International Journal of Teaching and Learning in Higher Education 2012, Volume 24, Number 1, 43-51
http://www.isetl.org/ijtlhe/ ISSN 1812-9129

Assessing the Quality of Problems in Problem-Based Learning

Nachamma Sockalingam
SIM University

Jerome Rotgans
National Institute of Education

Henk Schmidt
Erasmus University Rotterdam

This study evaluated the construct validity and reliability of a newly devised 32-item problem quality rating scale intended to measure the quality of problems in problem-based learning. The rating scale measured the following five characteristics of problems: the extent to which the problem (1) leads to learning objectives, (2) is familiar, (3) interests students, (4) stimulates critical reasoning, and (5) promotes collaborative learning. The rating scale was administered to 517 polytechnic students enrolled in problem-based curricula and the data collected were subjected to confirmatory factor analysis. The results revealed a good fit of the data with the hypothesized five-factor model. The coefficient *H* values of the five factors suggested acceptable factor reliability. Overall, the psychometric characteristics of the rating scale indicated adequacy of the instrument to measure the quality of problems in problem-based learning. Although there are other ways to assess problem quality, the ease of use and means to measure multiple indicators makes the problem quality rating scale useful.

The fundamental elements of problem-based learning (PBL) are problems, students and tutors (Majoor, Schmidt, Snellen-Balendong, Moust, & Stalenhoef-Halling, 1990; Williams, Iglesias, & Barak, 2008). Several studies point out that besides students' prior knowledge and tutors' performance, the quality of problems has the most significant influence on student learning (Gijselaers & Schmidt, 1990; Schmidt & Gijselaers, 1990; Van Berkel & Schmidt; 2000). Despite the significance ascribed to problems in PBL, surprisingly, there is a lack of validated instruments to measure their quality.

Problems are a set of descriptions of situations or phenomena demanding solutions or explanations, and are usually structured in textual format, sometimes with illustrations, pictures, videos, and simulations (Schmidt, 1983). In PBL, problems trigger the learning process. Problems are purported to achieve the objectives of PBL by engaging students in collaborative work and elaboration, thereby rekindling students' prior knowledge and promoting self-directed learning skills, and consequently leading to construction of new knowledge (Barrows & Tamblyn, 1980; Hmelo-Silver, 2004; Norman & Schmidt, 1992).

Generally, there are two approaches to measuring the quality of problems. One approach is to evaluate whether students are able to generate the same learning goals as intended by the curriculum. The degree of congruence between the two is considered to be reflective of problem effectiveness (Dolmans, Gijselaers, Schmidt, & Van der Meer, 1993; Mpofu, Das, Murdoch, & Lanphear, 1997). However, this method has its limitations in the sense that it addresses only one aspect of effective problems – that is, the extent to which a problem leads to formulation of the intended learning objectives. In addition, the procedure of comparing the student-generated learning goals with the faculty-intended learning objectives may be considered as time consuming and tedious. In a study by Dolmans et al. (1993), 24 expert raters were to compare a total of 51 faculty-intended learning objectives with the learning goals generated by 120 students for 12 problems. Assuming that each student comes up with five learning goals per problem, each rater would have to make 7200 comparisons for 12 problems and 120 students. To reduce the number of comparisons to be made, Dolmans et al. (1993) modified the protocol and allotted one group of 12 students (instead of 120) to each pair of raters. Although, this method provided detailed information about the extent to which a problem leads to the intended learning objectives, the practicality of the method to provide regular feedback about the quality of problems may be limited by the availability of time and resources.

An alternative approach is the administration of a self-report rating scale. To evaluate the quality of a course at the general program level, Schmidt, Dolmans, Gijselaers, and Des Marchais (1995), developed and validated a 58-item rating scale. Of the 58 items, five items measured the overall quality of all problems in the course. Considering that the measurement scope of the instrument was intended to be at the general program level, it may not be adequate in providing detailed feedback about individual problems.

