A rating scale for tutor evaluation in a problem-based curriculum: validity and reliability

D. H. J. M. DOLMANS[1], I. H. A. P. WOLFHAGEN, H. G. SCHMIDT & C. P. M. van der VLEUTEN

Department of Educational Development and Research, University of Limburg, Maastricht, The Netherlands

Summary. An instrument has been developed to assess tutor performance in problem-based tutorial groups. This tutor evaluation questionnaire consists of 13 statements reflecting the tutor's behaviour. The statements are based on a description of the tasks set for the tutor. This study reports results on the validity and reliability of the instrument. Confirmatory factor analysis showed that a three-factor model fitted the data reasonably well. The three factors are: (1) guiding students through the learning process, (2) content knowledge input, and (3) commitment to the group's learning. Generalizability studies indicated that the rating scales provide reliable information with student responses of existing tutorial group sizes. It is concluded that the tutor evaluation questionnaire is a fairly valid and reliable instrument that can be used in staff development programmes.

Key words: teaching/*stand; *education, medical, undergraduate; *curriculum; program evaluation; problem solving; sensitivity; Netherlands

Introduction

The tutor plays a key role in problem-based curricula. This notion is confirmed by research conducted by Gijselaers & Schmidt (1990). These authors postulated a causal model of problem-

based learning (PBL) to identify and measure the effects of important variables, such as group functioning, quality of the problems presented, time spent on self-study, student interest in the subject matter and student achievement. One of their findings was that tutor functioning has a direct causal influence on the functioning of small group tutorials which in turn influences students' interest in the subject matter. Tutor performance also has an indirect causal effect on student achievement. These results reflect the importance of tutors' abilities to guide tutorial groups in an adequate way.

For evaluation of tutor skills, the monitoring of staff conducting this role is required. It would be desirable to have an instrument available in order to collect information about the performance of the tutor. Such an instrument would enable the school to provide tutors with feedback. Training and remedial teaching could be provided to tutors based upon the shortcomings pointed out by the evaluation.

As a tutor evaluation questionnaire can provide teachers with feedback, items should reflect key features of the tutor role. This implies that the questionnaire should be based on the tasks set for the tutor at the medical school in which the instrument will be used, as well as on theoretical conceptions about the tutor role, as described in the literature (Barrows 1988). The role of a tutor in a problem-based curriculum mainly consists of guiding the tutorial group. The tasks of the tutor are to guide students through the learning process, to encourage students to attain a deeper level of understanding, to ensure that all students are involved in the group process, to monitor progress of individual students, to motivate students and to help the student group to deal

with their own problems of interpersonal dynamics (Barrows 1988).

In contrast to an abundance of literature describing desirable tutor skills, e.g. Barrows & Tamblyn (1980), Barrows (1988), Wilkerson (1992), Wilkerson, Hafler & Liu (1992), Kalishman & Mennin (1993), only few studies have been reported identifying important features of the tutor role. Most of these studies were concerned with tutor's expertise on the subject matter under discussion (De Volder 1982; Feletti *et al.* 1982; Swanson, Stalenhoef-Halling & van der Vleuten 1990; Davis *et al.* 1992; Moust & Schmidt 1992). De Volder (1982), for instance, found that tutor functioning as judged by students was positively related to expertise on subject matter and experience. Feletti *et al.* (1982) showed that good tutors were perceived by students as having a thorough up-to-date knowledge of the particular problem being studied and as encouraging them to review their academic progress.

Wilkerson (1992) and Moust (1993) investigated effective tutor behaviour stimulating student learning. Wilkerson (1992) concluded that two factors describing the skills perceived as most helpful by both tutors and students were maintaining positive interactions within the group and providing assistance in getting the work of the group accomplished. Moust (1993) found that, to facilitate student learning, a tutor should use terminology adapted from the students' level of competence. In other words, the tutor should ask questions in a language that students can understand. The tutor's interest in students' daily lives and personalities also appeared an important feature of effective tutor behaviour.

