COURSE CONTENT COVERAGE AS A MEASURE OF INSTRUCTIONAL QUALITY

DIANA H.J.M. DOLMANS
WIM H. GIJSELAERS
HENK G. SCHMIDT
University of Limburg, the Netherlands

In order to measure instructional quality, several methods have been proposed in the literature, among them student performance on achievement tests and student ratings of the quality of instruction. Both methods have in common that they tend to ignore course content coverage, although this is an important determinant of instructional quality. In this article a procedure is described which is used to assess students' actual learning activities. This procedure, the Topic Checklist, makes use of student ratings. Reliability, validity, and utility studies were conducted. The results suggest that the Topic Checklist is a reliable and fairly valid procedure to evaluate course content coverage and to detect problem areas in a course, providing feedback useful for carrying out improvements.

AUTHORS' NOTE: Parts of this article were presented in April 1991 as a paper, *Course Improvement Based on Course Content Data: An Explorative Study Conducted in a Problem-Based Curriculum* (Report No. TM 016 684), at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 334 220.)

EVALUATION & THE HEALTH PROFESSIONS, Vol. 16 No. 4, December 1993 448-471

Assessment of instructional quality in higher education is usually based upon two methods: (a) the measurement of educational outcomes, expressed as students' performances on various tests; and (b) measurement of the educational process by making use of student ratings. Each method has its own strengths and shortcomings. The first method focuses on measuring learning outcomes by means of achievement tests. Student performances on achievement tests are assumed to be a direct measure of student mastery of the course objectives. Scores on achievement tests reflect the actual knowledge of a group of students or an individual student in the subject matter domain. As such, they provide evaluative information on whether the educational goals are attained.

The second method focuses on the educational process rather than on educational outcomes. The most widely used method in this field takes the form of a questionnaire through which students are asked to rate several aspects of the program, such as adequacy of instructional materials and teaching skills. In this tradition it is assumed that the quality of a program influences educational effectiveness. As a consequence, evaluation should focus on measuring and improving the quality of various aspects of the program. This evaluation method is commonplace in many institutions of higher education. According to Marsh (1984), the purposes of this evaluation are to provide: (a) diagnostic feedback to faculty about their effectiveness, (b) a measure of teaching effectiveness to be used in promotion decisions, (c) information for students in the selection of courses, and (d) evaluation of the process for research on teaching. Several studies indicate that student ratings provide reliable and valid information on the quality of instruction (Costin, Greenough, & Menges, 1971; Feldman, 1977; Marsh, 1982). Besides, student ratings provide information for diagnosing instructional strengths and weaknesses which can be used to improve the program. As such, student ratings have had a positive impact on quality of teaching in higher education (Murray, 1990). In summary, student ratings are assumed to be an indirect measure of educational quality, because it is generally accepted that quality of instructional materials and lectures positively influences student learning, which in turn influences student mastery of course objec-

tives. Evaluation based on achievement test scores, on the other hand, is assumed to be a direct measure of educational effectiveness.

Despite the obvious advantages of using test scores for evaluating educational effectiveness, achievement tests have certain shortcomings for evaluation purposes. One of the major criticisms is that we do not know what factors might have contributed to or hindered the achievement of the course objectives. Low test scores on (parts of) the test may be due to various deficiencies which cannot easily be discerned. For example, poor performances may be a result of: (a) lack of effort by students, (b) difficulty level of items, or (c) inadequate match between course content and test content. First, lack of effort implies that students have spent less time studying the course content, which in turn influences the degree to which they master the course objectives, reflected by low test scores. In addition, low scores may be caused by test items that are too difficult to be answered correctly. Students might have spent a lot of time on studying the test content, but the high difficulty level of the items results in low scores. Finally, poor performances on specific parts of the test could indicate that the subject matter tested does not adequately correspond with the objectives set for the course. In this case, students are tested about topics not included in the course objectives. Besides, test items, although reflecting the course objectives, may not correspond with the instructional materials used, or with topics addressed during instruction (Schmidt, Porter, Schwille, Floden, & Freeman, 1983). In such cases, students did not have the opportunity to learn the information tested and are being penalized for the failures of instruction. Naturally, the same arguments hold for high performance scores. In summary, it can be concluded that test performance is a questionable measure to evaluate instructional quality because the source of performance cannot easily be discerned.

Although student ratings are accepted as reliable and valid measures, they also have certain shortcomings. One of the major objections is that these ratings do not reflect how well students have learned or have achieved the educational goals (Rotem & Glasman, 1979). For example, a certain lecture may be rated highly, although the content of the lecture did not correspond with the course objectives. This implies that the educational goals may not be realized. Although instruction

and materials used might be adequate, this will not guarantee that the course objectives are achieved by students. In other words, highly rated instruction does not compensate for missing content. This notion is confirmed by research conducted by Cooley and Leinhardt (1980).