Using Jonassen's theory of problem solving as a basis (Jonassen, 2000), Jacobs, Dolmans, Wolfhagen, and Scherpbier (2003) developed a 12-item rating scale to measure the complexity and structurednes of PBL problems. When the validity of the rating scale was examined by means of confirmatory factor analysis, results suggested an inadequate fit of the data with the hypothesized two-factor model. Instead, an alteration of the model from the two factor structure to a three-factor yielded a better fit. The altered model consisted of the factors: *too simple*, *too difficult*, and *too well-structured*. These factors were derived from the original two factors by splitting *complexity* into too simple or

too difficult, and *structuredness* into too well-structured or too ill-structured, subsequently combining too difficult and ill-structured to form the factor too difficult. Overall, the 12-item rating scale encompassing the three factors was concluded to be an adequate instrument to measure the two characteristics complexity and structuredness. Although the final three-factor model fitted the data reasonably well, it deviates significantly from the initially hypothesized two-factor model and raises concerns about the content validity of the rating scale, since it now measures an extra factor that seems to be conceptually different from what was initially intended.

Marin-Campos, Mendoza-Morals, and Navarro-Hernandez (2004) designed an 18-item rating scale to assess the three aspects of a PBL problem; (1) the extent to which the problem was correctly structured, (2) the extent to which the problem allowed students to carry out the expected learning activities, and (3) the extent to which the allocated time and resources were suitable for the students to work on the problem. Theoretical underpinnings of PBL (Schmidt, 1983; Dolmans, Snellen-Balendong, Wolfhagen, & van der Vleuten, 1997; Rangachari, 1998) served as the basis for the rating scale design. This rating scale was used to gather longitudinal feedback on 14 different problems from a group of 28 students. Compared to the earlier mentioned studies (Schmidt et al., 1995; Jacobs et al., 2003), this rating scale had the capability to yield more detailed feedback on individual problems. In addition, the internal consistency of the three factors seemed to be adequate when examined by means of Cronbach's alpha test. However there are two points to consider. Firstly, despite the reliability and usefulness of this rating scale to provide detailed feedback on individual problems, its validity remains to be tested. As this study involved only 28 students (from a medical course), validation involving a larger sample by means of factor analysis would still be needed. Secondly, the measurement scope of the rating scale could be extended further. For instance, various core learning activities such as identification of key learning objectives, the extent to which the problems encouraged group discussion, and interest triggered by the problem were treated as one factor (the extent to which problem allowed the students to carry out the expected learning activities). Differentiating the various learning activities is likely to provide comprehensive information about the influence of the problem on students' learning.

In summary, the two approaches used currently to assess the quality of problems are; (1) comparison of the student-generated learning goals with those intended by the curriculum, and (2) administration of a self-report rating scale to measure a selected set of problem characteristics. Both approaches have their advantages, but when it comes to practical considerations, administering a rating scale seems more feasible. Considering that the existing instruments only addressed a limited number of characteristics (i.e., two or three), we were motivated to develop and validate a more comprehensive problem quality rating scale.

To this end, we first developed a 56-item rating scale measuring eleven characteristics of effective problems in PBL. These characteristics were based on Sockalingam and Schmidt's (2007) study on students' perspectives of problems in PBL and theoretical underpinnings of PBL (e.g., Dolmans et al., 1997). Pilot testing of the rating scale showed that the data did not adequately fit the hypothesized 11 factor model and guided us in redesigning the rating scale to a shorter form of 32 items. The resulting 32-item rating scale was intended to measure the following five problem characteristics; (1) the extent to which the problem leads to formulation of intended learning objectives, (2) the extent to which the problem is familiar to students, (3) the extent to which the problem interests students, (4) the extent to which the problem promotes collaborative learning, and (5) the extent to which the problem stimulates critical reasoning. The objective of this study, therefore, was to validate and test the reliability of the 32-item rating scale. To this end, the rating scale was administered to 517 first year students at a polytechnic in Singapore. Subsequently, confirmatory factor analysis and reliability measures were carried out to examine the psychometric characteristics of the rating scale.

Method

Participants

The sample consisted of 517 participants (58% female and 42% male) with an average age of 18.69 (SD = 1.70) years. All participants were enrolled in a first year general curriculum in the academic year 2007/2008 at a polytechnic in Singapore.

