaluating the preparatory research report are based on the participants' personal experiences with, and insights into, the curriculum domain in question. Approval of the curricular consequences of the conclusions of the decisions, in other words the group decisions about the curriculum design specifications, is conceived of as curriculum legitimation. This concept, used frequently in German curriculum literature (Hameyer, 1983) is not very common in English curriculum literature, but it stands for the process of justification (Johnson, 1967) of curriculum design decisions. See for further details on decision-making processes in curriculum conferences in which argumentation and interaction are major characteristics Mulder & Nijhof (1987), Mulder, Nijhof & Remmers (1987) and Mulder & Thijsen (1989).

In several inquiries by the division of Curriculum Technology and Educational Administration of the Department of Education of the University of Twente, characteristics of curriculum conferences were studied (Nijhof & Mulder, 1986; Mulder & van Lent, 1988; Pieters & Mulder, 1989). The implemented curriculum conferences were indeed evaluated, but no clear insight into the decision-making process resulted from these studies, because these evaluations were not primarily focused on the decision-making process. As this decision-making process serves as the basis for legitimation of the curriculum design specifications, and it has to meet certain quality standards (Frey, Malliou, Langeheine & Horton-Kruger, 1988) it appeared necessary to look into the decision-making process separately. The reason for this interest in curriculum decision-making is the development of effective heuristics for this perennial problem in curriculum practice. A preliminary exploratory inquiry into this process analysis resulted in prototypes of two analysis-instruments in which reliability was tested. The results of this testing are described in this contribution.

The decision-making process in curriculum conferences can be displayed with Poole's Multiple Sequence Model of Decision Development (MSMDD): "Rather than picturing group decision as a series of phasic 'blocks' dropped one after another into sequence, it portrays development as a set of parallel strands or tracks of activity which evolve simultaneously and interlock in different patterns of time" (Poole, 1983). The several kinds of group activities form the core of the MSMDD: *task activities*, that serve the performance of the group task, *relational activities*, that manage the relations between the group members and *content activities* that reflect the substantial issues or arguments. For the way in which these three kinds of activities can be separated from each other, Poole refers to Bales' "Interaction Process Analysis Coding System" (Poole, 1981:7) for differentiating the task and relational activities, and Fisher's "Decision Proposal System" (Poole, 1981:7) for differentiating the content activities. In the next paragraph a description is given of the elaboration of Poole's MSMDD in this inquiry.

To summarize the important elements of the theoretical framework of this study, we can say that argumentation and interaction are assumed to be important characteristics of the curriculum decision-making process at a curriculum conference because of its legitimizing function. Argumentation and interaction are analyzed to create effective heuristics for curriculum decision-making. This inquiry was aimed at testing prototypes of instruments for analyzing argumentation and interaction.

1.1 Tracing and analyzing argumentation

The argumentation structure of a decision-making process can be made visible if the argumentation is reconstructed systematically, making clear what topics are discussed, what opinions and arguments are formulated and what arguments have been decisive for the specifications of the curriculum design. We conceive of argumentation with reference to Keough (1986: 19), as follows: argumentation in a text takes place if two criteria are met; (i) if there is a linguistically explainable opinion and (ii) if there are one or more overtly stated, clear arguments. When we try to detect argumentation defined in this way, in order to grasp the meaning of the verbal interaction during the curriculum conference, we run into serious difficulties, because it is not always obvious when argumentation takes place. Because it is not always immediately clear whether both criteria for argumentation are met, instrumental guidelines have been proposed to identify argumentation, such as argumentative indicators and opinion makers (van Eemeren, Grootendorst & Meuffels, 1984).

When argumentation is identified, the next step in the reconstruction of the argumentation structure is to divide the localized argumentation into meaningful components. Smith & Meux (1970) used the concept of episode for this structuring. An episode is defined as unity in a conversation or discussion, starting with a proposition that evokes a verbal exchange about a topic and is terminated by stopping the exchange about that topic. As such, an episode is a meaningful fragment of verbal interaction. The episode consists of an opening, a continuation and a termination stage; the termination can take place in different forms: a closing remark and a (modified) conclusion. It is also possible that there occurs an abrupt shift from one episode to another without a meaningful termination of the first episode. When the argumentation is structured in episodes, these episodes can be displayed and structured. Naess (1978) has proposed the possibility of reconstructing the argumentation per episode systematically, which is elaborated by Kopperschmidt (1985) in his "macrostructural argumentation analysis". A rather precisely-formulated opinion or proposition (coded as *FO* - the italics correspond with the categories in Figure 1) to which the arguments refer, forms the starting point of the analysis. The opinion that is stated by a participant of a curriculum conference faces advocates and opponents, expressed by formulating arguments (revealing *values* relating to the opinion). The arguments are connected to *FO* in such a way that it is apparent what relationship exists between each argument and *FO* on the one hand, and the other arguments on the other hand (this refers to the *direction* of the argument, either pro or anti). If an episode is terminated by a modified conclusion, the arguments that have been decisive are indicated. In this reconstruction of the argumentation of an episode 'other messages' (referred to as *other*) are also coded that are not argumentative, but that do influence the flow of argumentation.