From these descriptions, it becomes apparent that at least two aspects of a tutor's performance seem to be essential; content knowledge input and commitment to the group's learning. In addition, theoretical conceptions about the tutor's role stress the importance of the tutor to guide students through the learning process. Based on these considerations, a tutor evaluation questionnaire was developed reflecting three aspects of a tutor's performance: (1) guiding students through the learning process, (2) content knowledge input, and (3) commitment to the group's learning.

In this study, the results will be presented of a confirmatory factor analysis which has been conducted to assess the validity of the tutor evaluation questionnaire. In addition, the results of generalizability studies will be presented to estimate the instrument's reliability. Potential sources of error variance were analysed and used to estimate the number of required student responses to obtain reliable information for one tutor.

Method

Subjects

Data were collected during 19 6-week courses in the academic year 1992–1993: five first-year courses, five courses in the second-year, five courses in the third-year and four courses in the fourth-year. In total 18 tutorial groups participated in each course. The number of students participating in each tutorial group was either nine or 10. The total average response rate was 81%. Consequently, not all students rated their tutor, therefore some tutors were judged by less than nine or 10 students. The total number of tutors involved in this study was 293.

Instruments

A pilot study was conducted to identify key tutor skills. In the academic year 1990–1991, 100 students and 150 tutors, divided among the first four curriculum years, received a list of 16 items specifying behavioural characteristics of the tutor. Both students and tutors were asked to rate whether each item was assumed to be an important indicator of a tutor's performance. Furthermore, they were asked to indicate whether each item was clearly stated. These 16 items were based on the tasks set for the tutor and on studies investigating effective tutor-behaviour stimulating student learning, as described in the introduction section. This pilot study resulted in a list of 13 statements: six items related to the tutor's task to guide students through the learning process, four items about the tutor's content knowledge input and three items about the tutor's commitment to the group's learning.

At the end of each course, students were asked to rate whether their tutor demonstrated the

behaviour described in each statement: insufficiently (1), neutral (2), sufficiently (3). A 'not applicable' response option was added, which could be selected if, for instance, students initiated the activity described by the statement and tutor intervention in this respect was not necessary. The items of the tutor evaluation questionnaire and the underlying factors are shown in Table 1. The mean score on each item for the total group of 293 tutors varied between 2·78 (SD = 0·32) and 2·18 (SD = 0·57), on a three-point scale. The mean scores and standard deviations for each of the 13 items are also shown in Table 1. The low mean score for item 6 may reflect a general tendency among tutors to evade providing adequate feedback.

The feedback that is sent to individual tutors contains the scores of the students on individual items, expressed as the average percentage of students that rated the behaviour described in each statement as insufficient, neutral or sufficient. These percentages per item are averaged across the 13 items, resulting in a total average percentage of insufficient, neutral and sufficient for each tutor. In addition, the average percentage sufficient minus insufficient is computed for each tutor. This aggregated score varies between −100% and 100%. This score is used for a final qualification: extremely poor (the aggregated score is lower than or equal to −76); poor (higher than −76 and lower than or equal to −51); insufficient (higher than −51 and lower than or equal to −11); doubtful (higher than −11 and lower than or equal to 10); sufficient (higher than 10 and lower than or equal to 50); good (higher than 50 and lower than or equal to 75), and very good (higher than 75). In addition, students are asked to give an overall judgement (ranging from 1 to 10, 6 being 'sufficient') of the performance of the tutor.