Educational Context

The sole instructional method used in the polytechnic is PBL. To obtain a diploma certification, students are required to complete approximately 30 modules. To complete their course work requirement, students are encouraged to take four or five modules every semester for three years. Each module consists of 16 problems which are delivered in 16 weeks (one semester). In this approach, students work on one problem per day (Alwis & O'Grady, 2002). The typical class size is 25, in which students work in groups of five. Each class is facilitated by one tutor. The class starts with the presentation of a problem. Students discuss in their teams what they know, do not know,

and what they need to find out. In other words, students activate their prior knowledge, come up with tentative explanations for the problem, and formulate their own learning goals (Barrows, 1980; Hmelo-Silver, 2004; Schmidt, 1993). The tutor oversees the discussion. A period of self-study follows the first meeting. During the study period, students individually and collaboratively try to find information to address the learning goals (Hmelo-Silver, 2004). The class then meets again for a second meeting to discuss their findings and seek guidance from the tutor. This second meeting provides an opportunity to clarify learning goals, misconceptions and learn from each other. The class then breaks again for a second self-study period. This study period allows the students to find out more information and compile their findings. At the end of the day the teams come together as a class to present, elaborate, and synthesize their findings.

Instrument

Problem quality rating scale. We first designed a 56-item rating scale to assess eleven characteristics of effective problems. This rating scale was based on Sockalingam and Schmidt's (2007) study on characteristics of problems in PBL and theoretical underpinnings (e.g., Dolmans et al., 1997). The eleven characteristics are that problems should (1) be of suitable format (such as length of text and use of visuals), (2) be sufficiently clear, (3) lead to the intended learning objectives, (4) be familiar to students, (5) be of appropriate difficulty level, (6) be applicable/relevant (for instance, to other modules/ future work), (7) interest students, (8) promote self-directed learning, (9) stimulate critical reasoning, (10) encourage teamwork, and (11) trigger elaboration. This rating scale was piloted with 185 first year students. Confirmatory factor analysis showed the data did not adequately fit the hypothesized factor model. This is not uncommon in developing a new rating scale/questionnaire (Byrne, 2001). We then analyzed the covariance matrix for items that did not contribute significantly to the underlying factors, or were highly correlated. Items that shared higher correlation with other factors; that is items which cross-loaded were combined to form a single factor, taking the conceptual validity into consideration. For instance, three of the characteristics, (1) suitable format of problem (such as length of text and use of visuals), (2) the extent to which the problem is clear, and (3) the extent to which the problem leads to formulation of intended learning objectives were combined to form a single factor "the extent to which the problem leads to formulation of intended learning objectives." Similarly, two other characteristics; (4) the extent to which problem promotes teamwork, and (5) the extent to which problem triggers elaboration were combined to form a single factor of "the extent to which the problem promotes collaborative learning." Next, items that did not contribute significantly to the underlying latent factor were dropped. This led to too few items for three of the characteristics. Given that initially these characteristics were only represented by four items, the three characteristics had to be excluded. The excluded characteristics were (6) the extent to which the problem promoted self-directed learning, (7) difficulty level of the problem, and (8) the extent to which the problem is applicable/useful. The remaining three characteristics of effective problems, (9) the extent to which the problem is familiar to students, (10) the extent to which the problem interests students, and (11) the extent to which the problem stimulates critical reasoning, were considered to be unique and were used as individual factors in the rating scale. This resulted in a 32-item rating scale, measuring five characteristics of the problems. The five factors of the rating scale are (1) the extent to which the problem leads to formulation of intended learning objectives, (2) the extent to which the problem is familiar to students, (3) the extent to which the problem interests students, (4) the extent to which the problem promotes collaborative learning, and (5) the extent to which the problem stimulates critical reasoning. For details of the items, see the Appendix. All items were assessed on a 5-point Likert scale: 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree).

Procedure

The rating scale was administered electronically and participants were informed to think about the problem that they had worked on for the day (problem P11) when responding to the rating scale. Participants had fifteen minutes to complete the rating scale.

