Teaching and Learning in Medicine
1995, Vol. 7, No. 2, 82–91

APPLIED RESEARCH

Theory-Guided Design of a Rating Scale for Course Evaluation in Problem-Based Curricula

Henk G. Schmidt, Diana Dolmans, and Wim H. Gijselaers
University of Limburg
Maastricht, The Netherlands

Jacques E. Des Marchais
University of Sherbrooke
Sherbrooke, Quebec, Canada

In this article, the development and evaluation of a rating scale for course improvement is described. Unlike other attempts in this area, design was guided by an explicit theory describing the teaching and learning processes taking place in problem-based curricula. Data were collected in two curricula using the problem-based learning approach. A confirmatory factor analysis was conducted to test to what extent the model underlying the rating scale described the data. In addition, generalizability and utility studies were conducted. The results indicate that the rating scale data fitted the underlying model reasonably well. Most factors of the model showed good generalizability, not only across facets judged but also across items and judges. The potential utility of the instrument was illustrated by analyses of some aspects of one of the curricula. It was argued that, although the instrument developed is useful only within the context of problem-based curricula, the general approach followed here may be of use in every educational institution interested in program evaluation through systematic questioning of student-observers.

Evaluation of the efficiency of curricula for health professions education may serve various goals. One purpose of curriculum evaluation is accountability. Educational institutions are required to provide information concerning the quality of the health professionals they graduate. Graduates must be trusted to possess at least a sufficient level of competency before entering the health care system, and medical schools are responsible for demonstrating that their graduates actually possess the required competencies.

In addition, program evaluation has another role to play: to provide information suitable for proposing and carrying out improvements of an ongoing program. Outcome data, however useful, hardly provide information that is sufficiently specific to undertake small-scale, course-level curriculum improvement. Thus what is needed, in addition to outcome studies as an instrument for accountability, is a strategy that deals with the curriculum in its own right, provides information about its shortcomings, and enables a school to implement improvements whenever and wherever necessary.

Both the University of Limburg at Maastricht, The Netherlands, and the Université de Sherbrooke, Québec, Canada, use an approach to program evaluation for problem-based curricula that may satisfy these requirements. The approach is characterized by five distinctive steps.

1. After each course or "unit," a standardized rating scale is administered to students participating in that course. The rating scale is used as a global screening

The preparation of this article was supported by a grant of Le Fonds de Dotation de la Faculté de Médecine, Université de Sherbrooke while the first author was a visiting professor at that institution.

Parts of this article were presented at the Annual Meeting of the American Educational Research Association, San Francisco, March 1989.

Correspondence may be addressed to Henk G. Schmidt, PhD, Dean, Faculty of Health Sciences, Maastricht University, Post Office Box 616, 6200 MD Maastricht, The Netherlands.

device; it enables the evaluators to find out where weaknesses may reside within the program.

2. If shortcomings are detected, more in-depth investigations may be carried out. These investigations include more detailed analyses of the ratings, interviewing student panels, scrutinizing the learning resources provided, discussing possible sources of problems with the staff responsible, and analyzing the data resulting from the end-of-course examination.

3. Results are reported to teachers, students, and the curriculum committee that oversees the program. This report usually includes suggestions for improvement.

4. Support is provided to carry out necessary changes.

5. In the next academic year, the effects of changes carried out are measured, using the same instrument.[1]

In this article, emphasis is on the design and the measurement characteristics of the rating scale used for program evaluation in these schools. In addition, its use as a source of information for program improvement is illustrated. It should be noted that the way in which the rating scale was constructed deviates from other approaches common to the field,[2] in that a theory of how students learn in problem-based learning (PBL) and how this learning is affected by instructional procedures was used to guide the design process. Usually, a strictly empirical approach is applied in which the production of items is based on the designer's intuitive notions about what constitute important elements of the instructional process, whereas the selection of items for inclusion in the final version of the scale is conducted through exploratory statistical techniques. Such a "bottom-up" procedure takes a certain risk of resulting in a rating scale that does not reflect the important elements of the actual learning situation but rather measures mere idiosyncrasies about education that appear to be salient just because they are shared by students and designer alike. A theory-guided or "top-down" approach may avoid this confusion.

A Theory of PBL

PBL can be characterized as follows: A collection of carefully constructed problems is presented to small groups of students. These problems usually consist of a description of a set of observable phenomena or events that are in need of some kind of explanation. In medical education, they usually take the form of a description of a patient presenting a complaint and having a number of signs and symptoms. The task of the group is to discuss these problems and produce tentative explanations for the phenomena, described in terms of some underlying process, principle, or mechanism. In addition, students may be required to formulate questions, order additional laboratory information, or propose a management plan, depending on the nature of the material presented to them. Essential to the method is that the students' prior knowledge of the problem is, in itself, insufficient to understand it in depth. During initial analysis, dilemmas will arise and questions will come up that can be used as learning goals for subsequent, individual, self-directed learning.[3,4] While analyzing a problem in a prescribed, systematic fashion, the group is guided by a tutor, usually a member of the faculty. His or her task is to stimulate the discussion, to provide students with some subject-matter information whenever necessary, to evaluate progress being made, and to monitor the extent to which each group member contributes to the group's objectives. References, audiovisual aids, occasional lectures, and skills training are included as learning resources relevant to the understanding of the problems.