An even deeper insight into the decision-making process can be reached by carrying out yet a further analysis of the reconstructed argumentation that consists of labelling the argumentation structure (*arg. structure*) in the message of a group member (van Eemeren & Grootendorst, 1982), labelling the expressed opinions (*content argumentation*) and arguments themselves (*arg. type*) (Toulmin, 1958; Putnam & Geist, 1985; Propper, 1987; Schellens, 1985), noticing reserves (*reserve*) or limitations (*convince*) in an argument (van Eemeren, Grootendorst & Kruiger, 1984; Putnam & Geist, 1985) and localizing the position of the opinion (*opinion position*) in the argumentation (van Eemeren & Grootendorst, 1982). At this level of analysis we deal with verbal expressions of different participants, that relate to one opening opinion or proposition and one closing proposition or conclusion. After study of the literature and some explorations of empirical material existing of protocols of the curriculum conferences, an instrument has been developed for coding argumentation during curriculum conferences on a specific form (see Figure 1). The instrument is called MARS (Macro Argumentation Structure).

TEXTNUMBER :
FO (content):
type:
sender:

RATERS :

DATE :

cate-gory	sen-der	***reaction***		***argumentation***					***rest***	***termi-nating remark***	***first***	***(modified) conclusion***				
propo-sition		direc-tion	value	arg. struc-ture	arg. type	con-vince	posi-tion opinion	con-tent arg.			yes (1)		no type	yes/no		
											sender	prop.		sender	prop.	type
1																
2																
3																
4																
5																
6																
7																
8																
9																
10																
11																
12																
13																
14																
15																
16																
17																
18																
19																
20																
21																
22																
23																
24																
25																

[1] Yes, no and yes/no here means: the (modified) conclusion is based on previous propositions (yes), or not (no) or on previous propositions and not on previous propositions (yes/no). Under these three different conditions different categories had to be rated.

Figure 1: Rating form belonging to the argumentation analysis instrument MARS (for all components of the lower part of the form precoded alternatives existed).

1.2 Coding Verbal Interaction

As stated earlier, an open dialogue between the participants of curriculum conferences is assumed to be an important characteristic of curriculum decision-making. Two well-known phenomena, namely coalition-forming and dominance, are conceived of as menaces to open deliberation.

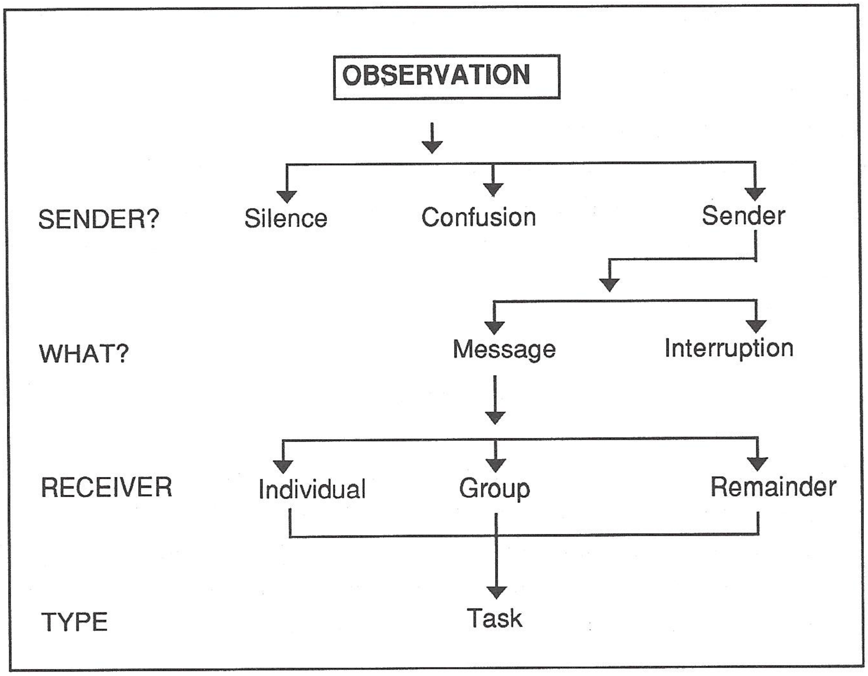

Figure 2: Composition of the categories of the observation instrument CODOM (for all components of the instrument precoded alternatives existed).