Cooley and Leinhardt (1980) conducted a study, the Instructional Dimensions Study, which attempted to identify effective classroom processes. They investigated a model of classroom processes identifying variables that attempt to explain the variation in student performances. These variables or constructs include: opportunity to learn, motivators, instructional events, and structure. Opportunity to learn deals with how time is spent in classrooms and the overlap between test and curriculum. This variable consists of two measures: (a) teachers' estimates of what was covered during a certain period, and (b) a result of content analysis of the curricular materials used in the classroom. Motivators consist of aspects encouraging learning. Instructional events include the content, frequency, quality, and duration of instruction. Structure considers the organization of the curriculum. The results of their study revealed that the variable opportunity to learn correlated most highly with achievement gain. They concluded that students perform better on a test if they have been taught the materials covered by the test. An implication of the importance of opportunity to learn is, according to Cooley and Leinhardt (1980), that evaluation studies should include information on how time is spent and the degree of overlap between what is tested and what is taught.

In summary, methods assessing educational quality focus on either educational outcomes or educational processes. Both methods have in common that they tend to ignore course content coverage as an important determinant of instructional quality. However, ignoring what is actually covered during a course seems to be in contradiction with the generally accepted finding that emphasizing different curricular content produces correspondingly different patterns of achievement (Madaus & Kellaghan, 1992). As Cooley and Leinhardt (1980) showed, opportunity to learn is the strongest and most consistent predictor of achievement. As a consequence, evaluation of educational effectiveness should also focus on what actually happens during instruction or what content is covered during instruction.

Indeed, several attempts have been made to develop evaluation procedures considering course content coverage. Schmidt et al. (1983) describe a procedure measuring curricular content which consists of constructing a detailed taxonomy. Test items and curricular materials are examined and classified according to the taxonomy. The result is a visual representation of the areas covered. Next, the distribution of topics across the categories of the taxonomy is compared. These researchers conclude that the taxonomy is useful for classifying achievement tests and curricular materials in terms of their overlap. Leinhardt and Seewald (1981) investigated two other approaches measuring overlap between what is tested and what is taught: (a) instruction-based estimates and (b) curriculum-based estimates. The instruction-based measure consists of teachers' estimates of students' opportunity to learn the content tested by test items. Teachers were asked to estimate the percentage of students who had been taught the minimum material necessary to pass each item of an achievement test. Leinhardt and Seewald concluded that teachers' estimates are an important variable in predicting test performance. Curriculum-based measures of overlap use a computer-based curriculum analysis technique. Each item of a test is analyzed to assess what kind of information is needed to answer the item correctly. A dictionary of test relevant information is then constructed. In addition, a dictionary of the content of curricular materials provided to students is constructed. The dictionaries are next matched to determine what percentage of the test had been covered through curricular materials (Leinhardt & Seewald, 1981). This procedure is often referred to as content analysis (Holsti, 1969; Krippendorf, 1981; Weber, 1985). Leinhardt and Seewald (1981) concluded that curriculum-based estimates and instruction-based estimates do equally well in predicting test performance. In summary, the results of these studies provide evidence for the importance of course content coverage when assessing instructional effectiveness. However, the primary purpose of the procedures described thus far is to select the test that best reflects the curriculum. As a consequence, they merely focus on how much of the test was covered by instruction, not on how much instructional time was actually spent on the intended course content. Besides, these procedures are too

time-consuming and too costly to implement in program evaluation (Leinhardt & Seewald, 1981).

The purpose of the present study is to investigate an evaluation procedure including course content coverage as an important educational variable. In particular, attention is paid to time actually spent by students on the intended course content. Data are collected at the student level for several reasons. First, students are the primary recipients of instruction. As such they provide an important and unique perspective to judge course content coverage (Braskamp, Brandenburg, & Ory, 1984). Second, time needed in learning differs across students, depending on student aptitude, quality of instruction, and student ability to understand instruction (Carroll, 1963). Third, students differ in time they are willing to spend in learning (Gettinger, 1984). These differences among students require that information about course content coverage be collected at the level of the individual student. In order to guarantee successful implementation of such a procedure in program evaluation, the procedure should not be time-consuming or costly, should be easy to implement in educational settings, and should provide suggestions for educational improvement. Questionnaires on which students rate what content was covered and how much time was spent seem to meet these requirements. First, student ratings are easy to collect, not very time-consuming, and fairly effective in providing results that can be used to make instructional improvements (Rotem & Glasman, 1979). Second, student self-assessments appear to be at least as good as predictors of academic performance as other assessment methods, as claimed by Shrauger and Osberg (1981).

The procedure described in this article makes use of a Topic Checklist (TOC), a list of topics which reflect the intended course content. These topics were derived from the intended course content. Because the intended course content differs across courses, a new list of topics needs to be generated, for each course, following the procedure described above. The list of topics can be seen as a blueprint of the course content. At the end of the course, students are asked to rate to what extent they mastered each topic and how much time they spent studying each topic. These data provide detailed information about course content covered by students. If students fail to master certain

topics, instructors should ask themselves how well they covered this content. As such, these scores may also provide feedback suitable for carrying out improvements.