The program-evaluation questionnaire is based on a model proposed by Schmidt and Gijselaers,[5–7] describing the instructional and learning processes going on in PBL. Their theory has been formulated in the models-of-school-learning tradition represented by authors such as Carroll, Bloom, and Cooley and Leinhardt.[8–10] Figure 1 summarizes the important variables of the Schmidt and Gijselaers model of PBL.

The model outlined can be considered a causal and quantitative representation of the learning going on in a problem-based context. The arrows indicate the direction of the causal influence. The coefficients are regression weights, representing the strength of the causal influence. According to the model, an increase in the magnitude of one of the variables characteristically causes an increase of the magnitudes of other variables. For instance, this theory predicts that an improvement in the quality of the problems presented to students, all other things being equal, will result in improved group functioning. Better executed small-group tutorials, in turn, influence study time, which leads to higher achievement of the students involved. The role of the other variables involved can be interpreted in much the same way.[6]

In summary, it is assumed here that variables relevant to the instructional process in a problem-based curriculum generally entertain a unidirectional, causal, relationship with student achievement. An implication of this assumption, important to the issues raised in this article, is that achievement can be improved by improving the quality of these instructional variables.

However, to be able to improve on the quality of instruction, one needs to be able to measure the relevant variables in a way that takes into account the limitations of the instructional context. For instance, it is usually impossible to monitor student learning in a classroom situation for more than a short time. In the context of PBL, these limitations are even more apparent because most of the learning takes place individually. Scheduled activities usually include 4 to 6 hr a week of small-group

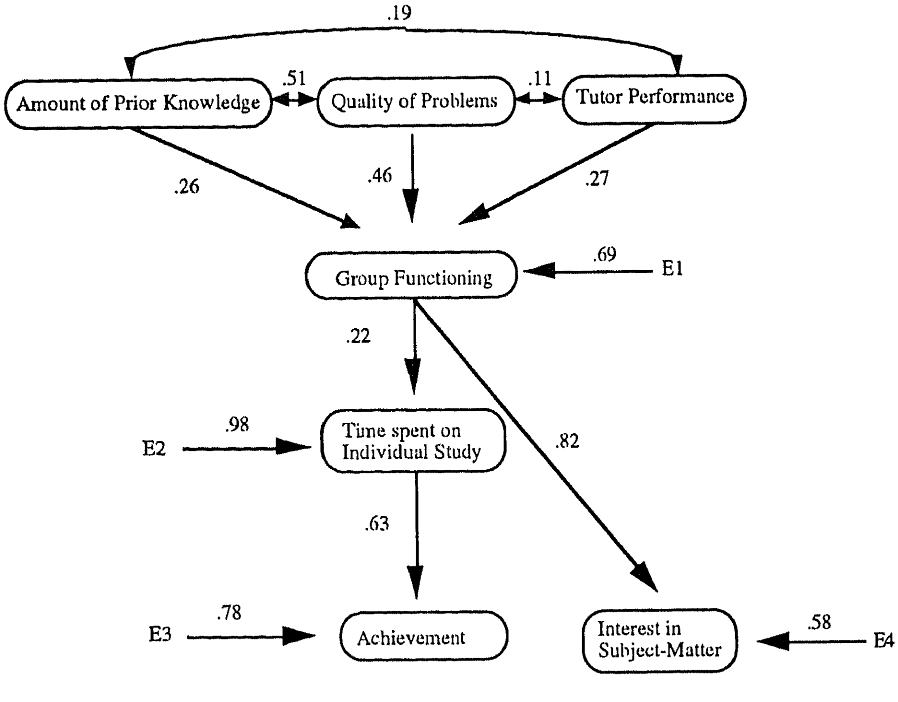

Figure 1. *Causal model of problem-based learning adapted from Schmidt and Gijselaers.*[7]

tutorials and 4 to 10 hr of lectures, skills training, and laboratories. Hence the only suitable method to carry out large-scale program evaluation is making use of students as observers of the ongoing learning activities. It has been shown repeatedly that student ratings are sufficiently reliable and valid to be used as indicators of the quality of instruction.[1,11]

Method

Subjects

Subjects were 1,800 students of the Health Sciences at the University of Limburg, the Netherlands. In addition, 286 tutors were involved in the program. Most of these tutors ran more than one tutorial group. For each 6-week course, students and tutors were both assigned to a different tutorial group in a random fashion. For the utility study, data were used from 95 1st-year students of the Faculté de Médecine of the Université de Sherbrooke, Québec, Canada.

Description of the curriculum. The 4-year Health Sciences curriculum consists of a large number of courses of equal length. Data of 98 courses taught in the academic year 1989 to 1990 were included in the analysis, which is 82% of the total number of courses. Sufficient data were unavailable from 21 courses. The courses were taught following the same general problem-based format: Students met with their tutor in small-group tutorials twice a week for 2 hr. An average of 6 hr of additional activities were scheduled, such as skills training and occasional lectures. The remaining time was spent on self-study. The Sherbrooke curriculum, from which 1987 through 1988 data were used, had similar features.