Procedure

To achieve a fully balanced design convenient for statistical analysis, a random sample of six students was selected from the total number of

Table 1. Items of the tutor evaluation questionnaire (the observed variables) and their common factors, mean scores (scale 1–3) and standard deviations (SD), $n = 293$

	Mean	SD
Factor 1 *Guiding students through the learning process*	2·51	0·36
1. The tutor demonstrates to be well-informed about the process of problem-based learning.	2·78	0·32
2. The tutor stimulates all students to participate actively in the tutorial group process.	2·38	0·48
3. The tutor stimulates a careful analysis of the problems.	2·62	0·43
4. The tutor stimulates the generation of specific learning issues useful for self-study.	2·55	0·37
5. The tutor stimulates an extensive reporting on information collected during self-study.	2·52	0·43
6. The tutor stimulates evaluation of the tutorial group process.	2·18	0·57
Factor 2 *Content knowledge input*	2·61	0·39
7. The tutor has an understanding of the subject matter covered in the course.	2·67	0·43
8. The tutor assists students in distinguishing main issues from minor issues.	2·58	0·39
9. The tutor uses his or her expert knowledge appropriately.	2·59	0·45
10. The tutor contributes towards a better understanding of the subject matter.	2·59	0·46
Factor 3 *Commitment to the group's learning*	2·62	0·39
11. The tutor gives an impression of being motivated.	2·69	0·42
12. The tutor shows interest in our learning activities during the course.	2·54	0·42
13. The tutor shows commitment with respect to group functioning.	2·62	0·40

students rating that tutor. If a tutor was judged by less than six students, the particular tutor and his group were excluded from the analysis. As 19 courses were included in this study, and 18 tutorial groups participated in each course, about 342 (19 × 18) tutors were initially involved. In the data set used for analysis 293 tutors were included (49 tutors participating in the courses under study were excluded owing to balancing). In total, 293 tutors were each rated by six students. Thus, 1758 student judgments (293 × 6) were available. Both students and tutors were randomly assigned to the tutorial groups.

Statistical analysis

A confirmatory factor analysis was carried out to assess the adequacy of the theoretical tutor performance model outlined above. For the confirmatory factor analysis, data were aggregated at the tutorial group level by computing average scores across students for each tutor. In total, 293 tutors or cases were involved in the confirmatory factor analysis. Table 2 presents the resulting correlation matrix used as input for confirmatory factor analysis. In the confirmatory factor model, as specified in this study, all common factors were correlated, observed variables one through to six were affected by the first common factor, observed variables seven to 10 were affected by the second common factor, variables 11, 12 and 13 by the third common factor. Table 1 contains the mean scores and standard deviations for each factor. Furthermore, all observed variables were assumed to be affected by a unique factor (error in each variable), and no pairs of unique factors were correlated. The Lisrel VII program (Jöreskog & Sörbom 1990) was used to determine whether the data confirmed this model. As an alternative check, a one-factor model and a two-factor model were tested.

Results

Each tutor receives a final qualification varying from extremely poor to very good, based on the aggregated score across items. In Table 3 the percentages of tutors that received a qualification are summarized. As can be seen in this Table, 10·3% of the tutors received a qualification below sufficient. It is worth mentioning that all tutors at the medical school of the University of Limburg are obliged to attend training before they can fulfil the tutor role.

Construct validity

Table 2 shows that the correlation coefficients between the observed variables varied between 0·52 and 0·88 ($n = 293$), except for item 6 'the tutor stimulates the evaluation of the group process' which correlated between 0·26 and 0·59 with the other variables. The correlation between common factor 1 and 2 was 0·76, between factor 1 and 3, 0·81 and between factor 2 and 3, 0·65. Thus, all three factors were highly correlated.

A model is assumed to fit the data if three conditions are met: (1) the χ^2 divided by the degrees of freedom should be lower than 2, a

Table 2. Correlation matrix between 13 items ($n = 293$)

	1	2	3	4	5	6	7	8	9	10	11	12	13
1													
2	0·66												
3	0·67	0·79											
4	0·63	0·67	0·77										
5	0·65	0·73	0·83	0·73									
6	0·50	0·59	0·44	0·40	0·49								
7	0·68	0·58	0·74	0·68	0·70	0·35							
8	0·58	0·58	0·74	0·71	0·71	0·26	0·77						
9	0·65	0·60	0·71	0·63	0·67	0·27	0·74	0·78					
10	0·62	0·58	0·74	0·64	0·73	0·27	0·82	0·77	0·80				
11	0·64	0·71	0·70	0·65	0·73	0·52	0·60	0·54	0·56	0·62			
12	0·58	0·73	0·68	0·60	0·67	0·52	0·56	0·52	0·57	0·60	0·85		
13	0·62	0·72	0·69	0·63	0·70	0·52	0·55	0·55	0·58	0·60	0·88	0·86	