Which subject matter is actually covered by students during a course is especially important in a new approach to professional education such as problem-based learning (Barrows & Tamblyn, 1980).[1] Students in a problem-based curriculum select their own topics for study and decide for themselves how much time will be spent on studying each topic. As a consequence, students' actual learning activities may not cover the curriculum content expected to be covered by teachers. Therefore, it is obvious that problem-based programs need adequate evaluation instruments. These instruments should provide information about what content is covered by students during a course and provide useful feedback to carry out instructional improvements.

The purpose of the present study is to examine the TOC's reliability and validity. In addition, the practical utility of the TOC in improving educational quality in the setting of a problem-based medical curriculum will be illustrated.

METHOD

SUBJECTS

The study was conducted at the medical school of the University of Limburg, the Netherlands. The first 4 years of the problem-based curriculum are structured as a series of 6-week courses. In total, 142 students in the 1991-1992 academic year attended a 6-week course on normal pregnancy, delivery, and normal child development. The course is organized around 12 problems focusing on subject matter domains related to normal child development, including childbirth, vaccination schemes, psychological and social aspects of child development, normal rates of child growth, normal stages in secondary sexual characteristics, puberty, adolescence, and psychosexual development. Thus, these 12 problems reflect the course content and can be seen as instructional subunits.

MATERIALS

At the beginning of the course, topics were derived from the 12 problems reflecting the intended course content. The TOC contained a list of 144 topics covering the course content as intended by the teachers. Examples of topics related to a problem about the development of the fetus are: blood circulation in the fetus, oxygen supply of the fetus, influence of smoking on the fetus's body weight, and organ development of the fetus. In addition, eight topics not directly related to the course content were included to estimate response set effects. Thus, the total number of TOC topics was 152. For each topic a Likert-type question was formulated. Students were asked to indicate whether they mastered each topic *not at all* (1), *insufficiently* (2), *reasonably well* (3), *sufficiently* (4), or *well* (5). Second, they had to indicate whether they had spent *no time at all* (1), *little time* (2), *a reasonable amount of time* (3), *much time* (4) or *very much time* (5) on studying each particular topic. The first question was intended to measure students' perceptions about the degree to which they mastered the subject matter specified by the topic. The second question intended to measure students' subjective perception about time spent learning a particular topic.

Furthermore, an end-of-unit examination was administered to the students. The end-of-unit examination included 179 items of the true-false type and was developed by teachers responsible for the development of the course. Students' scores on this achievement test consisted of the percentage of items correctly answered. The coefficient alpha for this achievement test was .86, indicating that the internal consistency was high.

In addition, at the end of the course students were asked to estimate the number of hours per week actually spent on study during the course.

PROCEDURE

At the end of the unit, students were required to fill in the TOC. Subsequently, an achievement test was administered to the students.

The number of students participating in this study was 98 out of 142, a response rate of 69%.

RESULTS

RELIABILITY

Generalizability studies were conducted to estimate the TOC's reliability. One of the advantages of generalizability theory over classical test theory is that it recognizes multiple sources of error, such as differences among students and differences in topics, instead of only a single undifferentiated error component (Brennan & Kane, 1979). Generalizability theory is based upon analysis of variance (Crick & Brennan, 1983). The TOC consists of a list of 144 topics answered by 98 students. In terms of generalizability theory, this is a design in which topics are crossed with students (Brennan & Kane, 1979; Crick & Brennan, 1983). The object of measurement is topics, because the purpose of the TOC is to evaluate course content coverage. This implies that differences in topic scores are the most important source of variability to be identified. Both the student sample and the topic sample are treated as random, because they both are considered to be exchangeable with any other sample of the same size drawn from the universe (Shavelson & Webb, 1991). According to generalizability theory, students are considered to be the facet. As a consequence, this is a one-facet study. This design has four sources of variability: (a) differences in topics (object of measurement), (b) differences among students, (c) differences arising from educational and experiential histories of students or topic by student match, and (d) error effects or unidentified events such as student attention. In a single-facet design, the third and fourth component cannot be disentangled (Shavelson & Webb, 1991). In Table 1, the sources of variability are summarized for both mastery rating scores and time rating scores. The first source of variance in this table involves topics, and the second source involves students. The third source of variance in this table consists of the topic by student interaction and unidentified sources of variance or error.

TABLE 1
Results from G Study: Topic by Student (T*S) Design

Source	df	*Estimated Variance Component*	*Standard Error*	*Percentage of Total Variance*
Mastery scores				
Topics (T)	143	0.2957	0.0356	25.0
Student (S)	97	0.1443	0.0212	12.2
TS, error	13871	0.7407	0.0089	62.7
Time scores				
Topics (T)	143	0.2614	0.0315	24.6
Student (S)	97	0.1752	0.0255	16.5
TS, error	13871	0.6244	0.0075	58.9

The estimated variance component for the topic source reflects the magnitude of error in generalizing from a topic score to the universe of topic scores. The standard error indicates the accuracy of the estimated variance component. The standard error is relatively low because of the large number of topics (144) and students (98). As can be seen in the last column of Table 1, the percentage of variance associated with topics is 25.0% for the mastery rating scores and 24.6% for the time rating scores. This percentage is the true variance or the variance of interest. This percentage is higher than the percentage of variance for students, 12.2% and 16.5%, respectively. The largest effect, however, is the topic by student interaction effect and the error effect, 62.7% and 58.9%, respectively. The interaction effect indicates that the relative standing of topics changes from student to student. In other words, topics are ordered differently by different students. This result is comparable with findings in studies about testing in which the error component also proved to be a large source of variance.