To trace the coalition-forming and dominant group members during the decision-making process, the quantity as well as the quality of the group interaction has been analyzed (Shaw, 1981). The quantity of interaction can be used to determine the communication pattern (Schmuck, 1977) and to localize group members who interact with or speak to one another relatively frequently (sender/receiver). It is possible to determine whether participants communicate in a positive manner or whether there is a certain amount of coalition forming, from information about the quality of interaction. Dominant participants initiate (often negative) and receive relatively much (often positive) communication. Extra indications that can be used are: response time, interrupting others (*what*), length of eye contact, voice volume and tone. The execution of the task (*type*) and relational activities performed by one participant can possibly also be considered a dominance aspect. After research of the literature and some explorations of the empirical material which exists from video registrations of curriculum conferences, an analysis instrument has been developed for tracing COalition forming and DOMinance (CODOM) of participants (see Figure 2 for the composition of the categories).

Thus in this paper we shall concentrate on the instrumentation of argumentation and interaction analysis of curriculum decision-making. Two prototypes of instruments were developed and tested for their reliability. In this paper we shall report the design (section 2) and the results (section 3) of this reliability testing. In section 4 we shall present some tentative conclusions.

2. Research Design

MARS is tested in varying independent pairs by four raters, and CODOM is tested by three independent observers. None of these raters and observers was informed about the characteristics of decision-making processes during curriculum conferences and the specific research questions of this inquiry. All persons were undergraduate research assistants of the Department of Education.

One can distinguish different types of reliability, particularly with this type of data:

- retest reliability; this is reliability of behavior of people (do they behave similarly across time, in a similar situation?);
- inter-observer reliability; this is the correspondence of the judgement of raters of the same material;
- internal consistency; do the different items of the instruments measure the same?

In our study only inter-observer or inter-rater reliability is at stake. Within this type of reliability, one can also distinguish several approaches regarding the object of reliability testing:

- inter-rater or observer reliability with regard to the definition of the unit of analysis (opinions and arguments for instance);
- inter-rater or observer reliability with regard to the rating of these units (coding senders, arguments and so on).

We shall give a very brief description of the tasks of the raters and observers now, to show the types of reliability testing that were carried out in this study. Later on in this paper we shall describe the reliability testing procedure more extensively.

During the argumentation analysis of unstructured episodes raters had to differentiate argumentation and information within the episodes, as well as having to mark opinions and arguments within argumentative verbal interaction, and the content of the opening proposition. In the argumentation analysis of both unstructured and structured episodes raters had to code the type and the sender of the opening proposition, as well as having to code the skeleton of the argumentation and to label arguments and opinions. During the observation analysis observers had to code certain categories or interactions between the participants of the curriculum conference.

So we can say that the first part of the argumentation analysis of the unstructured episodes was tested for inter-rater reliability with regard to the definition of the unit of analysis (in this case opinions, arguments, and the opening proposition). And the common part of the argumentation analysis of the unstructured and structured episodes was tested for inter-rater reliability with regard to the coding of the opening proposition, opinions, arguments, and the skeleton of arguments. During the observation analysis only the latter type of inter-observer reliability was tested (with regard to the coding of the interaction categories).

2.1 Material

From existing material (typed protocols and video-registrations) of one curriculum conference (Mulder, 1988) that lasted twelve hours, a stratified sample was drawn to determine the research material. The strata consisted of the three different terminations of an episode; without termination, with terminating remark, or with (modified) conclusion. For the raters a typed protocol of seventy-five minutes of the decision-making process was available, in which the episodes were marked by the researchers, and for the observers the corresponding video registrations were selected.

We should point out here that the material that was available was necessarily limited. Although registrations of three curriculum conferences were made, for the purpose of this inquiry registrations from only one curriculum conference were used. So we have to note that the material that has been selected is from one single group of participants and from a single subject for a specific subsystem of education. There was a practical reason for this choice: not all video-registrations were typed out literally at the moment of the study. Further analyses will be done that include this other material as well. At the time however, we only used a selection of the verbal interactions during one curriculum conference, which limits the generalization of the results of this study. But for the formative use of the results, which is the further development of analysis instruments for decision- making during curriculum conferences, the design was useful.