Analysis

First, the 32 items of the rating scale were parcelled, that is combined in groups of two or three based on semantic overlap (Bandalos & Finney, 2001; Little, Cunningham, Shahar, & Widaman, 2002). A total of 14 parcels were formed. Parcelling is a common measurement practice used in latent variable analysis. A parcel can be defined as the average of the two or three indicator items (Little et al., 2002). A detailed description of each of the 14 parcels, accompanied with the indicator items, is given in the Appendix. Next, descriptive statistics for all items and parcels, and correlation matrix for the five factors were generated. Subsequently, confirmatory factor analysis was carried out using AMOS 5 to examine whether the data fitted the hypothesized five-factor model (Arbuckle, 2003).

The analysis was carried out with three different types of samples: First, with an *exploration sample* (N = 209), to conduct an initial analysis of the hypothesized model, and then with a second *construct validation sample* (N = 208) to retest the model and cross-validate the second sample with the first. The cross-validation was done by means of a difference in Chi-square test (Byrne, 2001). As such, the models for the two samples were tested with both unconstrained and constrained factor loadings. Significant differences in Chi-square value between the constrained and unconstrained models in relation to the difference in degrees of freedom reveals the extent to which they differ. After the cross-validation was completed, we retested the five-factor model with the third *main sample*, which is the combined sample of the first two. For all three samples, parameter estimates were generated using maximum likelihood and tests of goodness of fit. Chi-square accompanied by degrees of freedom, sample size, p-value, root mean square error of approximation (RMSEA), and comparative fitness index (CFI) were used as indices of absolute fit between the models and the data. The Chi-square is a statistical measure to test the closeness of fit between the observed and the predicted covariance matrix. A small Chi-square value, relative to the degrees of freedom, indicates a good fit (Byrne, 2001). A Chi-square/*df* ratio of less than 3.00 is considered to be indicative of a good fit (Byrne, 2001). RMSEA is sensitive to model specification and is minimally influenced by sample size and not overly affected by estimation method (Fan, Thompson, & Wang, 1999). The lower the RMSEA value, the better the fit. A commonly reported cut-off value is .06 (Hu & Bentler, 1999). In addition to these absolute fit indices, the comparative fit index (CFI) was calculated. The CFI value ranges from zero to one and a value greater than .95 is conventionally considered a good model fit (Hu & Bentler, 1999).

Finally, Hancock's coefficient H was calculated for each of the five factors using the *main sample*. The coefficient H is a construct reliability measure for latent variable systems that represents an adequate alternative to the conventional Cronbach's alpha. According to (Hancock & Mueller, 2001) the usefulness of Cronbach's alpha and related reliability measures is limited to assessing composite scales formed from a construct's indicators, rather than assessing the reliability of the latent construct itself as reflected by its indicators. The coefficient H is the squared correlation between a latent construct and the optimum linear composite formed by its indicators. Unlike other reliability measures the coefficient H is never less than the best indicator's reliability. In other words, a factor inferred from multiple indicator variables should never be less reliable than the best single indicator alone. Hancock recommended a cut-off value for the coefficient H of .70.

Results

Descriptive statistics were calculated for the items and parcels; no outliers or other abnormalities were observed. The correlations between the five factors ranged from .29 and .65 (see Table 1).

As a next step, we tested whether the data fitted the hypothesized five-factor model. We did this for three samples, first, with the exploration sample, followed by the validation sample and finally with the main sample. The model fit statistics for all three samples are summarized in Table 2.

The results demonstrated that the data fitted the five-factor model well. The Chi-square/*df* ratio for the main sample, (N = 517), was 2.06, $p < .01$, RMSEA = .05 and CFI = .98. All factor loadings, ranging from .59 to .81, were statistically significant and thus contributed significantly to the respective latent variable. The test for invariant factorial structures revealed that there was no significant difference in the underlying factor structure between the exploration sample and the validation sample (see Table 3).

