ED 367 715 TM 021 252

AUTHOR Gijselaers, Wim
TITLE Analyses of Tutor Behavior at Different Time Points and within Different Departments.
PUB DATE Apr 94
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Background; *Behavior Patterns; Context Effect; Correlation; *Departments; Discussion Groups; Foreign Countries; Medical Education; *Medical School Faculty; Reliability; *Time Factors (Learning); Tutorial Programs; Tutoring; *Tutors
IDENTIFIERS Netherlands; *Problem Based Learning; Subject Content Knowledge

ABSTRACT

Questions about the necessity of tutors' content experience and what skills are required to make tutor behavior effective have received increased attention in the literature about problem-based learning. The present study attempts to examine the effects of context-specific variables through analysis of the generalizability of tutor behavior and the relationship between organizational background and tutoring. Subjects were 427 members of the academic staff of a Dutch medical school. Data were gathered about 2,299 tutorials over 8 consecutive years, with staff members guiding approximately 2 to 4 tutorials each year. Two studies were conducted, one to determine the stability of tutor behavior across different discussion groups; the other to explore influences on tutoring of departmental background. Results of the first study indicate that stability and generalizability of tutor behavior is low over multiple time points. Correlations between two consecutive tutorials were approximately 0.20. The second study shows that tutor behavior is related to departmental background. It is concluded that what a tutor does in a discussion group depends on context-specific characteristics and on organizational background, both features that should be considered in future research. Four tables are included. (Contains 17 references.) (Author/SLD)

Analyses of Tutor Behavior at Different Time Points and Within Different Departments.

Wim Gijselaers
Department of Educational Development and Educational Research
University of Limburg
P.O. Box 616
6200 MD Maastricht
the Netherlands

Paper presented at the Annual Meeting of the American Educational Research Association, 1994, New Orleans LA, April 4 - 8.

Abstract.

Background. Questions about the necessity of tutors' content expertise and what skills are required to make tutor behavior effective, have received increasing attention in the literature about Problem-based learning. However, studies on tutoring show inconsistent research findings about the link between tutor characteristics and student outcomes. Several researchers argue that context-specific variables (instructional characteristics and organizational context) may explain why the link between tutor behaviors and student outcomes is not consistent. The present study attempts to examine effects of context-specific variables through analyzing the generalizability of tutor behavior, and the relationship between organizational background and tutoring..

Method. Subjects were 427 members of the academic staff of the medical school at the University of Limburg, the Netherlands. During a period of eight consecutive academic years data were gathered about 2299 tutorials. On the average, staff members guided about 2 - 4 tutorials each academic year. The first study addressed the question whether tutor behavior is stable across different discussion groups. The second study explored the influences of departmental background on tutoring.

Results. Study 1: The results of this study suggest that stability and generalizability of tutor behavior is low over multiple time points. Correlations between two consecutive tutorials were about .20. Generalizability analysis showed similar results. Study 2: The findings of the second part of the study show that tutor behavior is related to departmental background. Substantianl score ranges exist between highest vs lowest rated deparments.

Conclusions. The conclusion may be drawn that what a tutor *does* in a discussion group depends on context-specific characteristics (course features, subject-area, requirements set by a discussion-group) and on organizational background. This finding suggests that future research on tutor functioning relying on multiple-group multiple-tutors designs, should also consider context-specific characteristics and organizational background.

Introduction.

Questions about the necessity of tutors' content expertise and what skills are required to make tutor behavior effective, have received increasing attention in the research literature (e.g. Albanese & Mitchell, 1993). Empirical evidence for the importance of the tutor role in problem-based learning has, for example, been found by Gijselaers and Schmidt (1990). These authors found, that group functioning was to a considerable extent (32.3% of the variance) explained by functioning of the tutor. In addition, they found that students' motivation correlated moderately ($r = .32$) with tutor's behavior. Gijselaers and Schmidt (1990) concluded that tutor behavior substantially influences the process of problem-based learning in discussion groups.

The basic issue in studies on tutoring is whether tutor characteristics can be identified that affect student achievement and lead to or constitute effective tutoring. This line of research seeks for context-free generalizations about what kind of tutor behavior is required in discussion groups. Empirical studies which seek to identify essential characteristics of tutor behavior that are required in problem-based tutorials, may in general be divided into two categories (Albanese & Mitchell, 1993). The first category contains empirical studies that are concerned with the question what kind of tutor skills are needed in the tutorial process to fulfil the tutor's tasks efficiently (e.g. Wilkerson, 1992; Moust, 1993). These studies aim to identify skills that are important in guiding the work of discussion groups. For example, Moust (1993) found that an essential tutor skill is the tutor's ability to use terminology and vocational language that is nearly equivalent with students' level of competence. He defined this ability as cognitive congruence.