2.2 Procedure

For testing the reliability of the argumentation analysis with MARS, the research material was divided into unstructured and structured episodes. In the structured episodes the research indicated (i) the types of messages that were intended for analysis and (ii) the opinions and arguments within the argumentative messages. Unstructured material was the literal verbal interaction between curriculum conference participants in which the messages were not prepared by the researchers in such a way that it would be clear to the reader of the protocol what information is argumentative and what messages can be interpreted as opinions and arguments. The distinction between unstructured and structured episodes was made because the training showed differences regarding the analysis and interpretation stage and the reliability testing of the coding stage is dependent on the results of these stages.

During the reliability testing four raters coded the research material (parts of the typed literal protocol) in pairs during three sessions (A, B, and C) of four hours. The pairs of raters varied per session. In every session two unstructured and two structured episodes were analyzed. This means that six unstructured and six structured episodes were analyzed by two varying pairs of raters.

The analysis of the unstructured episodes was broken down into three stages: (i) reading the episode and consequently analyzing the types of messages (for instance information and argumentation); (ii) interpreting each argumentative message within the episode by marking arguments and opinions; (iii) coding the different opinions, arguments, and argumentation structure. Both the first stages were carried out individually within each pair. After both raters were ready with Stage Two, they discussed and decided about the analysis and interpretation of the episodes. The results of Stage One (reading the episode and consequently analyzing the types of messages) and Two (interpreting each argumentative message within the episode by marking arguments and opinions) were registered in the typed protocols of the selected sections of the curriculum conference. The coding of the argumentation structure, arguments and opinions in Stage Three were carried out together. The results of Stage Three (coding the different opinions, arguments, and argumentation structure) were registered on a coding form (see Figure 1).

For the structured episodes the raters had to carry out coding of the argumentations, structure, opinions and arguments only; the researchers had carried out the analysis stage and the interpretation of the argumentative message already. The procedure for coding the structured material corresponded with Stage Three of analyzing and coding unstructured material. The raters had at their disposal a manual in which the instrument (MARS) was described, as well as the report of the preparatory study, used by the participants in the curriculum conference.

The interaction analysis was carried out by three observers in three group sessions of about one hour. Each observer coded twenty five minutes of the video registrations individually per session. After every ten minutes there was a pause of five minutes. The scores were entered every ten seconds by a hand computer (Canon X-07) and a dedicated computer programmer. Apart from that, the observers had at their disposal a manual in which the instrument was described and a group picture of the participants with their codes.

2.3 Training

Training of the research assistants was provided before analyzing, observing and coding began. The rater's training took approximately twenty five hours and consisted of theory and practice components. The raters learned to go through states: (i) reading and analyzing types of messages (ii) interpreting argumentative messages by marking opinion(s) and argument(s) and (iii) coding. An important result of the training was that it was clear that the raters had problems with Stage One and Two of the argumentation analysis (analyzing types of messages and marking opinions and arguments in argumentative messages). The inter-rater reliability (expressed in the correspondence of analysis and marking) for these stages was very low. Therefore, a distinction was made between structured (= prepared) and unstructured (= authentic) episodes during the reliability testing. In structured episodes the researchers had already performed the analysis and interpretation, which implied that the raters had only to go through the coding stage. For the unstructured episodes the raters had to go through all stages (analysis, interpretation and coding) themselves.

The observers training took approximately twenty four hours. The observers learned to observe and code video registrations simultaneously.

In both training sessions, practice materials were used that had not been used during the reliability testing.

2.4 Reliability Testing

The reliability of the two prototype analysis instruments was tested by Cohen's Kappa-coefficient (van der Sijde, 1987). As a criterion for an acceptable reliability a Kappa of .50 was chosen. This corresponds with Karlin's (1980) study on the "Decision Proposal & Modification Coding" (DMP), which is an elaboration of Fischer's model in which an analysis and interpretation stage can also be distinguished. During the execution of the stages mentioned, Karlin learned that a dependent relationship existed between these stages and some categories in his analysis instrument, which implied that a Kappa value of .80 was unattainable. The same principle holds for MARS as for Karlin's DPM. The same reliability criterion is chosen for CODOM.

In those cases where analysts had to mark parts of the text as being opinions, arguments or opening propositions (where reliability regarding the definition of the unit of analysis is at stake), and no pre-coded alternatives of categories of analysis existed, a percentage of agreement was used as a quantitative indicator of correspondence between the analysts.

3. Results

First, the results of the reliability testing of MARS are presented and, secondly, those of CODOM. Initially, the results of the rating of the unstructured and structured episodes are presented separately for MARS, because of the difference in stages that have to be performed by the raters (see for a depiction of the results of the reliability testing of MARS Table 1).

Table 1: Data for the reliability testing of MARS for unstructured and structured episodes.