Table 3. Distribution of qualification on tutor performance (mean = 5·82, SD = 1·18)

Qualification	Percentage of tutors ($n = 293$)
1 Extremely poor	0·7
2 Poor	0·7
3 Insufficient	4·8
4 Doubtful	4·1
5 Sufficient	22·2
6 Good	34·8
7 Very good	32·8

P-value that differs from zero; (2) the root mean square residual should be lower than 0·07; and (3) the goodness-of-fit index and the adjusted goodness-of-fit index, which takes into account the number of degrees of freedom, should be higher than 0·80 (Saris & Stronkhorst 1984).

Despite the high correlation coefficients between the three factors, a one-factor model did not fit the data (χ^2 [65 d.f.] = 819·86, P = 0·000, a root mean square residual of 0·076, a goodness-of-fit index of 0·621, and adjusted goodness-of-fit index of 0·47). The results of a two-factor model in which variables 1 through 6, 11, 12, and 13 were affected by one factor and variables 7 to 10 were affected by another factor showed that only the second condition specified by Saris & Stronkhorst (1984) was satisfied (χ^2 [64 d.f.] = 534·40, P = 0·000, a root mean square residual of 0·064, a goodness-of-fit index of 0·74, and adjusted goodness-of-fit index of 0·62). The three-factor model in which observed variables 1 through 6 are affected by the first common factor, observed variables 7 to 10 are affected by the second common factor, and variables 11, 12 and 13 are affected by the third common factor showed the following results: χ^2 [62 d.f.] = 240·62, P = 0·000, a root mean square residual of 0·047, a goodness-of-fit index of 0·887, and adjusted goodness-of-fit index of 0·835. Thus, for the three factor model, the first condition specified by Saris & Stronkhorst (1984) is not satisfied, whereas both other conditions are fulfilled.

In order to further cross-validate the proposed models, the data set was split up in two sets. Set one consisted of a random set of 10 courses (161 tutors) and set two consisted of the other nine courses (132 tutors). With regard to the first data set, the three-factor model fitted the data reasonably well, i.e., the first condition specified by Saris & Stronkhorst (1984) was not satisfied, the second was satisfied, and the third was almost satisfied (χ^2 [62 d.f.] = 189·63, P = 0·000, a root mean square residual of 0·055, a goodness-of-fit index of 0·846, and adjusted goodness-of-fit index of 0·773). With respect to the second data set the first condition was not satisfied, whereas the second and third condition were satisfied (χ^2 [62 d.f.] = 137·88, P = 0·000, a root mean square residual of 0·048, a goodness-of-fit index of 0·861, and adjusted goodness-of-fit index of 0·796).

In general the results of the confirmatory factor analyses indicate that the three-factor model shows a better fit than the one-factor model and the two-factor model, because two out of three conditions were consistently satisfied for the three-factor model. In other words, the three-factor model sems to show a reasonable fit.

The relationship between the overall judgement of a tutor and the scores on individual items also provide information about the construct validity of the questionnaire. The overall judgement correlates highly with all 13 items (all were above 0·54, P < 0·001), with the exception of item 6 'the tutor stimulates evaluation of the group process' which correlated 0·40 (P < 0·001).