The estimated variance components presented in Table 1 can be used to compute reliability indices. Because the ultimate concern of the TOC is to draw inferences about topics with reference to other topics, or about relative decisions, the reliability coefficient can be computed by means of the fraction between true variance or variance

caused by the object of measurement and the true variance plus error variance. This error variance should be divided by the number of students. The error variance in this study is composed of the topic by student interaction and the random events. In other words, only components influencing the ordering of topics are included. The variance component for students is not included in this computation, because this component does not affect the relative standing of topics. The student component only indicates whether students vary in degree of content covered, averaging over all topics (Shavelson & Webb, 1991). For the mastery rating scores, this computation revealed a percentage of .98 and for the time rating scores also .98. This generalizability coefficient indicates the correlation between the topic scores on this TOC and any other randomly selected TOC including 144 topics referring to the same course content.

As already mentioned, the estimated variance components can be used to compute reliability indices. On the basis of these components, the number of students needed to obtain reliable ratings of content coverage as measured by the TOC can also be determined. In generalizability theory this is called a decision study. The results of this study are shown in Table 2. Table 2 represents the reproducibility of TOC scores as a function of the numbers of student responses (generalizability coefficient G) and the corresponding standard error of measurement (SEM). The generalizability coefficient indicates how many students are required to obtain a minimal G of .80. SEM also provides important information with regard to the reliability of the TOC scores. The SEM can be used to estimate confidence intervals for particular scores. For example, the 95% confidence interval of a score can be estimated by multiplying the SEM by 1.96 (Ferguson, 1981). Starting from the assumption that a difference of at least .5, on the 5-point scale, between topics is required to obtain practical significant differences between topic scores, the SEM should be lower than or equal to 0.26 (.5/1.96) at the level of 95%. Taking into account this practical significance level, at least 15 students are required to obtain reliable results.

The reliability results presented thus far are computed at the topic by student level. As already described above, topics were derived from the 12 problems reflecting the course content. Consequently, topics

TABLE 2
Reproducibility of TOC Scores as a Function of the Numbers of Student Responses (generalizability coefficient G) and the Corresponding Standard Error of Measurement (SEM)

	Mastery Scores		*Time Scores*	
Number of Students	*G*	*SEM*	*G*	*SEM*
5	0.67	0.3849	0.68	0.3534
10	0.80	0.2722	0.81	0.2828
15	0.86	0.2222	0.86	0.2040
20	0.89	0.1924	0.89	0.1767
30	0.92	0.1571	0.93	0.1443
98	0.98	0.0869	0.98	0.0798

could be clustered around problems. This made it possible to carry out the analysis described above at the problem level. The number of topics corresponding with each problem ranges between 6 and 29, as shown in Table 3. Table 3 also contains a summary of the mean scores and standard deviations for each problem for the entire group of students.

In order to carry out this analysis for each problem, average cluster scores were computed at the student level. In terms of generalizability theory, this is an all random clusters-crossed-with-persons design. The object of measurement in this study is problems or clusters. Both the student sample and the cluster sample are treated as random. In Table 4, the sources of variability are summarized for both mastery rating scores and time rating scores. As can be seen in the last column of Table 4, the percentage of variance associated with clusters is 21.1% for the mastery rating scores and 20.7% for the time rating scores. This percentage is the true variance or the variance of interest. This percentage is lower than the percentage of variance for students, 35.3% and 41.4%, respectively. The largest effect for the mastery scores is the cluster by student effect and the error effect, 43.6%. The largest effect for the time scores is the student effect, 41.4%.

The estimated variance components in Table 4 can be used to compute reliability indices, following the same procedure as described above with respect to the topic by student level. On the basis of these

TABLE 3
Mean Scores and Standard Deviations for Each Problem for Both the Mastery Scores and the Time Scores

		Mastery Scores		*Time Scores*	
Problem	*Topics (n)*	*Mean*	SD	*Mean*	SD
1	29	3.31	.41	2.70	.45
2	11	3.45	.53	2.92	.60
3	9	3.34	.60	2.61	.56
4	13	3.79	.44	3.18	.56
5	21	3.07	.48	2.49	.46
6	5	3.49	.53	2.97	.58
7	5	3.43	.68	2.94	.70
8	15	3.47	.54	2.94	.57
9	16	3.15	.61	2.51	.59
10	7	3.29	.53	2.47	.49
11	7	3.65	.54	2.95	.64
12	6	2.68	.68	2.16	.58
Total	144	3.34	.39	2.74	.43

TABLE 4
Results From G Study: Cluster by Student (C*S) Design

Source	df	*Estimated Variance Component*	*Standard Error*	*Percentage of Total Variance*
Mastery scores				
Clusters or problem (C)	11	0.0815	0.03264	21.1
Student (S)	97	0.1367	0.02144	35.3
CS, error	1067	0.1688	0.00730	43.6
Time scores				
Clusters or problem (C)	11	0.0845	0.03376	20.7
Student (S)	97	0.1688	0.02583	41.4
CS, error	1067	0.1548	0.00670	37.9

components, the number of students needed to obtain reliable ratings of content coverage as measured by the 12 problems can be determined. These results are shown in Table 5.