Finally, the reliability of the factor was determined by calculating Hancock's coefficient H (Hancock & Mueller, 2001). The coefficient H values ranged from .66 (critical reasoning) to .78 (collaborative learning), with an average of .75. The values are indicative of a moderate to good reliability of the rating scale. The mean values, standard deviations, as well as reliability coefficients of the five factors are presented in Table 4.

Discussion

The objective of the present study was to validate and test the reliability of a rating scale to measure the quality of individual problems in PBL. To that end, a 32-item rating scale, based on students' conceptions about five characteristics of effective problems (Sockalingam & Schmidt, 2007) and theoretical underpinnings (e.g., Dolmans, et al., 1997) was developed. The rating scale was tested with 517 first year students in Singapore context. The factor structure of the rating scale was analyzed by means of confirmatory factor analysis using AMOS 5 (Arbuckle, 2003). Results of the confirmatory factor analysis revealed a good fit of the data with the hypothesized five-factor model. The standardized regression weights of all fourteen parcels were statistically significant, suggesting that the parcels contribute significantly to the underlying latent constructs. The coefficient H values for the five factors were satisfactory and indicative of a reasonably reliability. Cross-validation of the rating scale using two samples showed that there

Table 1
Correlation Matrix of the Five Factors

Factor	1	2	3	4	5
1. Learning issue	_				
2. Familiarity	.65**	_			
3. Interest	.60**	.56**	_		
4. Collaborative learning	.47**	.29**	.39**	_	
5. Critical reasoning	.49**	.38**	.56**	.51**	_

Note. ** Correlation is significant at the .01 level.

Table 2
Goodness-of-Fit Statistics of the Five-factor Model

Sample	*N*	X^2	*df*	X^2/ *df*	CFI	RMSEA
Exploration sample	209	76.34	64	1.19	.99	.03
Construct validation sample	208	130.95	64	2.05	.94	.06
Main Sample	517	131.69	64	2.06	.98	.05

Note. CFI = comparative fit index; RMSEA = Root mean square error of approximation.

Table 3
Cross Validation of Factor Structure

Model description	X^2	*Df*	X^2_{diff}	df_{diff}	Statistical significance
Hypothesized five-factor model	207.29	128	_	_	_
Model with measurement weights constrained	214.39	137	7.11	9	NS**

Note. **Not significant at the .05 level.

Table 4
Descriptive Statistics and Reliability Coefficient of the Five Factors

Factor	Mean	*SD*	Coefficient *H*
1. Learning issue	3.24	.60	.75
2. Familiarity	2.99	.60	.77
3. Interest	3.26	.66	.77
4. Collaborative learning	3.66	.61	.78
5. Critical reasoning	3.70	.51	.66

Note. ** Correlation is significant at the .01 level.

was no significant difference in the factor loadings and hypothesized five-factor model between the two groups. In summary, the psychometric characteristics of the 32-item rating scale seemed to be adequate for measuring students' conceptions about the five characteristics of effective problems.

The five factors of the rating scale are (1) the extent to which the problem leads to formulation of intended learning objectives, (2) the extent to which the problem is familiar to students, (3) the extent to which the problem interests students, (4) the extent to which the problem promotes collaborative learning, and (5) the extent to which the problem stimulates critical reasoning.

The first factor, the extent to which the problem leads to formulation of intended learning objectives, measures whether the problem instruction is clear, whether the keywords and clues that are embedded in the problem text allow students to identify the intended learning objectives, and come up with a logical approach to address the problem. This factor, to some extent, represents Jacob et al.'s (2003) complexity, Marin-Compas et al.'s (2004) two factors on problem structure and problem allowing expected learning

activities, and addresses largely the objective of Dolmans' approach to evaluating the effectiveness of problems by means of comparing student-generated learning goals with intended learning objectives (Dolmans et al., 1993). Of course, the use of self-report measures has its limitations. The indicator items and parcels used in the rating scale may not be as exhaustive as phenomenological approach. However, considering administrative issues, use of a rating scale is far less time-consuming and more practical.