The second category consists of studies that examine whether tutors must be experts to realise a certain degree of directiveness which, in turn, is assumed to influence student learning (e.g. Davis, Nairn, Paine, Anderson & Oh, 1992; Eagle, Harasym, Mandin, 1992; Moust, 1993; Schmidt, van der Arend, Kokx, & Boon, 1993a; Schmidt, van der Arend, Moust, Kokx, & Boon, 1993b). The debate about the degree of content expertise required for effective tutoring in discussion groups finds its origins in the works of Barrows (1980, 1985). He claimed that staff members who are good tutors can succesfully tutor in any course or area. Consequently, according to Barrows (1980, 1985), content expertise is not a necessary prerequisite for tutors and therefore skills for small group work are far more important. However, this assumption is increasingly questioned because of recent conflicting research findings. For example, Schmidt et al (1993a, 1993b) found that students who were guided by content experts realised a slightly higher achievement level and spent more time on self-directed learning. Studies from Eagle, Harasym and Mandin (1992) and Davis, Nairn, Paine, Anderson and Oh (1992) show that if discussion groups are guided by content-expert tutors, increased student achievement is found, more student-generated learning issues are produced, and more time is spent on self-study. However, other studies don't confirm these results. For example, Swanson, Stalenhoef-Halling, and van der Vleuten (1990) found in their large scale study (including 230 tutors and 600 students) that there was nearly no relationship between professional background of tutors and test performance of students. These researchers came to the conclusion that "it appears that physicians, biologists, and social scientists can all serve as tutors for problem-based learning groups without adverse effects on student learning" (Swanson, Stalenhoef-Halling, and van der Vleuten, 1990 pp. 133). As such, these results provided an empirical confirmation of Barrows' claims about effective tutoring. However, Schmidt (1993b) contends that given the conflicting findings, it is at least questionable that tutors do not necessarily need content knowledge.

More recently, Schmidt (1994) pointed out that inconsistencies in tutor expertise research may be explained by course specific characteristics, especially degree of course structure. He suggests that students will always try to find a minimum level of instructional structure in problem-based courses. Structure may be provided by the quality of problems and the internal relationships between problems. However, if this instructional structure is missing - for example by lack of adequate prior knowledge, or provision of ill-structured problems - students will attempt to ask for a higher degree of directiveness of the tutor. Schmidt (1994) argues that in such a situation a tutor may only respond effectively if a tutor has a sufficient degree of context-expertise. Only content-experts will be able to provide students with information that may help them to get a better understanding of the subject area of an ill-structured course.

It is interesting to note that Schmidt's (1994) explanation of the inconsistent and contradictory results in studies on tutoring, seem to parallel previous and similar research on teaching work. This line of teaching research also attempted to link specific teacher characteristics or teaching behaviors to student outcomes (e.g. Darling-Hammond, Wise & Pease, 1983). At best behavior patterns were found that contribute partially to student achievement. These authors contend that these inconsistencies may be explained by interaction effects (e.g. teacher - pupil interaction) and effects of context-specific course variables (e.g. class size, subject area, instructional quality). Factors like school organization, social variables, instructional variables interact to influence teacher behavior, student behavior and student learning. Several studies have indeed demonstrated that teachers respond specifically under different situations (e.g. Shavelson & Dempsey-Atwood, 1976; Rogosa, Floden & Willet, 1984). Teachers adjust their behavior to the changing needs of the teaching context.

In conclusion, studies on tutoring and research on teaching show inconsistent research findings about the link between teacher or tutor characteristics and student outcomes. Several authors (e.g. Darling-Hammond, Wise & Pease, 1983; Rogosa, Floden & Willet, 1984; Schmidt 1994) argue that context-specific variables (subject area, course structure) may explain why the link between tutor or teacher behaviors and student outcomes is not consistent. If the context-specific view on teaching provides a more valid perspective, it may be worthwhile to adopt this approach in the area of research on tutoring and analyse tutor behavior from this perspective.

The present paper describes an empirical study on tutoring focusing on the effects of context-specific factors. The analyses of tutor behavior are formulated through two major questions. Is the behavior of tutors consistent over time? and are organizational factors influencing tutor behavior? The first question refers to issue whether tutor behavior identified in one tutorial is also found in other tutorials. Examining the stability of tutor behavior may provide a first insight into the question whether tutors change instructional techniques or styles to suit particular tutorials, in different subject areas, or in different contexts or courses. The second question addresses the issue whether tutor's individual beliefs about how should be taught are influenced by organizational factors such as departmental background.

Before turning to the statistical analyses, it is useful to pay some detailed attention to both questions. Figure 1 contains an example of observations on tutor behavior from the data used in the present study: 10 occasions, three randomly chosen tutors.