Design of the rating scale. Based on the theoretical notions concerning PBL outlined in the introduction, a 58-item rating scale was constructed (see Appendix). For each of the constructs mentioned in Figure 1, a set of items was written that covered various facets of the variable concerned (because the focus of the study was on the rating scale, the achievement variable is ignored here). Most of the items consisted of a statement and a Likert scale ranging from 1 to 5, to which the students could respond by encircling a number: 1 *(totally disagree),* 2 *(rather disagree),* 3 *(neither agree nor disagree),* 4 *(rather agree),* and 5 *(totally agree).* The Schmidt and Gijselaers[7] model was extended to adapt it to program-evaluation purposes. As can be deduced from Figure 1, Schmidt and Gijselaers chose to ignore the possible role of supporting resources in PBL, such as skills training, lectures, and literature references in their original studies. In the present study, these elements were included. It was assumed that judgment of the quality of these aspects of the learning situation would be influenced by prior knowledge of the students and the quality of the problems presented, because these two input variables would moderate the extent to which students would benefit from these resources. For instance, if students had insufficient prior knowledge, the lectures would not be so effective as they would if sufficient prior knowledge were available. In addition, it was assumed that these resources would influence output variables such as achievement and interest. A second extension was that the intrinsic interest variable was included in a somewhat broader concept "perceived relevance of learning," because the interest item (Item 5) seemed to tap the same underlying construct as a number of other items specified in the Appendix. Figure 2 displays the model tested. The Vs in the figure stand for the measured variables, and their numbers refer to the items of the questionnaire. Two variables of the original model are each represented by one item: "Amount of Prior Knowledge" (Item 2) and "Time Spent on Individual Study" (Item 57). Appendix A displays the rating scale as it was presented to the subjects. Items 19 and 20 were excluded because they require a qualitative response. Items 53 to 56 have not been submitted to statistical analyses because they inquired about individual study habits, a topic considered less relevant for program-evaluation purposes.

Procedure. After each course, the rating scale was administered to all students in the University of Limburg Health Sciences curriculum. The total number of rating scales returned was 4,757. The average response rate was 53%. Because the rating scale was filled in by