Generalizability studies

Generalizability studies were conducted to estimate the reliability of the aggregated score, the 13 items and the three factors (Brennan & Kane 1979; Crick & Brennan 1983). The analyses were conducted at the level of individual students. In total 293 tutors were involved that were each judged by six students. In other words, 1758 cases (293 × 6) were included. For the aggregated score an all random students-nested-within-tutors design was used, with tutors as universe of generalization or object of measurement. This design allows variance component estimation of two sources: (1) difference between tutors (T) (object of measurement), and (2) differences between students nested within tutors and general error (S:I, e) (Shavelson & Webb 1991). As different tutors are judged by different students, it is impossible to determine whether students

Table 4. Estimated variance components of the aggregated score (scale −100 to +100)

Source	DF	Estimated variance component	Standard error	Percentage of total variance
Aggregated score Tutors (T)	292	1038·24	99·57	50·9
Students within Tutors (S:I), e	1465	999·86	36·92	49·1

differed in their judgements overall (student effect), or whether tutors are rank-ordered differently by students (tutor-by-student interaction effect).

In Table 4, the sources of variability and the corresponding estimated variance components are summarized. The percentage of variance associated with tutors for the aggregated score is 50·9. Approximately one-half of all variance can be attributed to variation between tutors. This percentage is the true variance or the variance of interest and apparently this instrument is able to discriminate tutor behaviour to a fair extent.

The estimated variance components presented in Table 4 were used to estimate reliability indices. Although the interpretation of scores from this instrument can be used in both an absolute and relative way, the present design yields equal reliability estimates for both interpretation perspectives (as students are nested within tutors). Hence, all variance components were included in the observed variance definition. For the aggregated score this computation revealed a reliability coefficient of 0·86, with an average group size of six students judging the

tutor. This coefficient indicates the expected correlation between tutor scores derived from similar, but not identical ratings using a different random sample of students.

Table 5 provides the reliability coefficients of tutor rating scores as a function of the numbers of student responses and the corresponding standard error of measurement (SEM). The reliability coefficient indicates how many students are required to obtain a minimal generalizability coefficient of 0·80. The standard error of measurement (SEM) also provides relevant information with regard to the reliability of the tutor rating scale. The SEM can be used to estimate confidence intervals for individual scores. For example, the 95% confidence interval of a score can be estimated by multiplying the SEM by 1·96 (Ferguson 1981). Taking the arbitrary standpoint that at least a difference of 40 points is required to obtain reliable results for the aggregated score, the SEM should be lower than or equal to 20·41 (40 divided by 1·96) at the level of 95%. Taking into account both the practical significance level of 40 points for the aggregated score and a generalizability coefficient of 0·80,

Table 5. Generalizability coefficients and standard errors of measurement (SEM), as a function of the number of student ratings for the aggregated score (scale −100 to +100)

Student responses	Generalizability coefficient	SEM
1	0·5094	31·6205
2	0·6750	22·3591
3	0·7570	18·2561
4	0·8060	15·8103
5	0·8385	14·1411
6	0·8617	12·9090

four students are required to obtain a reliable aggregated score.

Generalizability studies were also conducted to estimate the reliability of the 13 items and the reliability of the three factors. The variance components that were included in these analyses were: (1) difference in tutors (T) (object of measurement); (2) differences in items (I); (3) differences between students nested within tutors (S:T); (4) interaction between tutors and items (TI); and (5) interaction between items and students nested within tutors and general error (IS:T, e) (Shavelson & Webb 1991). In Table 6, the sources of variability and the corresponding estimated variance components are summarized for both the 13 items and the three factors. The percentage of variance associated with tutors for the 13 items is 25·6. For the factors 2 and 3 this percentage is somewhat higher, 30·8 and 30·7 respectively. For factor 1 this percentage is 20·9.

Approximately one-fourth of all variance can be attributed to variation between tutors, which indicates that the items and the three factors discriminate tutor behaviour to a fair extent.