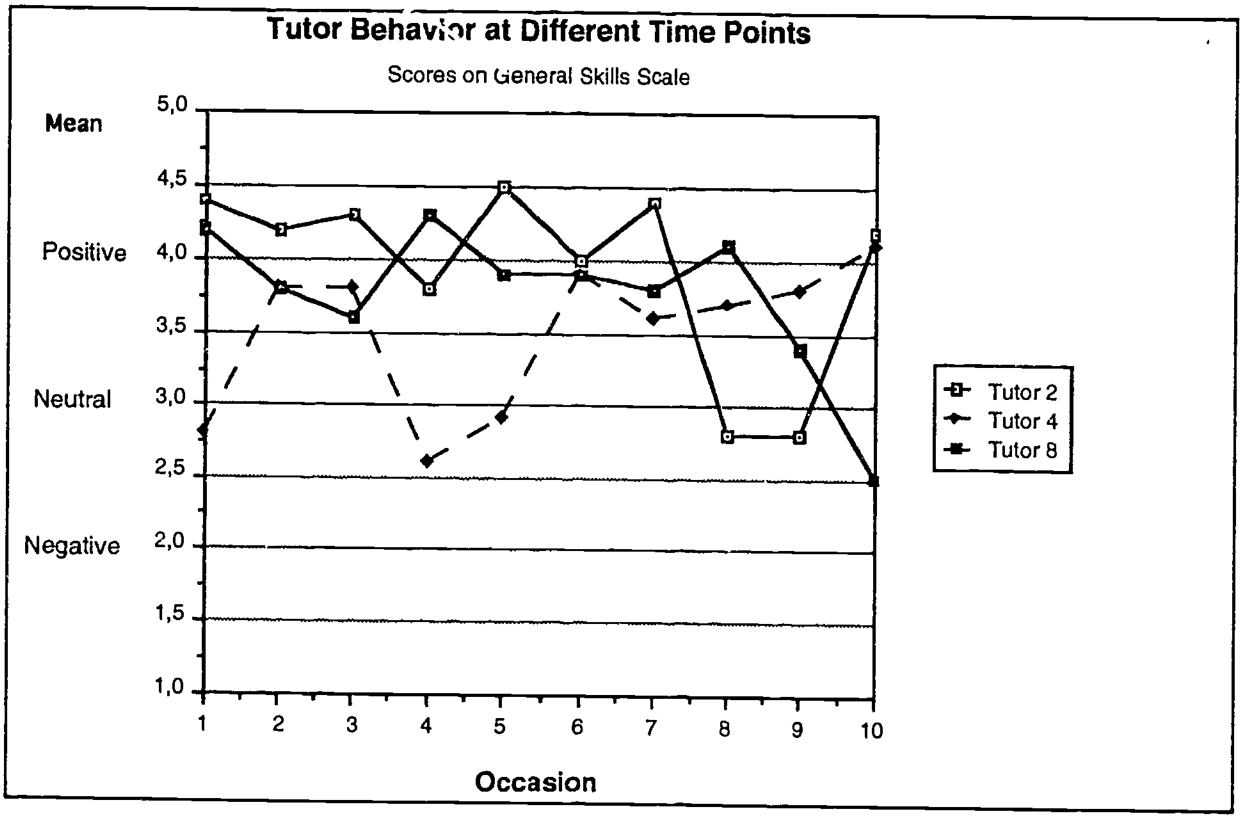

Figure 1: Tutor behavior at different time points

Every time point in figure 1, defined as occasion, reflects the teaching behavior of three different tutors as rated by their students with end - of - course evaluation questionnaires. The scores range from 1 to 5 on scale measuring general tutor skills. Scale scores were calculated as the mean score of seven five-point Likert-type items from the end - of - course evaluation neasuring tutor behavior (procedural skills to guide a discussion group efficiently). A score of 1 reflects students' rating that a tutor didn't show procedural group. A score of 5 expresses that these skills were shown by a tutor.

Figure 1 indicates that the sample of three tutors exhibits different kinds of behavior at different time points. Clearly their behavior is not stable. The individual time paths show that tutors' individual differences are not consistent over time: rank orders of time paths are not maintained over time (paths do intersect), and the distance between paths changes. In the present study the stability of tutor behavior is addressed in two analyses: 1) calculation of correlations of target behavior over multiple time points, and 2) generalizability analyses of tutor behavior.

With respect to the first question it is relevant to mention that Gijselaers (1988) conducted an earlier study about the stability of tutor behavior. In this study the behavior (measured with end-of-course student ratings) of 326 tutors was analyzed in two different courses. He found that the median correlations between time-one and time-two observations of teaching ranged about .20. Gijselaers (1988) came to the conclusion that tutor behavior is fairly unstable over different courses. He suggested that tutors probably tend to change their behavior depending upon the situational requirements set by course features and characteristics of the discussion group.

Next to the question of stability of tutor behavior, an additional question in the present study is whether organizational variables influence tutor behavior. Effective tutoring requires knowledge on the part of the tutor on how to apply against prescribed actions and to engage intentionally in tutor behavior which is supposed to be coherent with the goals of problem-based learning. Darling-Hammond et al (1983) point out that teacher behavior also depends on the knowledge that a course of action is the correct one, and that this course of action is both worthwhile and possible (this is also defined as efficacy). These authors report that research findings have shown that an individual's sense of efficacy can be influenced by interactions with others as well as by organizational factors: the value of rewards and the expectancy of achieving objectives. In this conception, tutor behavior may be affected by organizational factors such as departmental background. Different departments may share different values among individual staff members about how students should be taught in problem-based programs, and how much academic work should be spent on teaching or research. Some staff members may feel that a participatory role is required, others believe that a more directive role is necessary.

The present paper describes two studies, conducted at the medical school of the University of Limburg, Maastricht, the Netherlands. The first study is concerned with thc question of the stability of tutor behavior. The stability and generalizability of tutor behavior is examined over different time points. Proving the stability of tutor behavior would falsify the hypothesis that teaching, or in this particular case tutoring, is context dependent. The second study addresses the question whether organizational variables exist that influence tutoring. In this particular case the effects of departmental background, as operationalization of organizational context, are assessed.

Method

Subjects. Subjects were 427 members of the academic staff of the medical school at the University of Limburg, the Netherlands. During a period of eight consecutive academic years data were gathered about 2299 tutorials, guided by 427 staff members. On the average, staff members guided about 2 - 4 tutorials each academic year. The minimum amount of available data about individual tutors consisted of one observation for 95 staff members. The maximum amount of observation occasions was equal to 29 tutorials guided by 1 staff member over a period of about 8 academic years.

Description of the curriculum. The six-year medical curriculum consists of a four-year period (containing 20 problem-based courses) and a two year period of clinical clerkships. The present study used data of the four-year problem-based period. Each course follows the same problem-based format. Students met with their tutor in small-group tutorials, twice a week for two hours. At the beginning of each six-week course period students were randomly assigned to tutorial groups.