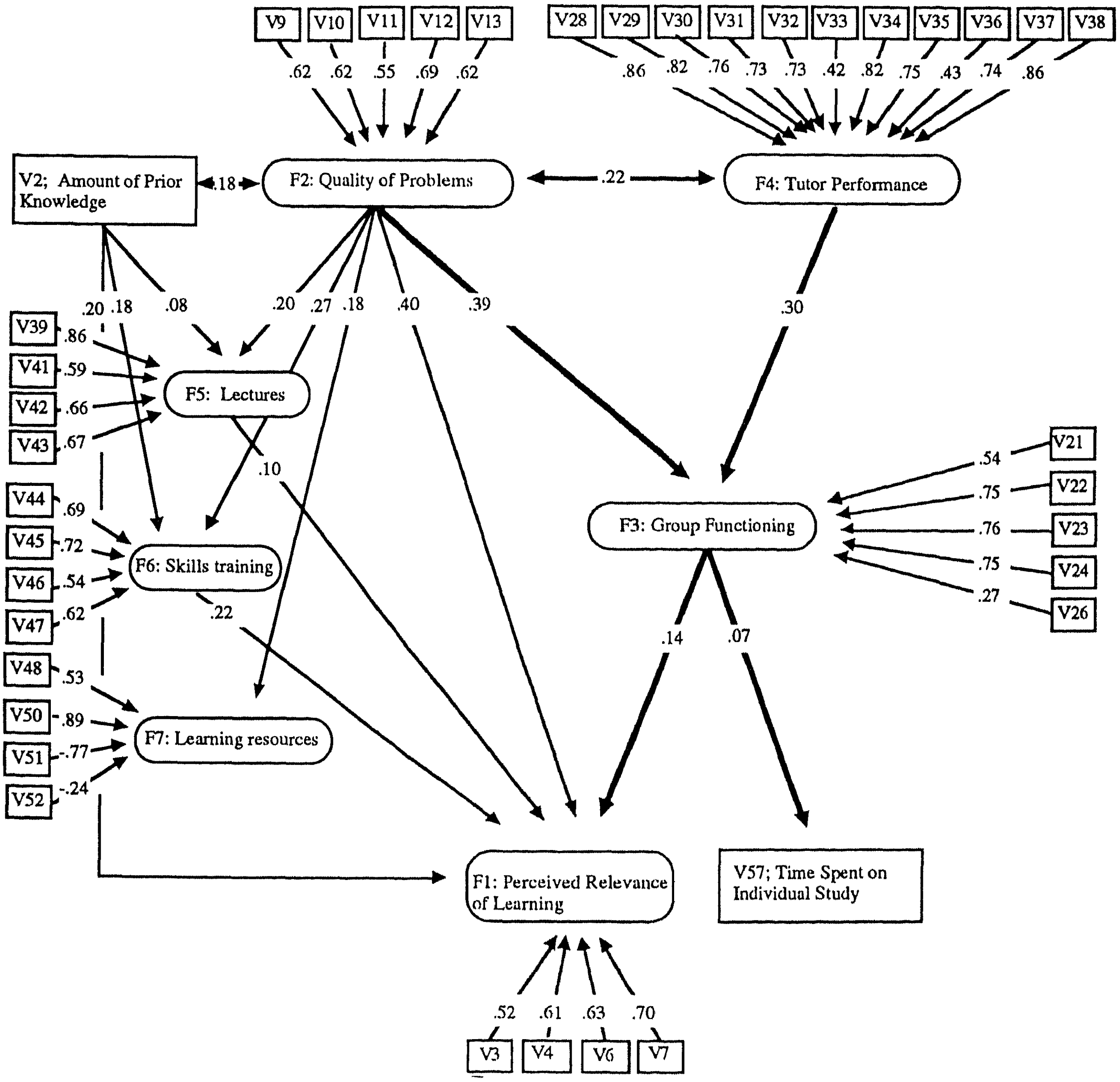

Figure 2. *Combined confirmatory factor analysis and path model of the rating scale. Latent factors are displayed in elliptic boxes, measured variables, in rectangular boxes. Only significant path coefficients between the seven factors, V2, and V57 are displayed. Causal paths predicted by the Schmidt and Gijselaers study*[7] *are displayed with bold arrows.*

most students on more than one occasion, the measurements cannot be considered independent. This may inflate results of statistical analyses carried out. Therefore, it was decided to aggregate the data at the tutorial-group level. Because subjects were randomly assigned to different tutorial groups for each course, group averages can be considered independent measurements. The total number of groups involved was 810. To evaluate the measurement characteristics and to demonstrate the possible use of the rating scale, a series of statistical and qualitative analyses was carried out. First, confirmatory factor analyses were conducted, using Bentler's structural equations approach.[12] The goal was to find out whether individual items fitted the hypothesized underlying factorial structure. This approach integrates confirmatory factor and path analysis and, therefore, was particularly suitable for our aim.

Second, generalizability studies were carried out to assess interrater agreement and other measurement characteristics of the rating scale. As indicated, the data were aggregated at the tutorial-group level, and for each variable, a mean score per tutorial group was computed. To achieve a fully balanced design convenient for generalizability studies, a random sample of four tutorial groups per course was selected from the total number of tutorial groups. If a course was rated by fewer than four tutorial groups, the particular course was excluded from the analysis. In the data set used for analysis, 50 courses were included (56.5% of the courses were excluded because of balancing). The generalizability of tutor performance, however, was studied using judgments of individual students rather than of groups. This was done because tutor performance is not a dimension that is supposed to vary by course but one that varies by group. To that end, 457 tutorial groups were selected, each having at least six students. Subsequently, for each group, six students were randomly selected to create a balanced design. The resulting data were subjected to analysis of variance.

Finally, the use of the rating scale was investigated using data from 95 1st-year students of the Sherbrooke curriculum. The goal was to illustrate how the rating-scale data can be used in analyzing various components of a curriculum, spotting weaknesses and suggesting improvements.

Results and Discussion

Confirmatory Factor Analysis

The chi-square statistic is most often used to evaluate the fit of data to a model. A nonsignificant chi-square value is considered a sign of "fit." In this study, χ^2(df = 689, N = 810) = 1,270.35, $p < .001$. Further relaxation of the model and removal of items showing relatively high residuals did not further improve the fit. These findings suggest that the model does not adequately represent the data. A problem, however, with analyses using chi-square for the evaluation of model adequacy is that this statistic is quite sensitive to violations of its distribution, particularly in relatively small samples.[12] Therefore, other statistics of fit have been developed that are less sensitive to violation of assumptions underlying the chi-square distribution. One of these statistics is the Comparative Fit Index (CFI).[12] Because the CFI takes into account attributes of the unrestricted model relative to the model under test, it is reported here. For the model tested, CFI = .99. A value larger than .90 may be considered an indicator of good fit. In addition, average standardized residuals were below .07, and the chi-square divided by the degrees of freedom is less than 2, both of which are considered indicators of reasonable fit.[13] Figure 2 graphically displays the results of the analysis.

Only the factor structure and path coefficients between the latent variables, prior knowledge, and time spent have been displayed; error terms are left out. The results suggest that the latent variables F1 to F7 explain most of the measured variables involved fairly well. Ten items had to be removed from the analyses, mainly from the quality-of-problems section. Regression weights, symbolizing the extent to which a measured variable is explained by its latent factor, are generally well over .50; 17 of 38 variables have regression weights higher than .70. Because $R^2 = 1 - E^2$, the regression weights in a standardized solution such as the one presented may be interpreted as correlations or "loadings" between the variable and its underlying factor. The path coefficients among the latent variables and between the latent variables, Prior Knowledge and Time Spent, on the other hand, are generally not impressive, although significantly different from zero. In addition, the underlying theoretical model of PBL as developed by Schmidt and Gijselaers[5-7] holds reasonably well. In conclusion, the rating scale, developed on the basis of theoretical notions, seems to capture the various elements of the learning taking place in problem-based curricula quite well.