The estimate variance components presented in Table 6 were used to estimated reliability indices. For the 13 items the reliability coefficient was 0·98. For factors 1 to 3 these coefficients were 0·85, 0·94 and 0·92, respectively. In Table 7 the reliability coefficients of the 13 items and the three factors are presented as a function of the number of student responses and the standard error of measurement (SEM). If it is assumed that at least a difference of 0·5 points at the three-point scale is required to obtain reliable results, the SEM should be lower than or equal to 0·26 (0·5 divided by 1·96). Taking into account this practical significance level of 0·26 and a generalizability coefficient of 0·80, Table 7 demonstrates that for the 13 items, one student is sufficient to obtain

Table 6. Estimated variance components of the 13 items and the three factors (scale 1–3)

Source	DF	Estimated variance component	Standard error	Percentage of total variance
13 items				
Tutors (T)	292	0·1182	0·0099	25·6
Items (I)	12	0·0021	0·0000	0·5
Students within Tutors (S:I)	1465	0·0137	0·0005	3·0
Tutors by Items (TI)	3504	0·0635	0·0026	13·8
IS:T, e	17580	0·2639	0·0028	57·2
Factor 1				
Tutors (T)	292	0·1099	0·0107	20·9
Items (I)	12	0·0002	0·0000	0·0
Students within Tutors (S:I)	1465	0·1198	0·0044	22·8
Tutors by Items (TI)	3504	0·0833	0·0044	15·9
IS:T, e	17580	0·2119	0·0035	40·4
Factor 2				
Tutors (T)	292	0·1473	0·0129	30·8
Items (I)	12	0·0001	0·0000	0·0
Students within Tutors (S:I)	1465	0·0521	0·0019	10·9
Tutors by Items (TI)	3504	0·0526	0·0044	11·0
IS:T, e	17580	0·2267	0·0048	47·3
Factor 3				
Tutors (T)	292	0·1427	0·0127	30·7
Items (I)	12	0·0001	0·0000	0·0
Students within Tutors (S:I)	1465	0·0705	0·0026	15·1
Tutors by Items (TI)	3504	0·0659	0·0057	14·2
IS:T, e	17580	0·1859	0·0049	40·0

Table 7. Generalizability coefficients (G) and standard errors of measurement (SEM), as a function of the number of student ratings for the 13 items and the three factors (scale 1–3)

Student responses	Item 1–13		Factor 1		Factor 2		Factor 3	
	G	SEM	G	SEM	G	SEM	G	SEM
1	0·90	0·117	0·49	0·346	0·74	0·228	0·67	0·266
2	0·95	0·083	0·65	0·245	0·85	0·162	0·80	0·188
3	0·96	0·068	0·73	0·200	0·89	0·132	0·86	0·153
4	0·97	0·058	0·79	0·173	0·92	0·114	0·89	0·133
5	0·98	0·052	0·82	0·155	0·93	0·102	0·91	0·119
6	0·98	0·048	0·85	0·141	0·94	0·093	0·92	0·109

reliable results. The results in Table 7 furthermore show that at least five students are required to obtain reliable results for factor 1 and two students are required to obtain reliable results for factors 2 and 3.

Conclusion

The purpose of this study was to investigate the validity and reliability of the tutor evaluation questionnaire. The results of the confirmatory factor analyses indicate that a three-factor model comprising 13 items fits the data better than a one-factor or a two-factor model. For the three-factor model two out of three statistical conditions specified by Saris & Stronkhorst (1984) were at least satisfied, which indicates a reasonable fit. However, all three factors are highly correlated with each other. This suggests that the factor scores provide little evidence of differential performance in the three areas.

Generalizability analyses indicated that the tutor evaluation questionnaire is a reliable instrument. The reliability coefficients for the aggregated score was 0·86, based on six student ratings. For the aggregated score, at least four students are required to obtain reliable results. The reliability coefficient for the 13 items was 0·98, based on six student ratings. One student seems to be sufficient to obtain reliable results at the level of 13 items. For factors 1 to 3 the reliability coefficients were 0·85, 0·94 and 0·92 respectively. The number of students required to obtain reliable results were for factors 1 to 3, five students, two students and two students, respectively. This implies that tutor evaluations with this questionnaire provide reliable informa-

tion in most problem-based settings where group sizes of the one used in this study are quite regular.