TABLE 5
Reproducibility of Cluster or Problem Scores as a Function of the Numbers of Student Responses (generalizability coefficient G) and the Corresponding Standard Error of Measurement (SEM)

	Mastery Scores		*Time Scores*	
Number of Students	*G*	*SEM*	*G*	*SEM*
5	0.71	0.1836	0.73	0.1760
10	0.83	0.1299	0.85	0.1244
15	0.88	0.1061	0.89	0.1016
20	0.91	0.0919	0.92	0.0880
30	0.94	0.0750	0.94	0.0718
98	0.98	0.0414	0.98	0.0398

The G indicates how many students are required to obtain a minimal G of .80. The SEM can be used to estimate confidence intervals for particular scores. Starting from the assumption that a difference of at least .5, on the 5-point scale, between average cluster scores is required to obtain practical significant differences between problems, the SEM should be lower than or equal to 0.26 (.5/1.96) at the level of 95%. Taking into account this practical significance level, even a number of five students is sufficient. However, a sufficient reliability coefficient requires about 10 students.

In summary, at the cluster by student level, fewer students are required to obtain reliable results than at the topic by student level, whereas the variance of interest is lower and the variance of students is higher at the cluster by student level. This finding is attributable to the fact that, if relative decisions are applied, the reliability coefficient does not include the variance component for students, but only the true variance and the error variance. The error variance component is smaller at the cluster level than at the topic level, at the cost of the student variance component.

VALIDITY

In this section the TOC's content validity and criterion validity will be assessed. Ebel (1983) states that the basis for the content validity of any measurement lies in the rationale for its construction. In order

to assure the TOC's content validity, the process of sampling topics from the course content should be based on rational considerations. This process will be described. As mentioned earlier, the TOC topics were derived from the 12 problems. The process of generating topics consists of two stages. First, for each problem, the researcher specified a list of topics reflecting the intended course content. Second, this list of topics was presented to three teachers responsible for the development of the course. They first individually assessed whether these topics had been intended when they were constructing the course. Topics which were not intended according to all three teachers were removed and topics missing from this list were added. In the few cases in which the teachers disagreed—one teacher considered the topic as important to be listed, and the others failed to do so—the topic was discussed by the teachers to reach consensus. Because these teachers were responsible for the development of the course, they were assumed to be experts in examining which topics were to be studied by students. As a consequence, the TOC's content validity seems to be guaranteed.

In order to examine the TOC's criterion validity for each student, average TOC mastery and time scores of each problem were computed. In addition, the total average TOC time and mastery score of all 12 problems together were computed for each student. These total average scores for each student were correlated with the average percentage of test items correctly answered by each student and the number of hours spent on study during the course by each student. As already described in the materials section of this article, at the end of the course an achievement test was administered to the students and all students were asked to estimate the actual number of hours per week spent on study during the course. Based on these scores at the student level, several correlation coefficients were computed. Because not all data were available for all students, the number of students included in the analysis varies. The coefficients are summarized in Table 6. First, the correlation coefficient between total average TOC time scores and total average TOC mastery scores for each student is .61 ($p < .001$, $N = 98$), indicating a strong to moderate relationship between the two concepts according to students' subjective feelings.

Second, a relationship would be expected between total average TOC time scores and the average number of hours spent on study during the course. The total average time spent during the course by individual students varied between 3 and 35 hours per week; the total average time spent on study for all students was 18.72 hours per week ($SD = 6.75$). The correlation coefficient between the total average TOC time scores and the estimated hours spent during the course for each student is .42 ($p < .001$, $n = 64$), as shown in Table 6. This coefficient indicates that both measures evaluating the actual time spent during the course correspond moderately with each other. At the level of individual problems, this correlation coefficient varies between .16 (n.s., $n = 64$) and .50 ($p < .001$, $n = 64$).

Third, correlation coefficients were calculated between the total average TOC scores of each student and the corresponding scores on the achievement test. The percentage of items correctly answered by individual students varied between 32 and 70. The total average percentage of items correctly answered for all students was 53.9 ($SD = 7.67$). The correlation coefficient between the total average time spent on the TOC and the test scores is .22 ($p < .05$, $n = 94$). This coefficient is also shown in Table 6. At the level of individual problems, this correlation coefficient varies between .01 (n.s., $n = 94$) and .36 ($p < .001$, $n = 94$). The fairly low correlation between time spent on the TOC and student achievement is either due to the procedure used in this study or to the questionable quality of the achievement test. In this particular case, the achievement test scores seem to be questionable, because low correlation coefficients were found between time spent on some problems and their corresponding test scores. These results suggest that the achievement test does not adequately cover course content. TOC topics related to some problems seem to reflect the intended course content more adequately than TOC topics related to some other problems.