Stages	Unstructured episodes		Structured episodes
1. Analysis			
differentiating		kappa	
argumentation and information	.21		-
2. Interpretation			
		%	
marking opinions (total of 40)	.506		-
marking arguments (total of 70)	.526		-
3. Coding			
		%	
a. Opening proposition			
content marking FO	.667		-
		kappa	
type FO	.54		.38
sender FO	1.0		1.0
b. Skeleton	.45		.54
labeling argument	.83		.79
labeling opinion	.97		1.0
c. number of traced conclusions	2[1)]		1[1)]
correspondence marking			
belonging arguments	zero		unanimous

1) These figures are no Kappa values but absolute numbers

3.1 MARS: Unstructured Episodes

(1) The analysis stage, in which several types of messages are distinguished, shows a Kappa-value of .21 between both groups of raters.

(2) Whether or not those marks had been placed in the same sentences by both groups was traced at the interpretation, namely in which opinion(s) and argument(s) are marked in an argumentative message by the raters. This appeared to be so in the case of 50.6% of all instances (n=40) while for arguments this percentage was 52.6% (n=70). (Percentages are used here to indicate the level of correspondence; reliability testing with Kappa is not appropriate here).

(3) The coding stage is divided into: (3a) starting proposition (FO), (3b) argumentation (into which the category 'reaction' is inserted), and (3c) conclusions. The category 'senders' shows perfect correspondence over all episodes. The categories 'rest' and 'terminating remark' have not been analyzed further, because they comprise only one subcategory.

(3a) The raters marked FO (n=6) in 66.7% of the cases as well as the argumentation. (Percentages are used here to indicate the level of correspondence; reliability testing with Kappa is not appropriate here). The correspondence between type and sender of FO showed Kappa-values of .54 and 1.0.

(3b) Because the argumentation section consists of relatively many categories, three higher-level categories are used to present the results: 'skeleton of the argumentation structure' (consisting of 'direction', 'value' and 'argumentation structure'), 'labelling of the argument' (consisting of 'types of arguments', 'level of conviction' and 'level of reserve') and 'labelling of opinions' (consisting of 'location of opinion' and 'content of argumentation'). The skeleton of the argumentation structure shows a Kappa of .45. The category 'value' especially shows a below-average score between the group (Kappa = .16). Labelling the argument and opinion shows a Kappa value of .83 and .97.

(3c) The two groups of raters traced two conclusions. Of the seven propositions on which the conclusions are based, five were interpreted as arguments and two as 'miscellaneous'. The marking of the five arguments showed no correspondence between the groups.

3.2 MARS: Structured Episodes

The analysis and interpretation of the material to be coded from the structured episodes was performed by the researchers. They also marked the content of FO.

(3a) The Kappa values for correspondence between the raters on the type and the sender of FO (n=6) is .38 and 1.0.

(3b) The Kappa value on the skeleton of the argumentation structure is .54. Again the category value shows a below-average correspondence (Kappa = .14).

(3c) The raters have traced one conclusion. The propositions and types of arguments showed perfect correspondence.

3.3 MARS: Comparison of the Results of Unstructured and Structured Episodes

After determining the Kappa values, the reliability intervals were computed. Too little material for both conditions was left because the research material was divided into unstructured and structured episodes. The reliability intervals therefore proved to be too wide. The consequence of this was that the conclusions regarding MARS have to be interpreted with some restraint. Furthermore, the conclusions only refer to the argumentation part. The numbers on which the results for the starting proposition and the (modified) conclusions are based were too small (n < 7).

A comparison of the results of the three higher-level categories for argumentation of the unstructured and structured episodes show that the structured episodes resulted in relatively higher correspondence on the 'skeleton of the argumentation structure' and 'labelling the opinion'.

The gain in interreliability on the category skeleton is attributed to the greater correspondence between the groups of raters on the category argumentation structure (Kappa = .66 for the structured and .34 for the unstructured episodes). Kappa values for labelling the argument and opinions are .79 and 1.0.

The unstructured episodes on the other hand show higher correspondence on 'labelling the arguments'. The three higher-level categories show a rank order (see Figure 3). The criterion for acceptable reliability is reached when structured episodes are used. Applying MARS on unstructured episodes, however, resulted in too little inter-rater correspondence for the 'skeleton of the argumentation structure'.