Generalizability Studies

Generalizability studies were conducted to estimate the reliability of average tutorial group scores for Perceived Relevance of Learning, Quality of Problems, Group Functioning, Tutor Performance, Lectures, Skills Training, and Learning Resources. With the exception of the analysis of Tutor Performance, a random tutorial groups-nested-within-courses design was used, with courses as the universe of generalization. The course was selected as the object of measurement because the purpose of the whole exercise was to distinguish between courses on the various dimensions, to find out whether some courses are rated poorer on some dimension than others, and to determine how reliable these differences are. The design selected allows for variance component estimation of the following sources: (a) differences between courses, (b) differences between items, (c) differences between tutorial groups nested within courses, (d) interaction between courses and items, and (e) general error. Generalizability analyses were conducted for each separate factor. The generalizability coefficient was computed as the ratio of the true variance caused by the object of measurement to the true plus error variance. The error variance was composed of tutorial groups nested within courses, course-by-item interaction, items-by-tutorial groups within courses interaction, and random events. Thus only components influencing the ordering of courses are considered as error variance. The variance component for items was not included, because this variance does not affect the relative position of courses.

Only summary statistics are reported here (full data can be obtained from the first author). The percentage of variance associated with courses for Factor 1, Perceived Relevance of Learning, was 42.7; thus, approximately 40% of all variance in the perceived relevance of learning factor can be attributed to variation between courses. Apparently, courses can be distinguished quite well with regard to this factor. The same holds for Factor 5, Lectures; differences between courses explained 47% of the variance. For the other factors, differences between courses explained between 22.9% and 14% of the total variance. An exception was Factor 7, Learning Resources. This factor does not distinguish between courses at all.

The estimated variance components were used to estimate reliability indices. Although the interpretation of scores from the rating scale can be used in both an absolute and a relative fashion, this design design yields similar reliability estimates for both interpretation perspectives (because tutorial groups are nested within courses). Hence, all variance components were included in the observed variance definition. For Factors 1 to 5, this computation revealed a generalizability coefficient (or *G* coefficient) varying between .86 and .97, with an average group size of four tutorial groups (or six students in the case of tutor performance) rating each object. For Factor 6, the *G* coefficient was equal to .54, and for Factor 7, it was equal to zero. The *G* coefficient indicates the expected correlation between tutorial group scores derived from similar but not identical ratings, using a different random sample of tutorial groups (or students, in the case of tutor performance).

Table 1 provides the *G* coefficients as a function of the numbers of tutorial group responses and the corresponding standard error of measurement (SEM). The table indicates how many tutorial-group judgments are required to obtain a minimal generalizability coefficient of .80. (In the case of Tutor Performance, *G* is expressed as a function of number of students involved in the rating, rather than the number of groups.) Most factors have high generalizability coefficients. Exceptions are the generalizability of Skills Training and Learning Resources scores. These scales do not produce generalizable findings, at least not under our design conditions. It is not clear why these factors are less reliable. The generalizability data show, however, that both factors have fairly high interactions between tutorials groups within courses, suggesting that groups do not order courses in the same way.

The SEM also provides relevant information with regard to the reliability of the instrument. 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. Assuming that a difference between courses of .5 or greater on the 5-point scale can be considered a meaningful result, the SEM should be equal to or lower than .5 divided by 1.96, or .26. Taking into account these two requirements for reliable and sensitive measurement, one can conclude that, for Relevance of Learning, at least three tutorial groups are needed to obtain acceptable results; for Problem Quality and Group Functioning, at least two; and for Lectures, only one tutorial group is needed to obtain reliable results. One student judge is sufficient to rate a tutor reliably with our scale or an equivalent one. This is somewhat surprising because tutor behavior is supposed to be fairly variable

Table 1. *Rating Scale for Program Evaluation: Generalizability Coefficients (G) and Standard Errors of Measurement (SEM), as a Function of the Number of Tutorial Groups*

Number	Factor 1 Relevance		Factor 2 Problems		Factor 3 Groups		Factor 4[a] Tutor Performances		Factor 5 Lectures		Factor 6 Skills Training		Factor 7 Resources	
	G	*SEM*	*G*	*SEM*	*G*	*SEM*	*G*	*SEM*	*G*	*SEM*	*G*	*SEM*	*G*	*SEM*
1	.61	.28	.79	.16	.76	.15	.86	.18	.80	.22	.23	.54	.00	.87
2	.75	.20	.89	.11	.86	.11	.92	.13	.89	.16	.37	.38	.00	.61
3	.82	.16	.92	.09	.90	.09	.95	.11	.92	.13	.46	.31	.00	.50
4	.86	.14	.94	.08	.93	.08	.96	.09	.94	.11	.54	.27	.00	.34

[a] The generalizability coefficients of Factor 4, Tutor Performance, are expressed as a function of the number of student judges rather than of groups.

over sessions and, at first glance, appears quite difficult to measure. The data, however, show otherwise.