The second factor, the extent to which the problem is familiar to students, refers to students' familiarity with the context and content of the problem. The familiarity with the problem is the result of past experiences, subject-domain knowledge, and general knowledge. Inclusion of this factor in the rating scale seems reasonable considering the large body of research that suggests that prior knowledge strongly influences learning (Anderson, 1990; Dolmans, Wolfhagen, & Schmidt, 1996; Mamede, Schmidt, & Norman, 2006; Norman & Schmidt, 1992; Schmidt & Boshuizen, 1990; Soppe, Schmidt, & Bruysten, 2005).

The third factor, the extent to which the problem interests students, and the fourth factor, the extent to which the problem promotes collaborative learning, represent the same two factors as in Schmidt's general model of PBL (Schmidt & Gijselaers, 1990). In our case, however, we are more concerned about measuring the student interest and collaborative learning at the problem level to provide detailed feedback on individual problems. As such, the grain-size of our instrument is larger in order to detect differences between individual problems. Interest generated by the problem refers to the level of curiosity and engagement invoked by the problem. Collaborative learning promoted by the problem refers to the extent to which the problem triggers teamwork and elaborations such as brainstorming and discussions. This is also referred to as group functioning in PBL literature.

The fifth and final factor, the extent to which the problem stimulates critical reasoning, refers to the extent to which the problem triggers questioning, stimulates thinking and reasoning, as well as whether the problem allows for multiple solutions. The latter was referred to as structuredness by Jacobs et al. (2003). In our case, however, the fifth factor is broader, and includes questioning, thinking, and reasoning in the context of PBL problems (Kamin, O'Sullivan, Younger, & Deterding, 2001; Tiwari, Lai, So, & Yuen, 2006).

In conclusion, the five factors described above extend the measurement scope of the existing instruments. Besides the characteristics measured by the existing instruments (Jacobs et al., 2003; Marin-Campos et al., 2004; Schmidt et al., 1995), the problem quality rating scale discussed in this study includes four additional factors (The extent to which the problem is familiar to students, the extent to which the problem interests students, the extent to which the problem promotes collaborative learning, and the extent to which the problem stimulates critical reasoning). This study, therefore, may provide an instrument to measure the quality of problems in a more comprehensive manner than those available at present.

One important point to note in this study is that the administration of rating scale was post-experience; the problem quality rating scale was administered to the students after they had worked on the problem. In this case, students had retrospectively assessed the problem. Whether the rating scale could be used to predict the quality of problem remains to be tested. Given that there is communication between the students and the tutors and within the groups of students during the learning process on the content as well as the learning process (Hmelo-Silver, 2004), it is likely that the students' perceptions of the problem quality is molded by the students' learning experience with the problem. For instance, in PBL, the tutor would from time to time check on the students' progress and would feedback on the students' learning such as relevance of learning objectives, critical reasoning and collaboration as a team. The tutor would also summarize the learning objectives at the end of the lesson, which would allow students to compare their work with the faculty-intended learning objectives (Hmelo-Silver, 2004). However such indicators of student learning would be missing if students had not experienced the problem. Therefore, we feel that it would be more meaningful to collect feedback on the individual problems after students had worked on the problem (rather than before). Often, courses are evaluated at the module level (Schmidt, et. al., 1995) and this would not provide much information on which set of problems had not been effective. The problem quality rating scale would allow us to systematically collate data on various problem characteristics at an individual problem level and allow us to review the module at an individual problem-level. To further test the usability of the problem quality rating scale, future studies could look into administering the rating scale for a number of different problems from different subject domains and correlating students' assessment of the problem with their academic achievement.

References

Alwis, W. A. M., & O'Grady, G. (2002, December). *One day-one problem at Republic Polytechnic.* Paper presented at the Forth Asia-Pacific Conference on Problem-Based Learning, Hatyai, Thailand.

Anderson, J. R. (1990). *Meaning-based knowledge representations*. New York: Freeman.

Arbuckle, J. L. (2003). *Amos 5.0 update to the Amos user's guide.* Chicago, IL: Small Waters Corp.