Because both TOC mastery scores and achievement test scores intend to measure students' mastery of the course content, a relationship would also be expected between them. The correlation coefficient between the total average TOC mastery score and the average test score for each student is .35 ($p < .001$, $n = 94$), as shown in Table 6. With regard to the statements made above about the questionable

TABLE 6
Correlation Coefficients Between TOC Time and Mastery Scores, Achievement Test Scores, and Time Spent on Study

	TOC Time Score	*Test Score*	*Mastery Score*	*Hours Spent*
TOC time score	1	.22*	.61**	.42**
Test score		1	.36**	.11
TOC mastery score			1	.21
Hours spent				1

*$p < .05$; **$p < .001$.

validity of the achievement test, this fairly low coefficient is not surprising. In addition, Table 6 shows the correlation coefficient between the actual numbers of hours spent on study and test scores. The correlation coefficient between the average number of hours spent on study during the course and students' corresponding test scores is .11 (n.s., $n = 63$). This coefficient is lower than the correlation between TOC time scores and test scores .22 ($p < .05$, $n = 94$).

In summary, the correlation coefficient between total average TOC time scores and the achievement test scores is low. This result is probably due to the questionable validity of the achievement test. As a consequence, the fairly low correlation between total average TOC mastery scores and achievement test scores was not surprising. However, the relationship between TOC time scores and the actual number of hours spent on study is moderate, which seems to provide some evidence for the criterion validity of the TOC.

UTILITY

In a problem-based curriculum, problems are the starting point for students' intended learning activities. Teachers design these problems with certain topics in mind that students are expected to cover. Whether students undertake the learning activities planned by teachers is to a large extent determined by problem effectiveness. The effectiveness of a problem may be defined as the match between student-generated learning issues and the preset faculty objectives (Dolmans, Gijselaers, Schmidt, & Van der Meer, 1993). Ineffective problems

hamper the process of generating learning issues, implying that course content is not entirely covered by students.

Because the TOC topics were derived from 12 problems, it was possible to categorize the topics across the 12 problems and to compute average time scores and mastery scores for each problem. These results are shown in Table 3.

To judge whether the TOC is a useful procedure for indicating problem areas in course content, time spent on studying topics connected to a problem and the average mastery scores should differ across problems. One-way analysis was conducted to reveal these differences. The data consisted of the average time and mastery score of each problem for each individual student. The average mastery score on the TOC (144 topics) was 3.34 (SD = .39), and the average time spent was 2.74 (SD = .43). The average problem mastery scores varied between 2.68 (SD = .68) and 3.79 (SD = .44), as depicted on the left vertical axis in Figure 1. The average mastery scores differed across problems ($F(11,1146) = 25.92$, $p < .000$). The average time spent varied between 2.16 (SD = .58) and 3.18 (SD = .56). These results are also depicted in Figure 1 on the right vertical axis. One-way analysis indicated that time spent on certain topics also differed across problems ($F(11,1131) = 25.55$, $p < .000$). The average time spent on studying the eight filler topics that were included to estimate response-set effects was 1.78 (SD = .56). The average mastery score for these eight filler topics was 2.69 (SD = .54). Response-set effects seem to be less likely, because time spent was low for these eight filler topics. This finding suggests that students indeed fill out the questionnaire seriously. The total average time spent on each problem and the total average mastery scores as listed in Table 3 and Figure 1 indicate the degree to which students' learning activities cover the intended course content. Problems with relatively low average scores should be improved in order to ensure course content coverage. How curricula can actually be improved using information derived from the TOC will be illustrated.

The results in Figure 1 show that time spent on studying the different problems varies between *little time* (2) and *much time* (4). Mastery scores vary from *insufficiently* (2) to *sufficiently* (4). Problem 4 scores relatively high on both rating scales. The objectives specified