Procedure. At the end of each course a questionnaire was administered after the examination. Courses usually contained 18 groups, representing 8-10 students for each group. The surveys were completed by an average of 94 % of the students enrolled in each small group tutorial. The research that led to the development and validation of the evaluation instrument is described by Gijselaers (1988), and Gijselaers and Schmidt (1991). The questionnaire contained five-point Likert scale items (categories ranging from 1 = "entirely disagree" to 5 = "entirely agree") about various aspects of the course, including tutor's behavior. The section on tutor behavior consisted of 12 items. Previous validation research showed that one general factor determined this part of the questionnaire: general tutor skills. This factor contains the majority of the items (7 out of 12) of this section measuring general tutor skills (tutor's knowledge about principles of problem-based learning, overall functioning of the tutor, tutor's motivation, tutor's knowledge about course objectives, tutor's stimulating role in the tutorial group, tutor's support to spend time on self-study, tutor's role in evaluating the group process in the tutorial). Data were aggregated at the level of the tutorial group. The internal consistency of the general skills scale was .90 (7 items), the inter-rater reliability (intraclass-correlation) was equal to .75 (when based upon 7 ratings within a group).

Analysis Study One. To assess the stability and generalizability of tutor behavior two kinds of analyses were conducted. Stability refers to the extent that a tutor's behavior as measured at one point in time correlates with the same measure at a different point in time. The first analysis consisted of calculating correlations between observation occasions on the tutor's general skiils scale. A test-retest correlation approach was used to multiple time points. The stability measure in this study was the product-moment correlation between the measurements of the target behavior at time one and the measurements of the same target behaviors on the same set of tutors for time two.

The second analysis was conducted to assess generalizability of tutor behavior. The data for the generalizability analyses consisted of two or more observations of the target behavior for individual tutors. Generalizability refers to the extent that measures of tutor behavior are stable across different occasions (courses). Implicit in the notion of generalizability is the assumption that tutor behavior - or measured attribute - is in a steady state. Differences among scores on the measured behavior are assumed to be explained by one or more sources of error, and not to maturation, learning or deliberate and planned changes in tutor behavior. The measure of stability is the coefficient of generalizability (G-coefficient). The G-coefficient is analogous to the reliability coefficient in classical test theory (Shavelson & Webb, 1991). Generalizability theory distinguishes between decisions based on the relative standing or ranking of individuals and decisions based on the absolute level of their scores. The G-coefficient allows for making relative interpretations: are tutors ranked in a stable way across occasions? In addition, the Phi-coefficient (ϕ) allows for absolute interpretations: a score on the measure of target behavior is interpreted by its absolute level and not by how well a tutor's rating was relative to the other tutors attaining a certain rating. In the present generalizability analysis both coefficients were calculated.

Analyses Study Two. To assess the effect of organizational influences on tutor behavior, for each staff member departmental background was registered. The medical school of Maastricht is organized around 48 departments (falling within clusters of clinical, biomedical, or social sciences). For the present study data were used from tutors working in 35 departments. Oneway ANOVA's were conducted to assess the effects of departmental background on tutoring. Dependent variables were scores on the items about tutor behavior. The nominal variable "departmental background" served as independent variable.

Results

Stability and generalizability of tutor behavior. Correlations between all pairs of time points (T_1-T_{25}) were calculated for the scale measuring general tutor skills[1] . The frequency distribution of T(T-1)/2 correlations among all 25 time points was analyzed to calculate the median correlation. The median correlation provides an estimate of the stability coefficient of the target behavior. The median correlation for the targeted scale amounted about .02. This result indicates that the correlation between pairs of time points is generally near zero. The median correlations between time-one and time-two observations of tutoring ranged about .20. Correlations between time-one and time-two are one level below the diagonal of the correlation matrix. Inspection of table 1 reveals that only time-one time-two correlations are to some extent substantial, nearly all other correlations amount about zero. The results in table 1 show that stability of tutor behavior,

[1] The analysis didn't include data of three tutors for which measures of occassion 26, 28 & 29 were available, because of having less than 5 observations available for calculating correlations.

expressed as correlations between occasions, is extremely low. As such, these results are in line with previous findings described by Rogosa et al. (1984).

INSERT TABLE 1 ABOUT HERE

Generalizability theory provides a second approach to the assessment of stability of tutor behavior. The G study for tutor behavior scores had, using the terminology of generalizability theory, a crossed *p x i* design. In the present study persons are equal to tutors and items are equal to occasions. In the following analyses this design is referred to as a tutors (*t*) x occasions (*o*) design. Both, tutors and occasions, were regarded as random samples from the universe of occasions and tutors.

The expected mean squares equations were as follows (Shavelson & Webb, 1991):

$$E\,(MS_t) = \sigma^2_{to,e} + n_o\,\sigma^2_t$$

$$E\,(MS_o) = \sigma^2_{to,e} + n_t\;\sigma^2_o$$

$$E\,(MS_{to,e}) = \sigma^2_{to,e}$$

These equations are solved to obtain estimates of each variance component:

$$MS_t = \hat{\sigma}^2_{to,e} + n_o\,\hat{\sigma}^2_t$$

$$MS_o = \hat{\sigma}^2_{to,e} + n_t\;\hat{\sigma}^2_o$$

$$MS_{po,e} = \hat{\sigma}^2_{to,e}$$

Table 2 contains the results of generalizability analysis conducted on two data sets. The first data set contains data from 325 tutors on two different - consecutive - occasions. The second data contains data from 179 tutors who guided discussion groups on four different sequential occasions. The second data set is a sample of the first one. Two generalizability analyses were conducted to assess the stability of the estimated variance components. The results in table 2 show that the estimated variance components are for both data sets nearly the same. This is an indication for the robustness of the estimated variance components. The estimate variance component for occasions $\hat{\sigma}_o^2$ is in both data sets very close to zero, but negative. Therefore, these negative estimates are set to zero. Given the nearly similar estimates of variance components, the calculations based on the first data set were used for the D study (table 3).