Bandalos, D. L., & Finney, S. J. (2001). Item parcelling issues in structural equation modelling. In G. A. Marcoulides, & R. E. Schumaker (Eds.), *New developments and techniques in structural equation modelling* (pp. 269-296). Mahwah, NJ: Lawrence Erlbaum.

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

Byrne, B. M. (2001). *Structural equation modeling with Amos. Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

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. doi:10.1097/00001888-199303000-00013

Dolmans, D. H. J. M., Snellen-Balendong, H., Wolfhagen, I. H. A. P., & van der Vleuten, P. M. (1997). Seven principles of effective case design for a problem-based curriculum. *Medical Teacher, 19*(3), 185-189. doi:10.3109/01421599709019379

Dolmans, D. H. J. M., Wolfhagen, I. H. A. P., & Schmidt, H. G. (1996). Effects of tutor expertise on student performance in relation to prior knowledge and level of curricular structure. *Academic Medicine, 71*(9), 1008-1011.

Fan, X., Thompson, B., & Wang, L. (1999). The effects of sample size, estimation methods, and model specification on SEM fit indices. *Structural Equation Modeling, 6*, 56-83.

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

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent systems. In R. Cudeck, S. D. Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Jöreskog*. Lincolnwood, IL: Scientific Software International.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235-266. doi:10.1023/B:EDPR.0000034022.16470.f3

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Jacobs, A. E. J. P., Dolmans, D. H. J. M., Wolfhagen, I. H. A. P., & Scherpbier, A. J. J. A. (2003). Validation of a short rating scale to assess the degree of complexity and structuredness of PBL problems. *Medical Education, 37*(11), 1001-1007.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63-85. doi:10.1007/BF02300500

Kamin, C. S., O'Sullivan, P. S., Younger, M., & Deterding, R. (2001). Measuring critical thinking in problem-based learning discourse. *Teaching and Learning in Medicine, 13*(1), 27-35. doi:10.1207/S15328015TLM1301_6

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151-173.

Majoor, G. D., Schmidt, H. G., Snellen-Balendong, H. A. M., Moust, J. H. C., & Stalenhoef-Halling, B. (1990). Construction of problems for problem-based learning. In Z. H. Nooman, H. G. Schmidt, & E. S. Ezzat (Eds.), *Innovations in medical education: An evaluation of us present status B2—Innovations in medical education: An evaluation of us present status*. New York: Springer Publishing Co.

Mamede, S., Schmidt, H. G., & Norman, G. R. (2006). Innovations in problem-based learning: What can we learn from recent studies? *Advances in Health Sciences Education, 11*(4), 403-422.

Marin-Campos, Y., Mendoza-Morals, L., & Navarro-Hernandez, J. A. (2004). Students' assessment of problems in a problem-based pharmacology course. *Advances in Health Sciences Education, 9*, 299-307.

Mpofu, D. J. S., Das, M., Murdoch, J. C., & Lanphear, J. H. (1997). Effectiveness of problems used in problem-based learning. *Medical Education, 31*(5), 330-334.

Norman, G. R., & Schmidt, H. G. (1992). The psychological basis of problem-based learning: A review of the evidence. *Academic Medicine, 67*(9), 557-565. doi:10.1097/00001888-199209000-00002

Rangachari, P. K. (1998). Writing problems: A personal casebook. http://fhs.mcmaster.ca/pbls/writing/images/Writing Problems.pdf

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

Schmidt, H. G. (1993). Foundations of problem-based learning—Some explanatory notes. *Medical Education, 27*(5), 422-432.

Schmidt, H. G., & Boshuizen, H. P. A. (1990, April). *Effects of activation of prior knowledge on the recall of a clinical case*. Paper presented at the Meeting of the American Educational Research Association, Boston, MA.

Schmidt, H. G., Dolmans, D. H. J. M., Gijselaers, W. H., & Des Marchais, J. E. (1995). Theory-guided design of a rating scale for course evaluation in problem-based curricula. *Teaching and Learning in Medicine, 7*(2), 82-91.

Schmidt, H. G., & Gijselaers, W. H. (1990, April). *Causal modelling of problem-based learning*. Paper presented at the Meeting of the American Educational Research Association, Boston, MA.