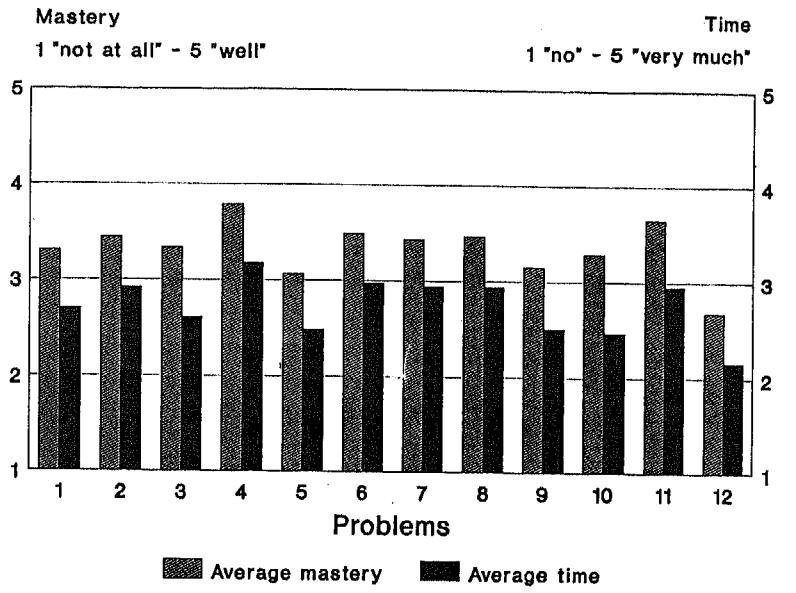

Figure 1: Average Problem Mastery Scores (Left Vertical Axis) and Average Problem Time Scores (Right Vertical Axis)

for this problem are related to different stages of delivery and risks for mother and child during delivery. Problem 12 scores relatively low on both scales. This problem deals with psychosexual development of adolescents. Problems 5, 9, and 10 also score relatively low on both scales. These problems are respectively related to the period within 10 days postpartum, psychological and social aspects of child development, and child growth. Examining rating scores on separate topics related to a low scoring problem provides information to improve the problem. For instance, Problem 12 consists of a description of a 20-year-old girl, living with her friend for several months. Because her boyfriend's former sexual preferences make her worry about their relationship, she visits her general practitioner. Scores on the TOC related to this problem revealed that students scored relatively low on DSM-III classification, psychosexual anamnesis and gender disorder of the transsexual type. Low scores on these topics would indicate that Problem 12 should be improved in order to ensure content coverage.

Adding cues to the text of the problem referring to these issues would improve its effectiveness. For instance, the following sentence could be added: What kind of diagnostic instruments are available to a general practitioner to collect information about the boyfriend's psychosexual development?

CONCLUSIONS

Instructional quality is usually assessed by two methods: (a) student performance on achievement tests and (b) student ratings of the quality of instruction. Both methods have in common that they tend to ignore course content coverage as an important determinant of instructional quality. However, several studies (Cooley & Lienhardt, 1980; Leinhardt & Seewald, 1981; Schmidt et al., 1983) provide evidence for the importance of studying course content coverage when assessing instructional effectiveness. As a consequence, evaluation of educational effectiveness should focus on what actually happens during instruction or on course content coverage.

In this article, the TOC procedure is described, yielding information about student learning activities during a course. This procedure provides information about what content is being studied by students, how much time is spent studying the intended course content, and the degree to which students master this content. Generalizability studies indicated that the TOC seems to be a reliable procedure. At the topic by student level, differences between topics or the variance of interest appeared to be higher than the variance caused by differences between students. On the contrary, at the problem by student level, the variance between students was higher than the variance between problems. A possible explanation for the relatively large effects with respect to differences between students, at the problem by student level, may be the variability between students' interest in the subject matter related to each problem or the variability between students' prior knowledge with regard to the various problems. Because of the clustering of topics around problems or subject matter domains, variability between students becomes more apparent at the student by cluster level than at the student by topic level. Furthermore, decision studies indicated that at

the topic by student level, a minimum of 15 students is required to obtain reliable results. At the cluster level, 10 students are enough to provide reliable results. The results at the cluster level are more reliable, if relative decisions are applied, because the reliability coefficient does not include the variance component for students, but only the true variance and the error variance. The error variance component is smaller at the cluster level than at the topic level, at the cost of the student variance component.

Validity studies revealed that the TOC's content validity was assured. The TOC topics were derived from the 12 problems reflecting the intended course content. Teachers responsible for the development of the course were asked to remove topics which were not intended and add topics which were lacking. As a consequence, the topics were chosen on the basis of explicit rational considerations. Criterion validity was examined by computing several correlation coefficients between TOC scores and achievement test scores and time spent on study. First, a correlation coefficient of .61 ($p < .001, N = 98$) between total average TOC time score and total average TOC mastery score for each student indicates a strong to moderate relationship between the two concepts, according to students' subjective feelings. Second, the correlation coefficient between two independent measures asking students to estimate time spent on learning activities during the course is .42 ($p < .001$, $n = 64$). This relationship is moderate and seems to provide some evidence for the criterion validity of the TOC. Third, the correlation coefficient between TOC time scores and the achievement test scores is .22 ($p < .05$, $n = 94$). This result is in accordance with results from studies examining classroom use of time and achievement, which show only moderate correlations (Frederick & Walberg, 1980). Lack of overlap between content tested and content presented to students may account for this moderate relationship. If time spent on the TOC should be a reliable predictor of student achievement, then modifications in the test are required. The correlation between TOC mastery scores and achievement test scores was also fairly low. This result is not surprising, considering the questionable validity of the achievement test.