Table 2: Comparison of ANOVA Estimates of Variance Components for the Tutor Behavior Dat Two Data Sets with (Two Occasions & 325 Tutors; Four Occasions & 179 Tutors)

Data Set I (325 tutors, 2 occasions)				*Estimated*	*Percentage*
Source of	*Sums of*		*Mean*	*Variance*	*of Total*
Variation	*Squares*	*df*	*Squares*	*Components*[a]	*Variance*
Tutors (*t*)	109.13	324	.3368	.0661 (.257)	24.4%
Occasions (*o*)	.082	1	.082	-.0003[b] (0)	≈ 0%
Interaction (*po,e*)	66.30	649	.2046	.2046 (.45)	65.6%
Data Set II (179 tutors, four occasions)					
Tutors (*t*)	79.38	178	.446	.0598 (.24)	22.5%
Occasions (*o*)	.507	3	.169	-.0001[b] (0)	≈ 0%
Interaction (*po,e*)	110.358	715	.2067	.2067 (.45)	67.5%

[a] Square roots of the estimated variance components are in parentheses.

[b] Negative estimated variance component set to zero.

The largest variance component, that for the interaction between persons and occasions (.2046), accounts for about 65% of the total variance in scores in both data sets. The variance component for tutors accounts for 24% of the total variance. A way to interpret the variance component of tutors is to relate it to the error component as a signal / noise ratio. This ratio is about (24.4/65.6) = 37% for the first data set and (22.5/67.5) = 33.3%. In relationship to the fact that calculations were based on relative few occasions (only two or four occasions), this result suggests that tutor behavior may be distinguished in a fairly reliable way. In conclusion, a "test" of tutor behavior that is one occasion in length, seems to do a relative fair job of reliable distinguishing among tutors' behavior. Another way to interpret the results is examine the magnitude of the square root of the variance component for tutors. The magnitude of a standard deviation can be used to give a rough approximation of the range of scores in a distribution. The square root of $\hat{\sigma}^2_t = .25$. So, the expected tutor scores on the general skills scale (ranging from 1 - 5) have at least a range of 1.96 (95% confidence interval) * .25 = ..49. Differences greater that .49 on the general tutor skills scale seem to make sense for making reliable distinctions between tutors within one occasion. The large variance component for the *person x occasions* interaction indicates that the relative standing of tutors differs considerably from occasion to occasion. Recall, that this is the major focus of the present study.

The results of this generalizability analysis were used to estimate in a *decision* study (D-study) how many occasions are needed to reach a value greater than .80 for the G-coefficient. D-studies use information from a G-study to design a measurement that minimizes error for a particular purpose. In this particular study the key issue is how many occasions are needed to obtain reliable information about tutor behavior. The number of occasions needed, serves as an indicator for the alleged stability of tutor behavior: the more occasions are needed, the less stable tutor behavior is in different tutorial groups (under the assumption that the measurement for tutor behavior within an occasion is reliable in a sufficient way). In this study the results from the G-study were used for the D-study.

For relative decisions the G-coefficient and ϕ for a single facet design are:

$$\mathbf{G} = \frac{\hat{\sigma}^2_p}{(\hat{\sigma}^2_p + \hat{\sigma}^2_{rel})} \qquad \hat{\phi} = \frac{\hat{\sigma}^2_p}{(\hat{\sigma}^2_p + \hat{\sigma}^2_{Abs})}$$

where:

$$\hat{\sigma}^2_{rel} = \frac{\hat{\sigma}^2_{to,e}}{n'_o} \qquad \text{and} \qquad \hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_{to,e}}{n'_o} + \frac{\hat{\sigma}^2_o}{n'_o}$$

Table 3 contains estimates of the G-coefficient and Phi coefficient for 1 to 14 occasions. The results of this analysis are nearly similar with the results of the correlational approach. The coefficient of generalizability (G-coefficient) for two occasions is equal to .43. This result indicates a low generalizability of targeted tutor behavior between time one and time two. The G-coefficient approaches a value greater than .80 only after 11 occasions.

Table 3: ANOVA Estimates of Variance Components for the Tutor Behavior Data

		G Study	*Alternative D Studies*				
Source of							
Variation	$\hat{\sigma}^2$	$n'_o =$	1	2	4	8	14
Tutors (*t*)	$\hat{\sigma}^2_t$		.0661	.0661	.0661	.0661	.0661
Occasions (*o*)	$\hat{\sigma}^2_o$		-.003 (0)	-.0002 (0)	-.0000	-.0000	-.0000
Interaction (*po,e*)	$\hat{\sigma}^2_{to,e}$		.2046	.1023	.05115	.02558	.01461
$\hat{\sigma}^2_{rel}$			.2046	.1023	.05115	.02558	.01461
$\hat{\sigma}^2_{Abs}$			.2046	.1023	.05115	.02558	.01461
G			.24	.39	.56	.72	.81
$\hat{\phi}$			.24	.39	.56	.72	.81

The ϕ–coefficient, which can be seen as a G-coefficient for absolute decisions, reaches nearly similar values as the G-coefficient for each time point. The results of the G and D study make clear that a one-occasion measurement would not provide a good estimate of a tutor's behavior. The results from both the correlational and generalizability analyses indicate that tutor behavior is in general fairly unstable. Tutor behavior is marked differently in different tutorial groups. The assumption of consistency in tutor behavior seems therefore not reasonable.