Sockalingam, N., & Schmidt, H. G. (2007, March). *Features of problems in problem-based learning: The students' perspective*. Paper presented at the HERDSA Annual Conference, Adelaide, Australia.

Soppe, M., Schmidt, H. G., & Bruysten, R. (2005). Influence of problem familiarity on learning in a problem-based course. *Instructional Science, 33*(3), 271-281.

Tiwari, A., Lai, P., So, M., & Yuen, K. (2006). A comparison of the effects of problem-based learning and lecturing on the development of students' critical thinking. *Medical Education, 40*(6), 547-554.

Van Berkel, H. J. M., & Schmidt, H. G. (2000). Motivation to commit oneself as a determinant of achievement in problem-based learning. *Higher Education, 40*(2), 231-242.

Williams, P., Iglesias, J., & Barak, M. (2008). Problem based learning: Application to technology education in three countries. *International Journal of Technology and Design Education, 18*(4), 319-335.

NACHAMMA SOCKALINGAM is at present a Lecturer at The Teaching and Learning Centre, SIM University, Singapore. This study was conducted when she was at Republic Polytechnic, Singapore.

JEROME ROTGANS is a Research Scientist at the Centre for Research in Pedagogy and Practice, National Institute of Education, Singapore.

HENK SCHMIDT is a Professor of Psychology and the Dean of the Faculty of Social Sciences at Erasmus University Rotterdam. He is also the appointed Rector Magnificus at Erasmus University Rotterdam, The Netherlands.

Acknowledgements

The authors would like to thank Republic Polytechnic, Singapore for support of this study.

Appendix
Detailed Description of the Five-Factors and 14 Parcels

Parcels	Statement
Factor 1: The extent to which the problem leads to formulation of intended learning objectives	
1. Clarity of the problem	1. I was clear about what the problem required my team and me to do
	2. The problem was clearly stated
2. Elements of clue or key words in problem	3. The problem provided sufficient clues/ hints
	4. The problem contained sufficient keywords
3. Structured approach to the problem	5. I was able to identify the key learning objectives from the problem
	6. I was able to come up with a satisfactory list of topics to explore on based on the problem
	7. I had a logical approach to the problem
Factor 2: The extent to which the problem is familiar to students	
1. Familiarity with content	1. I was familiar with the content of the problem even as I started to work on it
	2. I have personally experienced one or more situations described in the problem
	3. I could relate to the content of the problem based on my experiences
2. Relates to general knowledge	4. The problem statement fits well with my prior knowledge
	5. The subject matter of the problem reflected current affairs/issues around the world
3. Relates to subject-domain knowledge	6. I have done similar topic as in the problem before
	7. I had sufficient basic knowledge to identify suitable resources
Factor 3: The extent to which the problem interests students	
1. Triggers personal interest at the start	1. I was not interested to read the problem
	2. I was curious to find the answer
2. Engages in self-directed learning	3. The problem stimulated me to find out more information on the topic
	4. The problem stimulated me to work hard during the breakouts
3. Problem captivates attention	5. The problem was engaging throughout the learning process
	6. The problem captivated my attention throughout the day
Factor 4: The extent to which the problem promotes collaborative learning	
1. Problem triggers brainstorming	1. The problem triggered sufficient level of group discussion
	2. We brainstormed over the problem on what we needed to find out
2. Problem triggers team discussion	3. Everyone in the team participated in the discussion
	4. The problem stimulated us to discuss
3. Problem encourages team work	5. Team member's expertise in different subjects helped in solving the problem
	6. Our team worked efficiently
Factor 5: The extent to which the problem stimulates critical reasoning	
1. Problem stimulates thinking, questioning and reasoning	1. The problem triggered lots of questions in my mind
	2. I analyzed the information collected to respond to the problem
	3. The problem stimulated me to think and reason statement
2. Problem encourages multiple perspectives	4. The problem had more than one right answer
	5. There were many different viewpoints regarding the solution
	6. Team members had diverse opinions on the problem