Utility

Perhaps the best way to illustrate the utility of the rating scale as an instrument for course evaluation may be to demonstrate its actual use in the evaluation of the 1st year of a problem-based curriculum as developed by the Université de Sherbrooke medical school in Québec, Canada. As an example, students' responses to items concerned with the Quality of Problems were reviewed and interpreted. The data displayed in the figures are average factor scores; thus, the results can be interpreted as scores on the same 5-point Likert scale used for the individual items. The higher the score, the more students agreed that the particular characteristic mentioned was sufficiently present. Again, 3 is the neutral point. The horizontal axis displays abbreviations for the names of the seven courses composing the first year of the Sherbrooke curriculum: Biology 1, Biology 2, Growth and Development, Nervous System, Locomotion, Mental Health, and Community Health. The figures were constructed such that comparisons between courses can be easily made. Because no formal standards exist for sufficient instructional quality in the domain of PBL (nor in other domains), comparative data can be used to provide insights into the relative strengths and weaknesses of each course.[1] In the comparisons among courses, only differences larger than .5 were considered.

As displayed in Figure 3, the courses of the school's 1st-year curriculum show fairly large differences with respect to the problems used as a stimulus for learning. Ratings vary between neutral and high.

The rule here is that problems should be rated as high as possible. Because the highest average rating (the rating for the Nervous System course) indicates which ratings could in principle have been possible for the other courses as well, the data suggest that in at least four courses, improvements regarding the nature of the problems may be necessary. The question is, of course, which improvements? A general overview like the one presented in Figure 3 does not provide answers to this question. It points at where weaknesses in particular courses may reside but does not, in and of itself, provide suggestions for remediation. In this case, it may be useful to analyze the response patterns on individual items composing the factor of interest.

Figure 4 shows average scores on the following items: "The problems were clearly stated," "The problems were suitable for using a systematic approach," "The problems sufficiently stimulated group discussion," "The problems gave sufficient opportunities for formulating learning goals," and "The problems sufficiently stimulated self-directed learning." It suggests that the problems used in different courses may not suffer similar weaknesses. For instance, the community-health course's problems lacked sufficient clarity, as compared with problems in the other courses, whereas the problems in the locomotion course gave fewer opportunities for formulating learning issues. At this point, several strategies for improvement are possible. The first is to look into the problems themselves from the perspectives provided by the ratings. Often, the problems' shortcomings present themselves quite clearly when one reads them. In those cases, simple reformulation may be sufficient. Sometimes, however, the difficulty lies not so much in the problem formulation as in the instructional context within which the problems had to be understood by the students. For instance, according to students, their tutors in the locomotion course voiced ideas about the learning goals that could not be deduced from the problems themselves. Thus when the problems themselves do not clearly

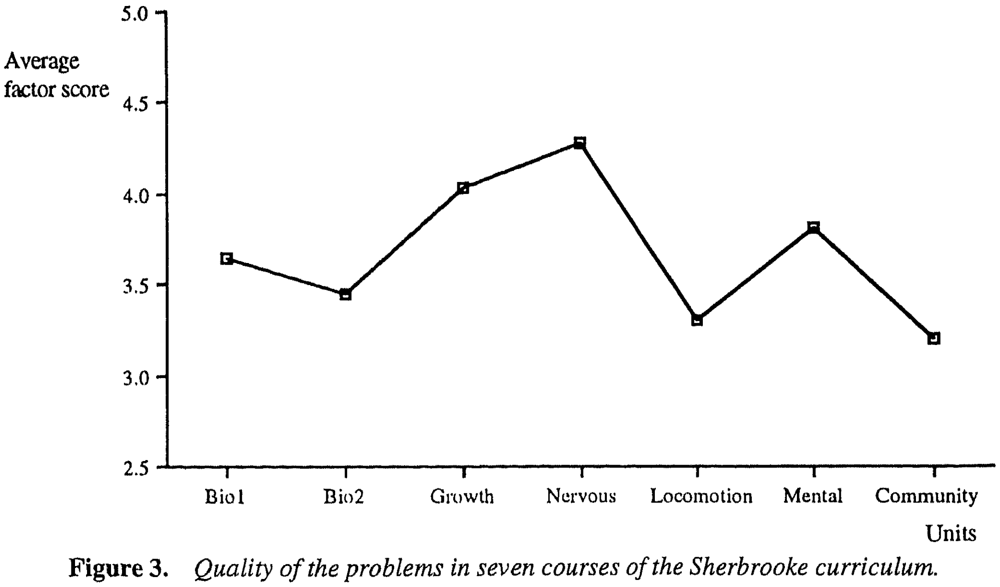

Figure 3. *Quality of the problems in seven courses of the Sherbrooke curriculum.*