Effects of departmental background. The second part of the study focused on the question whether organizational factors, in this particular case departmental background, influenced tutor behavior. Table 4 contains the results of Oneway ANOVA's conducted on student's ratings on tutor items by departmental background. The second column of this table contains the item text. The next two columns contain the F-ratio's and η^2, as measure for the strength of association between departmental background and tutor functioning, for the complete data set consisting of 35 department. Column 5 and 6 contain the F-ratio's and η^2 for a data set consisting of departments that got the 25% lowest and 25% highest scores on the overall rating item (item number 12). This additional analysis was conducted to get estimates of the strength of association when focusing on the relative highest or lowest rated departments. The final column contains the ranges on the items of the two highest versus lowest rated departments plus the corresponding department numbers. A significant effect (F (34,2129) > 1.40, $p < .05$) of departmental background on tutor behavior was found for 10 out of 12 items on tutor behavior. The average scores on individual items, on a five-point Likert-item, ranged between 3.1 and 4.7 for tutors working in different departments. It was found that the ratings of tutors from some departments differed systematically across items.

Table 4: Oneway ANOVA student's ratings on tutor items by departmental background.

	Column 2	3	4	5	6	7	
Item	The Tutor....	F-ratio (34,2129)	η^2	F-ratio (17,1036)	η^2	Ranges Mean Scores Low	 - High
						Dep. Numbers	
1	appeared to have a thorough understanding of the course' objectives	1.80*	.026	2.71*	.043	3.6 25/45	- 4.2 26/27
2	appeared to be aware of the principles of problem-based learning.	2.27*	.028	3.20*	.050	3.7 25/2	- 4.3 16/26
3	appeared to be enthusiastic about teaching in this group.	1.28	.020	1.96*	.031	3.6 25/34	- 4.2 10/26
4	encouraged to spend substantial efforts for this course.	2.16*	.034	3.29*	.051	3.1 25/34	- 3.8 26/27
5	posed regularly questions which stimulated group-discussion	1.84*	.028	3.05*	.047	3.4 25/45	- 4.0 10/27
6	used his content-expertise to guide the group discussion.	2.5*	.038	3.07*	.048	2.3 34/45	- 3.4 26/27
7	encouraged to make appointments about subject-matter to be studied.	1.9*	.029	2.94*	.046	3.2 25/30	- 3.8 10/24
8	checked whether appointments were realized.	1.87*	.028	2.51*	.041	2.4 25/34	- 3.0 24/26
9	encouraged communication with experts in the subject-area of this course.	1.62*	.025	1.25	.02	2.6 25/34	- 3.2 26/40
10	encouraged utilization of additional learning materials.	1.36	.022	1.44	.024	2.4 10/34	- 3.1 26/36
11	took regularly the initiative to evaluate the ongoing of the discussion-group	2.53*	.039	2.34*	.037	3.0 25/30	- 3.9 10/45
12	functioned, from an overall perspective, in a good way.	1.87*	.029	2.91*	.046	3.6 25/34	- 4.3 16/26
	General tutor skills scale (item 1, 2, 3, 4, 5, 11, 12)	1.66*	.026	3.0*	.047	3.4 25/34	- 4.0 27/26

Note: *$p < .01$

The η^2 's for effects of departmental background on tutoring are in general low. The values for η^2 range about .02 for the complete data set, and range about .05 when only including the relative high and low scoring departments. These values correspond with product-moment correlations of about .20. Despite these low values for η^2 , the results in table 4 suggest that some departments receive consistent low or high values. For example departments 25 (Paediatrics), 30 (Medical Sociology) and 34 (Ophthalmology) receive tutor ratings which are about .6 - or more - lower than the highest scoring departments 10 (Dermatology), 26 (Neurology), 27 (Clinical Psychiatry) and 40 (Social Psychiatry). In conclusion, it appears that staff members from low scoring departments share different styles of tutoring than high scoring departments.

Discussion and Conclusion

The results of the present study suggest that stability and generalizability of tutor behavior is low over multiple time points. More than ten observations of individual tutors were required to reach an acceptable degree of generalizability of tutor behavior. Correlations between time-one and time-two tutoring were about .20. These results indicate that individual tutors who are tutoring "favorable" in one discussion group, may get different (higher or lower) ratings in other discussion groups. For example, good tutors at occasion one may be rated as "poor" at occasion two. These patterns are also found when time trends of individual tutors are plotted in graphs. An obvious question is how this finding may be interpreted. It may be due to poor measurement instruments, it may also be due to the fact that tutors adjust their behavior to the changing needs of the discussion group. It seems unplausible that the used measurement instruments are poorly designed, given previous validation research (Gijselaers, 1988). This research showed that the end-of-course questionnaire is fairly reliable and valid for evaluative purposes. This questionnaire allows to make reliable and valid distinctions between tutors, discussion groups and courses.

Therefore a more plausible explanation may be sought in a context-specific view on tutoring. Individual tutors probably adopt their tutor behavior, as far as possible, to the requirements of a specific discussion group in a specific course. As such, the present findings confirm Schmidt's (1994) notion that depending on the degree of course structure, students require different tutor behavior. In general, different course settings may demand different behavior patterns. Schmidt

(1994) showed that effects of content-expertise from tutors on students outcomes become visible if the degree of course structure is taken into account. He suggests that relative poor structured courses require a more directive leadership from tutors. Content-expertise is especially required in those cases when courses don't contain sufficient information about how to manage and analyse problems in discussion groups. Consequently, content-expertise may be viewed as a necessary prerequisite to apply a more directive leadership in ill-structured courses.

Schmidt's (1994) findings suggest that course structure is an important variable that interacts with behavior of discussion groups and, in turn, tutor behavior. In the present article this theory is extended and also organizational variables, next to degree of structure, are added. The present study suggests that course structure may be one out of other variables (context-specific variables such as group composition, subject area and organizational factors), explaining why tutor behavior differs in different courses. It seems evident that further generalizability analyses are necessary to examine under what conditions (for example, same course - different group; same kind of courses; same mixture of students in a discussion group on matching variables such as motivation) tutor behavior reaches more stability.