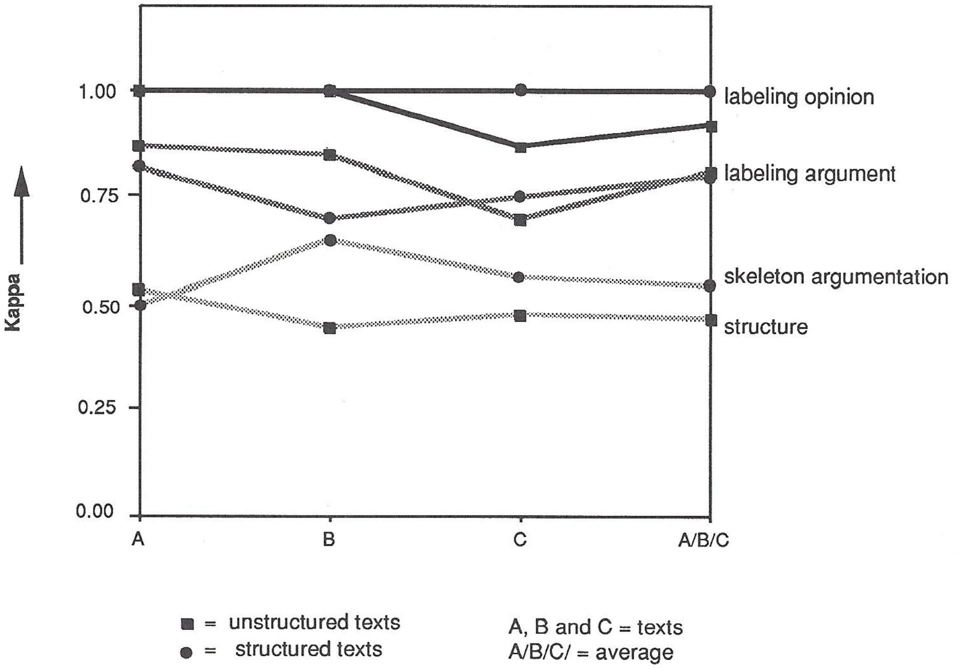

Figure 3: Total correspondence between raters about combined categories of MARS for unstructured and structured episodes.

The expectations of Karlin (1980) were confirmed here: the correspondence between raters using categories in which analysis and interpretation play a part ('skeleton of the argumentation structure') appears to be smaller than when using categories in which analysis and interpretation are not necessary ('labelling the argument' and 'opinion'). The differences between the coding stages regarding the two kinds of episodes are very small and therefore have to be interpreted carefully.

3.4 CODOM

The results of the observation analysis with CODOM regarding the average correspondence of the category sender show a Kappa value of .76. For the combined categories 'interruption' and 'message' the average Kappa is 1.0. The combined categories 'message' and 'receiver' show Kappa values of .38, the category 'type of the message' and 'receiver' shows a Kappa value of .38 and the category 'type of the message' a Kappa value of .52. Three of the four categories of CODOM have thus met the criterion of Kappa = .50. Only the categories 'message' and 'receiver' remained below the stated criterion.

4. Conclusions

During curriculum design processes in which curriculum design specifications have to be justified (or legitimized), careful decision-making is widely accepted to be important when certain criteria are to be met (Frey, Malliou, Langeheine & Horton-Kruger, 1988). But analyzing this decision-making with respect to two of these criteria, argumentation and interaction, appeared to be very difficult.

The intention of this paper however, was not to present the results and conclusions of the argumentation and interaction analysis of the curriculum conference as such. The main issue here was whether the instruments for argumentation and interaction analysis would result in reliable data. We can conclude now that this was partly the case. To be more specific, the results of this inquiry show that:

1. regarding argumentation analysis, application of the prototype of the argumentation analysis instrument resulted in satisfying inter-rater reliability (with respect to the definition of the unit of analysis) for:

   - marking opinions;
   - marking arguments;
   - marking the content of the opening proposition and satisfying inter-rater reliability (with respect to the coding of the unit of analysis) for unstructured episodes for:
   - coding the type of opening proposition;
   - coding the sender of the opening proposition;
   - labelling arguments;
   - labelling opinions;

   and satisfying inter-rater reliability (with respect to the coding of the unit of analysis) for unstructured episodes for:
   - coding the sender of the opening proposition;
   - coding the skeleton of the argumentation;
   - labelling arguments;
   - labelling opinions;

2. regarding interaction analysis, application of the prototype of the observation analysis instrument resulted in satisfying inter-rater reliability (with respect to the coding of the unit of analysis) for:

   - sender;
   - interruption/message
   - type of message.

All other elements of the instruments failed to result in reliable data as far as inter-rater/observer reliability is concerned.

Furthermore, we can conclude that argumentation and interaction are characteristics of deliberations during curriculum conferences that are very hard to operationalize and to analyze. Even a relatively intensive training did not result in satisfying reliability of all elements of the research instruments. Further analyses are necessary to detect the sources of unreliability. As yet we presume that regarding argumentation analysis the raters are not fully capable of analyzing verbal interaction when they lack:

- contextual information of the curriculum problems discussed;
- an historical description of the curriculum issues in the education subsystem that is being discussed;
- information about the present curriculum issues in this case.