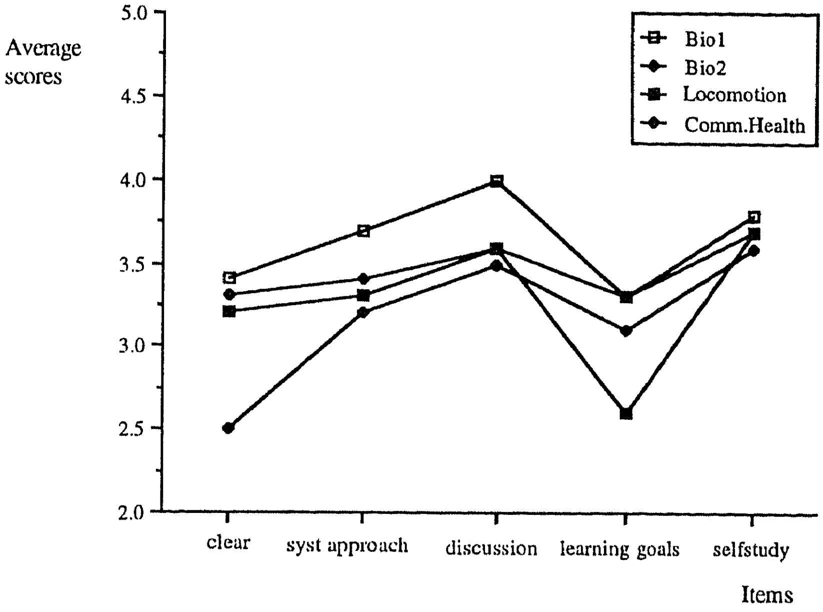

Figure 4. *Average scores of four courses of the Sherbrooke curriculum on five items constituting the quality-of-problems scale (items are displayed on the horizontal axis).*

reveal what is wrong with them, it may be useful to interview students and staff about what they see as the causes for the insufficient usefulness of the problems as a stimulus for learning.

Other dimensions measured by the rating scale displayed in Appendix A can be used to conduct similar analyses.

General Discussion

Student ratings are generally considered reliable and valid indicators of the quality of instruction, in particular when students are asked to judge elements of the instructional context that are readily observable, like teaching skill or the adequacy of instructional materials.[11] Two problems, however, limit their usefulness in everyday instructional development. The first is that student ratings are often insufficiently specific to provide guidelines for course improvement. For instance, knowing that his lectures were judged unfavorably does not help a teacher in finding ways to improve on their quality. A way to deal with this problem is to formulate descriptive statements derived from theoretical notions of what constitute important facets of the ongoing teaching and learning processes. It is, for instance, more useful to inquire about the extent to which the subject matter presented was adapted to the level of the students' prior knowledge than to ask whether they liked the teacher, because amount of prior knowledge influences the processing of new information,[14] whereas like or dislike of a teacher has no known influence on learning.

In this article, a study is reported that suggested ways in which a rating scale developed for problem-based medical curricula can provide information useful for curriculum improvement. The rating scale covered educationally important dimensions and was based on an explicit theory explaining the learning going on in problem-based curricula.

The measured variables submitted to confirmatory factor analysis appeared to cover most relevant constructs proposed in Schmidt and Gijselaers' theory of PBL.[7] In addition, most scales showed reasonable to excellent generalizability, even using no more than three average judgments. It can be concluded that in comparison to other instruments available for course evaluation, like the SEEQ,[2] the measurement characteristics of rating scale are to be considered favorable.

A second problem limiting the usefulness of student ratings in everyday instructional development is that, because no absolute standards exist for sufficient instructional quality, it is almost impossible to decide when remedial action is required with respect to a certain course. In this article, a solution to this problem based on comparisons among courses is proposed and illustrated. It was demonstrated that the highest ratings in a dimension of instructional quality can be used as a standard against which to judge the performance of other elements.[1]

Part of the present discussion is devoted to the rating scale's use in the improvement of courses. The example suggested that student ratings generally provide a simple yet informative means of detecting shortcomings in a course. A limitation of the approach should be noted here. The rating scale is a standard instrument applicable to a wide range of problem-based courses. This characteristic is both its strength and its weakness: It enables evaluators to compare across courses and put each course into the perspective of the others, but, on the other hand, it provides rather global information. It can point at where weaknesses in a course are to be found. For a detailed assessment of the nature of these weaknesses, however, further information is required, for instance, through interviewing of participants or analysis of the learning materials.

Of course, the rating scale described in this article is applicable only to those curricula that use PBL as their instructional approach. However, the evaluation strategy outlined has features that could be transplanted to other programs as well. First, the use of a theoretical perspective on learning and instruction in designing a questionnaire or rating scale seems to be mandatory if one wishes to detect and evaluate the critical components of the learning environment. Second, careful analysis of measurement characteristics contributes to the significance of findings. Too often, conclusions regarding the quality of instruction are based on questionnaires about which even elementary measurement information is lacking. And third, comparing courses on common characteristics may provide a way of circumventing the problem of absence of standards for instructional excellence.