Furthermore, the correlation coefficient between the average number of hours spent on study during the course and student test scores

is .11 (n.s., $n = 63$), whereas the correlation coefficient between TOC time scores and the achievement test is .22 ($p < .05$, $n = 94$). Consequently, TOC time scores seem to be better predictors for student achievement than the actual number of hours spent. This finding is confirmed in a review by Anderson (1985) of the literature about the relationship between time and student achievement. Anderson concluded that, when comparing the relationship between course content overlap and achievement with time spent in number of hours and student achievement, the former relationship tends to be stronger than the latter. A possible explanation is that, when measuring course content overlap, only time spent on the intended learning activities is considered, whereas the actual number of hours spent on study may include time spent on topics not directly intended to be studied. In summary, the correlation coefficients between TOC time and TOC mastery scores and between TOC time scores and the actual hours spent are moderate. The correlation between TOC time scores and test scores is fairly low. However, only a few studies have shown impressive evidence concerning the criterion validity of tests (Ebel, 1983). The major problem is the imperfect or uncertain validity of the criterion scores, according to Ebel (1983).

Moreover, it was shown that the TOC is an appropriate procedure to indicate problem areas in a course. Problems 5, 9, 10, and 12 scored relatively low. These problems were respectively related to the period within 10 days postpartum, psychological and social aspects of child development, child growth, and psychosexual development. Studying these problems in detail and comparing them with the other problems of the unit revealed that they were related to psychological and sociological issues, such as psychological well-being of the mother postpartum, social development of children, and psychological effects of being extremely tall. Problem 4, on the other hand, which addresses physiological processes of delivery, scored relatively high. These findings correspond with faculty observations that students tend to prefer biology issues at the expense of psychological ones. Time and mastery scores on separate topics related to each problem indicate what content is covered by students and contain cues for problem improvement, as was shown in the results section of this article. In summary, the TOC seems to be a reliable procedure requiring rela-

tively few raters to yield detailed information about the nature of the weaknesses of a course. It also provides feedback suitable for carrying out improvements.

NOTE

1. The principal idea behind problem-based learning is that learning should be organized around problems related to the profession, rather than around subjects derived from academic disciplines (Barrows & Tamblyn, 1980). Problems usually consist of a set of observable phenomena or events in need of some kind of explanation and management. Students analyze these problems, attempting to understand the underlying principles or mechanisms through small-group discussion. In doing so, they activate whatever they already know about the problems. However, students' prior knowledge in itself is not sufficient to attain a deep understanding. During discussion some questions usually remain unanswered, subsequently serving as a guide for independent and self-directed learning (Schmidt, 1983).

REFERENCES

Anderson, L. W. (1985). Opportunity to learn. In T. Husén & T. N. Postlethwaite (Eds.), *The international encyclopedia of education: Research and studies* (pp. 3682-3686). Oxford: Pergamon.

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

Braskamp, L. A., Brandenburg, D. C., & Ory, J. C. (1984). *Evaluating teaching effectiveness. A practical guide*. Beverly Hills, CA: Sage.

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

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis, 2*(1), 7-25.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 41*, 511-535.

Crick, J. E., & Brennan, R. L. (1983). *Manual for Genova: A generalized analysis of variance system*. Iowa City, IA: American College Testing Program.

Dolmans, D.H.J.M., Gijselaers, W. H., Schmidt, H. G., & Van der Meer, S. B. (1993). Problem effectiveness in a course using problem-based learning. *Academic Medicine, 68*(3), 207-213.

Ebel, R. L. (1983, Summer). The practical validation of tests of ability. *Educational Measurement: Issues and Practice*, pp. 7-10.

Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education, 6*, 223-274.

Ferguson, G. A. (1981). *Statistical analysis in psychology and education* (5th ed.). Auckland: McGraw-Hill.

Frederick, W. C., & Walberg, H. J. (1980). Learning as a function of time. *Journal of Educational Research, 73*, 183-194.

Gettinger, M. (1984). Individual differences in time needed for learning: A review of the literature. *Educational Psychologist, 19*(1), 15-29.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Krippendorf, K. (1981). *Content analysis. An introduction to its methodology*. Beverly Hills: Sage.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18*(2), 85-95.

Madaus, G. F., & Kellaghan, T. (1992). Curriculum evaluation and assessment. In P. W. Jackson (Ed.), *Handbook of research on curriculum* (pp. 119-154). New York: Macmillan.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Psychology, 52*, 77-95.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*(5), 707-754.

Murray, H. G. (1990). Student instructional ratings: The impact on quality of teaching in higher education. *Teaching News, 39*, 7-13.

Rotem, A., & Glasman, N. S. (1979). On the effectiveness of students' evaluative feedback to university instructors. *Review of Educational Research, 49*(3), 497-511.

Schmidt, H. G. (1983). Problem-based learning: Rationale and description. *Medical Education, 17*, 11-16.

Schmidt, W. H., Porter, A. C., Schwille, J. R., Floden, R. E., & Freeman, D. J. (1983). Validity as a variable: Can the same certification test be valid for all students? In G. F. Madaus (Ed.), *The courts, validity and minimum competency testing* (pp. 133-151). Hingham, MA: Kluwer-Nijhof.

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

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin, 90*(2), 322-351.

Weber, R. P. (1985). *Basic content analysis*. Beverly Hills, CA: Sage.