This kind of information could be inserted in the training for raters. In our study this was not the case, as training consisted mainly of teaching undergraduate research-assistants the meaning of the elements of the argumentation analysis instrument. Furthermore, in our study the number of episodes was restricted. If more episodes of more situations are used, the design of reliability testing would be more powerful.

Regarding interaction analysis we presume that the receiver, as a component of the verbal interaction model, could be eliminated, because speakers also try to influence the whole group, even when they are speaking to only one participant.

Regarding the reliability of the 'skeleton of the argumentation structure', the problem is that the raters have to interpret the intended meaning of the verbal information in the protocols. As oral language is often incomplete, subjectivity is difficult to eliminate from interpretation processes. The unstructured episodes therefore showed only a small inter-rater reliability. Perhaps other procedures should be tested to determine the reliability of the interpretative activities of raters. It may be possible for participants of curriculum conferences themselves to analyze and interpret their own propositions regarding the skeleton of the argumentation structure. Protocols and video registrations may be used for stimulated recall. Such data could be compared with interpretations of raters of typed protocols. Such an inquiry could lead to more knowledge about the thinking of curriculum developers. This thinking is little known at the moment, but it may be an important condition of the effectiveness of curriculum decision-making. It also may reveal knowledge about the attunement of interventions of chairpersons of curriculum conferences to the thinking-styles of participants.

If we interpret the results carefully against the background of decision-making processes in curriculum conferences, some questions arise regarding the verbal interaction of the participants. Are the participants able to interpret each other's propositions? Is the chairperson able to trace the argumentation process precisely and perform appropriate interventions according to this process? Are the conclusions drawn from the discussion reflections of the arguments that have been put forward? Are contra-arguments discussed profoundly enough? Is every participant adequately able to infer the direction of others' messages? These fundamental questions for curriculum conferences cannot be answered until the research instruments tested in this inquiry are optimalized further and applied in studies with more research material, possibly supplemented by other research instruments.

Finally, a systematic problem in analyzing deliberation in curriculum development processes has to be mentioned. Argumentation analysis requires that participants in the decision-making process formulate opinions and arguments. However certain opinions and arguments are not communicated overtly during the sessions for a number of different reasons. One of these reasons can be that the participants share common opinions and arguments that shape the discussions. Such underlying opinions and arguments remain covert for the analysts, although they are very important to understand the meaning of the decision-making processes. Another reason can be that some opinions and arguments are not articulated enough to bring them into the discussions. They remain in the intuitive realm. Also these intuitions can shape the interaction processes without becoming overt in the verbal interaction. Still another reason can be that participants know that certain opinions and arguments are not powerful enough in the discussions, because other participants have stated more convincing opinions and arguments. In such cases participants may withhold such arguments and opinions from the discussions although they can influence the individual's own attitude to the decisions on the curriculum design specifications. Finally, the different communicative competence of participants, may stimulate some of them to engage in and others to refrain from the discussions more and more. We assume however, that these phenomena are more often the case in curriculum committees with a more continuous character and more stable roles of the committee members than in curriculum conferences in which participants meet only once. Nevertheless, the factors just indicated, may define the limits of decision-making in curriculum conferences. And they may conceal important reasons for decision-making, which causes severe validity problems in the design of our study on the reliability testing of analysis instruments for decision-making in curriculum conferences. Further study is necessary to clarify these suppositions.

References

Beauchamp, G.A. (1981). *Curriculum theory* (4th Ed.). Itasca, Illinois: Peacock.

Eemeren, F.H. van and Grootendorst,R. (1982). *Regeln fuer eine vernuenftige Diskussion; ein Beitrag zur theoretischen Analyse der Argumentation zur Loesung der Differenzen.* Dordrecht: Floris.

Eemeren, F.H. van, Grootendorst, R. and Kruiger, T. (1984). *Argumenteren.* Groningen: Wolters-Noordhoff.

Eemeren, F.H. van, Grootendorst, R. and Meuffels, B. (1984). Het identificeren van enkelvoudige argumentatie. *Tijdschrift voor Taalbeheersing*, 6, 297-310.

Frey, K. (1982). *Curriculum conference*: an approach for curriculum development in groups. Kiel: Institute for Science Education.

Frey, K., Malliou, K., Langeheine, R. and Horton-Krueger, G. (1988). *Studies of the quality of the curricular process in the curriculum conference.* Zuerich/Kiel: Institut fur Verhaltenswissenschaft/Institut fur Paedagogik der Naturwissenschaften an der Universitaet Kiel.