References

1. Gijselaers WH. *Kwaliteit van onderwijs gemeten (Measuring instructional quality).* Unpublished doctoral thesis. Maastricht, The Netherlands: University of Limburg, 1988.
2. Marsh HW. SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology* 1982;52:77–95.
3. Barrows HS, Tamblyn R. *Problem-based learning.* New York, NY: Springer, 1980.
4. Schmidt HG. Foundations of problem-based learning: Some explanatory notes. *Medical Education* 1993;27:422–32.
5. Schmidt HG, Gijselaers WH. *Causal modelling of problem-based learning.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA, April 16–22, 1990.
6. Gijselaers WH, Schmidt HG. The development and evaluation of a causal model of problem-based learning. In Z Nooman, HG Schmidt, E Ezzat (Eds.), *Innovation in medical education: An evaluation of its present status* (pp. 95–113). New York, NY: Springer, 1990.
7. Schmidt HG, Gijselaers WH. Causal modelling of problem-based learning. *Instructional Science* in press.
8. Carroll JB. A model of school learning. *Teachers College Record* 1963;64:723–33.
9. Bloom BS. *Human characteristics and school learning.* New York: McGraw-Hill, 1976.
10. Cooley WW, Leinhardt G. The instructional dimensions study. *Educational Evaluation and Policy Analysis* 1980;7:7–25.
11. Cohen PA. Student ratings of instruction and student achievement: A meta-analysis of multi section validity studies. *Review of Educational Research* 1981;51:281–309.
12. Bentler PM. *EQS: Structural Equations program manual.* Los Angeles, CA: BMDP Statistical Software, 1989.
13. Saris WE, Stronkhorst LK. *Causal modelling in nonexperimental research. An introduction to the Lisrel approach.* Amsterdam, the Netherlands: Sociometric Research Association, 1988.
14. Anderson RC. The notion of schemata and the educational enterprise. In RC Anderson, RJ Spiro, WE Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 415–31). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1977.

Received 13 March 1989
Final revision received 17 March 1994

Appendix: Program Evaluation Rating Scale for Problem-Based Learning

General Impression of the Course

1. Taken together, I've worked in an agreeable way
2. The course's subject matter was adapted to my prior knowledge
3. The course's objectives were clear to me
4. The topics of this course were useful
5. The course's subject matter was difficult to understand
6. I have learned a lot during this course
7. I consider the subject of this course interesting
8. The course was well organized

The Problems

9. The problems were clearly stated
10. The problems were suitable for using a systematic approach
11. The problems sufficiently stimulated group discussion
12. The problems gave sufficient opportunities for formulating learning goals
13. The problems sufficiently stimulated self-directed learning
14. The problems helped me in integrating the basic with the clinical sciences
15. My expectations with regard to the contents of the course have been confirmed
16. I have studied independent of the course's schedule to a large extent
17. A sufficient variety of problems was available
18. I had enough time to complete the assignments
19. The following problems were poor: (fill in numbers)
20. The following problems were high quality: (fill in numbers)

The Tutorial Group

21. The tutorial group agreed explicitly on subject matter to be studied
22. Generally, everybody complied with the agreements
23. The meetings have been productive
24. Everybody actively contributed to the discussion
25. The meetings stimulated self-directed learning activities
26. The group discussion hardly influenced my choice of topics to be studied
27. I found the atmosphere in my group agreeable

The Tutor

28. The tutor displayed a fair understanding of this course's objectives
29. The tutor displayed knowledge of the principles underlying problem-based learning
30. One had the impression that the tutor liked his or her role
31. The tutor encouraged us to work hard
32. The tutor's questions stimulated the discussion
33. At regular intervals, the tutor evaluated with us the group's functioning
34. The tutor appeared to be sufficiently knowledgeable with respect to course's topics
35. The tutor used his subject-matter knowledge to help us
36. He intervened in ways that disturbed the progress of the group discussion
37. The subject-matter contributions of this tutor were relevant
38. Taken together, the tutor played his role well

The Lectures

39. The lectures provided structure to the course's subject matter
40. The topics treated were difficult to understand
41. The lectures linked up with the topics I studied
42. Generally, the topics were presented in a clear fashion
43. The lectures have been an indispensable part of this course

The Skills-Training Programs

44. The training in professional skills linked up well with the course's theme
45. The training in professional skills was offered in an instructionally sound fashion
46. The training in professional skills fitted within the time frame of the course
47. I think that the training in professional skills is relevant for this curriculum

The Learning Resources

48. I borrowed books and journals from the library regularly
49. The learning resources that I wished to consult were available sufficiently
50. I have only consulted the articles suggested by the staff
51. Because of the articles suggested by the staff, I was not encouraged to look for reading myself
52. The articles suggested by the staff were relevant for the various problems

Study Behavior

53. In view of the end-of-unit test, I confined myself to studying the literature suggested by the staff
54. The learning goals produced by the tutorial group were restricted to topics we thought would be part of the end-of-unit test
55. To get an impression of the topics and the difficulty of the end-of-unit test, I have studied tests from previous years
56. When the test date approached, I started spending more time in preparing for the test and less time on issues agreed on in the tutorial group

Open Questions

57. How much time on the average did you spend each week on independent study? (Fill in the answer in whole hours)
58. If you had to mark this course's program on a scale from 1 to 10 (6 is sufficient), what mark would you assign to this course?