The feedback that is sent to individual tutors does not contain information about the performances of the tutor with regard to the three factors, but only information at the level of individual items and the aggregated score. The reliabilities of the item scores and the aggregated score indicate that these scores can be used for staff development purposes. The high correlation between the three factors limit the usefulness of the three factor scores to highlight areas of a tutor's strength and weaknesses. As the feedback only contains information about the scores on individual items and the aggregated score, the results of this study imply that, if evaluations of a tutor are based upon less than four student ratings, caution is in order.

This high reliability of the aggregated score indicates that these scores can be potentially used in faculty development programmes. These scores provide a source of information that could be used in promotion, tenure and salary decisions. Further research is however needed to investigate the consistency of tutor performance over time. Finally, assessing tutors' performances in the tutorial group also requires training programmes for poor scoring tutors, as the ultimate purpose is to improve tutors' behaviour.

References

Barrows H.S. & Tamblyn R.M. (1980) *Problem-based learning: an approach to medical education.* Springer-Verlag: New York.

Barrows H.S. (1988) *The tutorial process*. Southern Illinois University School of Medicine: Illinois.

Brennan R.L. & Kane M.T. (1979) Generalizability Theory: A Review. *New Directions for Testing and Measurement* **4**, 33–51.

Crick J.E. & Brennan R.L. (1983) *Manual for Genova: A Generalized Analysis of Variance System*. American College Testing Program: Iowa.

Davis W.K., Nairn R., Paine *et al.* (1992) *Must Small-Group Facilitators Be Content Experts?* Paper presented at the American Educational Research Association, San Franciso, CA, USA.

De Volder M.L. (1982) Discussion groups and their tutors: Relationship between tutor characteristics and tutor functioning. *Higher Education* **11**, 269–72.

Feletti G.I., Doyle E., Petrovic A. & Sanson-Fisher R. (1982) Medical students' evaluation of tutors in a group-learning curriculum. *Medical Education* **16**, 319–25.

Ferguson G.A. (1981) *Statistical Analysis in Psychology and Education* (5th edn.). McGraw-Hill: Auckland.

Gijselaers W.H. & Schmidt H.G. (1990) Development and evaluation of a causal model of problem-based learning. In: *Innovation in Medical Education: An evaluation of its present status* (ed. by A.M. Nooman, H.G. Schmidt & E.S. Ezzat), pp. 95–113. Springer-Verlag: New York.

Jöreskog K.G. & Sörbom D. (1990) *Lisrel VII. User's Guide*. National Educational Resources: Chicago.

Kalishman S. & Mennin S. (1993) *The Tutor and the Tutorial Process: A View from Within a Problem-based Tutorial Group*. Paper presented at the American Educational Research Association, Atlanta, Georgia, USA.

Moust J.C. & Schmidt H.G. (1992) *Undergraduate students as tutors: Are they as effective as faculty in conducting small-group tutorials?* Paper presented at the American Educational Research Association, San Francisco, CA, USA.

Moust J.C. (1993) *De rol van tutoren in probleemgestuurd onderwijs. Contrasten tussen student- en docenttutoren.* [The role of tutors in problem-based learning. Contrasts between student- and staff-tutors] PhD thesis, University of Limburg, Maastricht, The Netherlands.

Saris W. & Stronkhorst H. (1984) *Causal modelling in nonexperimental research*. Sociometric Research Foundation: Amsterdam.

Shavelson R.J. & Webb N.M. (1991) *Generalizability Theory. A Primer*. Sage, London.

Swanson D.B., Stalenhoef-Halling B.F. & Vleuten van der C.P.M. (1990) Effect of tutor characteristics on test performance of students in a problem-based curriculum. In: *Teaching and assessing clinical competence* (ed. by W. Bender, R. Hiemstra, A.J.J.A. Scherpbier & R.P. Zwierstra), pp. 129–33. Boekwerk Publicaties: Groningen.

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

Wilkerson L.A., Hafler J.P. & Liu P. (1992) A case study of student-directed discussion in four problem-based tutorial groups. *Academic Medicine* **66**, 9, S79–S81.