In the present article it was hypothesized that next to characteristics of instructional variables, also organizational variables interact with tutor behavior. Following the theoretical framework by Darling-Hammond et al. (1983), the question was put forward whether organizational background influences shared beliefs and values among staff members within different departments. Darling-Hammond et al. (1983) contend that within different organizational units, teachers may have substantially different ideas about how to teach and what to teach. If this notion is true, one might expect that within medical schools different departments may adhere to different ideas about teaching. The organizational looseness in university organizations makes it possible for academic staff to follow their own preferences for teacher behavior, without paying attention to the goals of an educational program. For example, innovations outreaching the level of individual instructors, such as problem-based learning, are particularly vulnerable for multiple beliefs among staff members on how to act and react in small-group discussions. Some staff members may feel that a participatory role is required, others believe that more directive role is necessary. However, it seems that in the literature on problem-based learning (Albanese & Mitchell, 1993) no studies have been conducted to assess whether organizational variables, like departmental background, influence how staff members adhere to teaching goals of problem-based learning.

The second part of the present study may be considered as an attempt to examine effects of organizational variables on tutoring. The findings of the second part of the study show that tutor behavior is related to departmental background. Individual tutors from relative low rated departments show a different leadership style than tutors from highly rated departments. Tutors from low rated departments vs high rated department differ with respect to their motivation, knowledge about problem-based learning and course objectives, and the way they guide and evaluate discussion groups. Measures of strenght of association η^2 show that variance explained by departmental background is in general low. Howevr, the low η^2 for the data set including all departments is in part a statistical artefact. If the number of departments included in the analyses is relative high, substantial increases in department effects are required to get corresponding increases in η^2. This statistical artefact may be illustrated when the number of departments is decreased. Taking only the 25% lowest and highest rated departments into consideration, η^2 gets a doubled value. Therefore, a more helpful way for interpreting department effects is to take the ranges of mean scores into consideration. These ranges show that substantial differences exist between the relative lowest and highest rated departments.

An obvious question is how differences between departments with respect to tutor behavior may be explained. As such, it remains surprising that a purely administrative variable explains to some extent variance in tutor behavior. Explanations may be sought in inconsistent effects of staff development programs, differences in organizational structure and organizational culture. For example, it may be possible that some departments get low ratings because their staff members simply lack certain basic tutor skills due to poor staff development programs. In this view tutor behavior is not the result of shared reasonable beliefs that direct actions, but merely a matter of tutor competence or the way staff members have learned on how to teach in problem-based programs. Although such an explanation may be plausible, in the context of the present study such an interpretation lacks credibility. At the Medical School of the University of Limburg all staff members are required to follow staff development programs that train them for tutoring. It seems very unlikely that these staff development programs are consistently not effective for individual tutors working in certain departments.

Organizational structure (degree of coordination within departments, department size, balance between time spent on research, education and patient care) may also result in consistent relative poor or good tutoring. For example, departments subjected to much patient care have less time to spend on teaching and might therefore consider tutoring as teaching burden, resulting in less enthusiasm when tutoring. Organizational culture (values about teaching) may be another source of explanation. It is not unthinkable that within different departments different attitudes exists towards how effective tutors should behave, or about the validity of the problem-based learning approach requiring certain teacher behavior. The present investigation doesn't reveal clear patterns of shared organizational characteristics resulting in relative low or high ratings. Although incidental post-hoc interpretations may - of course - be made, the present analysis doesn't provide cues explaining the causes of low or high ratings in certain departments. Evidently, further research is required to seek for department characteristics influencing tutor behavior.

In conclusion, the present study shows that tutor behavior is not very stable, or generalizable across multiple time points. In addition, it was shown that departmental background influences tutor behavior. Therefore, the conclusion may be drawn that what a tutor *does* in a discussion group depends on context-specific characteristics (course features, subject-area, requirements set by a discussion-group) and on organizational background influencing his or her beliefs about when certain skills should be applied in any given point in time. This finding suggests that future research on tutor functioning relying on multiple-group multiple-tutors designs, should also consider context-specific characteristics and organizational background.

References

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: a review of literature on its outcomes and implementation issues. *Academic Medicine, 68*, 52-81.

Barrows, H.S., & Tamblyn, R.M. (1980). *Problem-based learning*. New York, NY: Springer Publishing.

Barrows, H.S. (1985). *How to design a problem-based curriculum for the preclinical years*. New York, NY: Springer Publishing.

Darling-Hammond, L., Wise, A. E., Pease, S. R. (1983). Teacher evaluation in the organizational context: a review of the literature. *Review of Educational Research, 53*, 285-328.

Davis, W. K., Nairn, E., Paine, M. E., Anderson, R. M., & Oh, M. S. (1992, April). *Must small-group facilitators be content experts?* Paper presented at the American Educational Research Association, San Francisco, California.

Eagle, C.J., Harasym, P.H., Mandin, H. (1992). Effects of tutors with case expertise on problem-based learning issues. *Academic Medicine, 67*, 465-469.

Gijselaers, W. H. (1988). *Kwaliteit van het onderwijs gemeten: studies naar de betrouwbaarheid, validiteit en bruikbaarheid van studentoordelen* [Measurement of educational quality: studies on the reliability, validity and utiiity of student ratings]. Unpublished doctoral dissertation, University of Limburg, Maastricht, the Netherlands.

Gijselaers, W. H. & Schmidt, H. G. (1990). The development and evaluation of a causal model of problem-based learning. In Z. Nooman, H. G Schmidt, E. Ezzat (Eds.), *Innovation in medical education: An evaluation of its present status*. (pp. 95-113). New York: Springer Publishing Company.