Hameyer, U. (1983). Systematisierung von Curriculumtheorien. U. Hameyer, K. Frey & H. Haft, (Hrsg) *Handbuch der Curriculumforschung.* (53-103) Weinheim: Beltz.

Johnson, M. (1967). Definitions and models in curriculum theory. *Educational Theory*, 17, 127-139.

Karlin, A.J. (1980). The development of a coding system for the assessment of organizational decision-making as negotiation. *Proceedings of the 30th Annual Meeting of the International Communication Association.* Acapulco, Mexico, May 18-23.

Keough, C.M. (1986). The nature and function of argument in organizational bargaining research. *Proceedings of the 36th Annual Meeting of the International Communication Association.* Chicago, Ill, May 22-26.

Kopperschmidt, J. (1985). An analysis of argumentation. T.A. van Dijk (Ed.). *Handbook of discourse analysis.* London: Academic Press.

Mulder, M. (1988). *De curriculumconferentie in het PRABO-project.* Enschede: University of Twente, Department of Education.

Mulder, M. and Lent, J. van (1988). *Kantoorautomatisering. Een onderzoek voor leerplanontwikkeling.* Lisse: Swets & Zeitlinger.

Mulder, M. and Nijhof, W.J. (1987). Performance requirements analysis and determination. Paper presented at the *4th Annual Workshop of NeTWork "New Technologies and Work"*, Bad Homburg, April 9-11. Enschede: University of Twente, Department of Education.

Mulder, M., Nijhof, W.J. and Remmers, J.L.M. (1987). An exploration of the curriculum conference. Two case studies. Paper presented at the *Annual Meeting of the American Educational Research Association.* Washington D.C., April 23. Enschede: University of Twente, Department of Education.

Mulder, M. and Thijsen, A. (in press). Decision-making in curriculum conferences. *Journal of Curriculum Studies.*

Naess, R. (1978). *Elementare Argumentationslehre.* Baarn: Ambo.

Nijhof, W.J. and Mulder, M. (Eds.) (1986). *Basisvaardigheden in het beroepsonderwijs.* 's-Gravenhage: Instituut voor Onderzoek van het Onderwijs.

Pieters, J.M. and Mulder, M. (Eds.) (1989). Produktie-automatisering. Een onderzoek naar curriculum en instructie in het Middelbaar Technisch Onderwijs. Enschede: Universiteit van Twente, Onderzoek Centrum Toegepaste Onderwijskunde.

Poole, M.S. (1981). Decision development in small groups 1: A comparison of two models. *Communication Monographs*, 48, 1-8.

Poole, M.S. (1983). Decision development in small groups III: A multiple sequence model of group decision development. *Communication Monographs*, 50, 321-341.

Proepper, I.M.A.M. (1987). Beleidsevaluatie als argumentatie. *Beleidswetenschap*, 1, 113-135.

Putnam, L.L. and Geist, P. (1985). Argument in bargaining. *Proceedings of the 35th Annual Meeting of the International Communication Association*. San Francisco, May 24-28.

Schellens, P.J. (1985). *Redelijk argumenteren*. Dordrecht: ICG Printing.

Schmuck, R.A. (1977). *The second handbook of organization development in schools*. Palo Alto, CA: Mayfield.

Schwab, J.J. (1987). The practical: A language for curriculum. I. Westbury, and N.J. Wilkhof, (Eds.). Science, curriculum and liberal education: Selected essays. (287-321) Chicago: University of Chicago Press.

Shaw, M.E. (1981). *Group dynamics. The psychology of small group behavior.* New York: McGraw-Hill.

Smith, B.O. and Meux, M.O. (1970). *A study of the logic of teaching.* Urbana: University of Illinois Press.

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Van Bruggen, J.C. (1987). The problems and possibilities of the impact evaluation of institutes for curriculum development. *Studies in Educational Evaluation*, 13, 231-246.

Van der Sijde, P.C. (1987). *Training the teaching script.* Almere: Versluys.

Wilms, W.W. (1984). Restoring Education to Vocational Education: A Role for Evaluation? *Studies in Educational Evaluation*, 10, 5-15.

The Authors

JACQUELINE te BRAKE studied at the Department of Education of the University of Twente and is currently working at the Regional Center for Adult Education in Enschede, The Netherlands. Special interests are curriculum theory, curriculum research and adult education.

MARTIN MULDER studied at the Department of Education of the University of Utrecht and is currently working at the Department of Education of the University of Twente, The Netherlands. Special interests are curriculum theory, curriculum technology, curriculum research and human resource development.