Gijselaers, W.H., & Schmidt, H.G. (1991, April). *Using students' ratings as measure for educational quality: the case of problem-based medical education.*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Ill.

Moust, J. (1993). *De rol van tutoren in probleemgestuurd onderwijs: contrasten tussen student- en docent-tutoren*. [The role of tutors in problem-based learning: contrasts between undergraduate teaching assistants and faculty]. Unpublished doctoral dissertation, University of Limburg, Maastricht, the Netherlands.

Rogosa, D., Floden, R., & Willet, J. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology, 76,* 1000-1027.

Schmidt, H.G., van der Arend, A., Kokx, I., & Boon, L. (1993a, April). *Peer versus staff tutoring in problem-based learning*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Schmidt, H.G., van der Arend, A., Moust, J., Kokx, I., & Boon, L. (1993b). Influence of tutor's subject-matter expertise on student effort and achievement in problem-based learning. *Academic Medicine, 68,* 784-791.

Schmidt, H.G. (1994). Resolving inconsistencies in tutor expertise research: lack of structure causes students to seek tutor guidance. Unpublished manuscript, University of Limburg, the Netherlands.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer.* California, Sage Publications.

Swanson, D.B., Stalenhoef-Halling, B.F., & Van der Vleuten, C.P.M. (1990). Effects of tutor characteristics on test performance of students in a problem-based curriculum. In W. Bender, R.J Hiemstra, A.J.J.A. Scherpbier, & R.P. Zwierstra (Eds.), *Teaching and assessing clinical competence* (pp. 129-134). Groningen, the Netherlands: Boekwerk.

Wilkerson, L. (1992, April). *Identification of skills for the problem-based tutor: student and faculty perspectives.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California.

Table 1: Correlations[a] of target behavior (General Tutor Skills) between 25 occasions, Means, Standard Deviations and Number of Observations.

	Occasion																								
Occasion	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	--																								
2	.24	--																							
3	.08	.20	--																						
4	.22	.29	.32	--																					
5	.06	.12	.18	.28	--																				
6	.01	.07	.07	.06	.13	--																			
7	.01	.13	.09	.02	.13	.28	--																		
8	-.20	.01	.01	.10	.14	.01	.08	--																	
9	.08	.01	.02	.07	-.07	-.08	.11	.35	--																
10	-.05	-.10	.15	-.03	-.11	.10	-.09	.20	.22	--															
11	.05	.04	.13	.14	.20	.04	.23	-.08	.10	.02	--														
12	-.01	-.01	.05	.19	-.05	.17	.05	.12	.04	.15	-.06	--													
13	.17	-.08	.14	-.01	.08	.06	-.12	.22	.32	.46	-.03	-.06	--												
14	.09	-.17	-.13	.02	-.06	-.15	-.08	.04	.16	.30	.05	-.01	.18	--											
15	.05	-.06	-.10	.08	-.25	.22	-.19	-.04	.38	.37	-.13	.43	.33	.23	--										
16	-.34	-.18	-.25	-.14	-.35	-.21	.39	-.05	.17	.07	.22	-.17	-.11	.09	.16	--									
17	.01	-.46	-.15	.04	-.43	-.25	.11	.02	.04	.06	-.12	.21	.17	.28	.33	.34	--								
18	-.14	-.41	.26	.31	-.05	-.22	.02	.07	-.25	-.19	.21	.19	-.08	.04	-.23	.01	.48	--							
19	.47	.06	.32	.22	.12	.14	.12	.34	-.10	-.20	-.15	.44	-.02	.16	-.07	-.08	.20	.26	--						
20	-.18	.31	-.13	-.22	.13	.27	.49	-.46	-.38	-.25	.22	-.27	-.51	-.39	-.27	.01	-.31	-.06	-.08	--					
21	-.11	-.01	-.31	-.25	.53	-.39	.23	.17	-.28	-.18	-.08	-.17	-.07	-.34	-.28	-.36	-.49	-.12	-.22	.34	--				
22	.02	-.42	-.42	.09	.18	-.29	-.31	-.45	.53	.26	.33	-.37	.08	.65	.08	-.27	-.34	-.28	-.40	.03	.08	--			
23	.01	.15	-.14	.02	-.25	.25	-.57	-.23	-.63	.35	.42	-.28	.39	-.30	-.03	-.30	.01	.13	-.64	.01	.11	-.07	--		
24	-.11	.33	-.32	-.05	-.24	.30	-.43	-.49	-.58	.36	.38	-.40	.07	-.21	-.05	-.36	-.41	-.20	-.82	.35	.23	.47	.85	--	
25	-.53	.15	.45	.49	.06	.13	-.60	.25	-.84	-.24	.92	.35	-.13	-.61	-.21	-.21	-.49	.17	-.74	-.22	.29	.04	.72	.59	--
Mean	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.9	3.8	3.9	3.9	3.9	3.8	3.8	3.8	3.9	3.8	3.9	3.9	4.0	4.0	3.8	3.9	4.0	4.0
Sd.	.5	.5	.5	.5	.5	.5	.5	.4	.4	.4	.6	.5	.5	.4	.6	.3	.5	.6	.5	.3	.5	.7	.5	.4	.5
N	423	328	247	188	159	134	111	92	80	72	61	55	44	37	30	28	23	20	18	17	15	12	9	8	5

Note: Underlined correlations - Signif. LE .01 (2-tailed). [a] Calculations based on pair-wise deletion of missing values. This explains why the number of subjects slightly differ with the number of subjects in the generalizability analyses.