J Sci Educ Technol (2010) 19:146–164
DOI 10.1007/s10956-009-9188-9

Development and Application of a Rubric for Evaluating Students' Performance on Newton's Laws of Motion

Mustafa Sabri Kocakülah

Published online: 16 September 2009

Abstract This study aims to develop and apply a rubric to evaluate the solutions of pre-service primary science teachers to questions about Newton's Laws of Motion. Two groups were taught the topic using the same teaching methods and administered four questions before and after teaching. Furthermore, 76 students in the experiment group were instructed about the features and use of the rubric and asked to construct a rubric, while 77 students in the control group were not. Students' solutions were evaluated with the agreed rubric by the instructor, an independent coder and the peers in the experiment class. The effectiveness of the rubric on students' achievement was examined by applying descriptive statistics and linear regression to scores obtained from both tests. *T*-test statistics and analysis of variance procedures were also used to analyze the reliability and validity of the assessments made. The results revealed that the developed rubric was used consistently by the instructor and peers and significant correlations ($p < 0.001$) were found among the scores. The inter-coder reliabilities were 0.98 and 0.93 in the pre- and post-tests with 76 peer coders. A generalizability study showed that the estimates of 16 peer coders on average matched the reliability of single-instructor assessments. It was concluded that the developed rubric was able to highlight the aspects of the problem solutions and helped increase students' achievement.

Keywords Newton's laws of motion · Peer assessment · Pre-service primary science teachers · Rubric

M. S. Kocakülah (✉)
Necatibey Education Faculty, Balıkesir University,
10100 Balıkesir, Turkey
e-mail: sabriko@balikesir.edu.tr

Introduction

Rubric is regarded as a popular pedagogical tool whose contribution to education has been recognized by teachers and teacher educators. Indeed, recent literature on science teaching and learning has focused on the development and use of a rubric as an alternative and a contemporary assessment tool for evaluating performance (Luft 1999; Popham 1997). A rubric describes various levels of performance that students are supposed to attain across a scoring scale. A rubric is therefore important for both teachers and students as it reveals the desired achievement for a performance task with an established set of criteria so as to score students' performances. The highest levels of student performance define a complete task and place students in a constructive learning and self-evaluation process (Hafner and Hafner 2003).

Luft (1999) identifies two perspectives at the beginning of rubric development: In the first perspective, goals and standards should be clarified for students by the rubric designer, while the description of 'acknowledged' standard should be depicted for a scientifically literate person in the second perspective. From the researcher's perspective, the focus should be more on the experiences and the development level of the students to develop a rubric rather than on whether the criteria originated from the designer or from external standards. In addition, purpose, clarity, feasibility, generalizability and suitability for improvement are among other criteria for rubric development (Herman et al. 1992).

As Hafner and Hafner (2003) pointed out, studies involving the use of rubrics in the literature have focused on their construction, validity and reliability analyses as well as their generalizability for scoring various teaching processes [laboratory performance (Rutherford 2007), constructivist classrooms (McClure et al. 2000)] and techniques [concept

maps (Shaka and Bitner 1996) and laboratory flow diagrams (Davidowitz et al. 2005)] from the researchers'/teachers' viewpoint. Although these studies investigate with experienced coders the aspects of effective teaching practices along with the production of an assessment tool to indicate students' progress throughout or by the end of the course, they offer scarce information about the students' ability to use rubrics to assess the performance of their classmates and gender-based differences in students' performances as evaluated by their teachers and peers.

Only the study of Hafner and Hafner (2003) has revealed the relationship between teacher's and students' performance assessment scores obtained from a constructed rubric. They examined the effects of possible gender-based differences in student achievement on oral presentation skills and reported that the rubric used as a scoring tool could be considered as 'gender neutral' since there was no difference between the genders.

Lorenzo et al. (2006) draw attention to the fact that although the gender gap has been closing in many science and technology fields, the largest gender difference in students' performance remains in physics. Therefore, the authors investigated the possibility of reducing the gender gap in conceptual understanding in an introductory calculus-based mechanics course and initially provided instruction with traditional teaching methods both in lectures and sections, while shifting to a highly interactive teaching style year by year during the 7 years of the study. Traditional lectures of the first year were replaced by Peer Instruction (PI) including short (10–15 min) mini lectures and a set of conceptual questions were discussed by students in small groups to address conceptual difficulties and to actively involve the students. The implementation of PI was improved by using a research-based mechanics textbook which was developed by Mazur (1997) and refined the in-class questioning/discussion strategy over the following 5 years (Crouch and Mazur 2001). During the final year, traditionally presented section meetings were also replaced by the composition of tutorials and cooperative quantitative problem solving activities to increase student engagement. Thus, the courses studied are classified into three groups: traditional (T), partially interactive (IE1) and fully interactive (IE2). It was found that an increased degree of interactivity significantly decreased the gender gap from pretest to posttest. In addition, the use of PI in the lectures (IE1) is reported to reduce the posttest gender gap to 7.8%, which amounts to less than two-thirds of the pretest gender gap. Full interactivity (IE2) cut the posttest gender gap to 2.4%, which was a quarter of the pretest gap, and the gender gap for this group was not statistically significant ($p = 0.429$) after instruction.

Lorenzo et al. (2006) attributed the reduction in gender gap to the introduction of interactive instructional methods that offer students opportunities to interact and explain their ideas, provide feedback on their understanding through the conceptual questions and tutorials, alternate between structured teaching and group works with peer discussions, encourage collaboration among peers, and provide a less competitive classroom culture. Therefore, the correlation between the teacher's and students' performance assessment scores and the effect of gender on students' performances remains to be further analyzed in different studies. In other words, there is a need to investigate the following questions: 'to what extent can the rubrics be considered as an assessment tool for the students?', 'do rubrics yield valid and reliable outcomes in assessing performance when employed by student coders?' and 'does a student's gender affect how she or he assesses performance using the rubric?'

The present study describes the development and application of a rubric to evaluate students' performance in Newton's Laws of Motion (NLM). In particular, it investigates the effectiveness of the rubric when used by students to solve problems about NLM, and thus, attempts to reveal whether they achieve better in problem solving than those who do not use the rubric in their problem solving efforts. The study also aims to evaluate the necessity of drawing free body diagrams, to assess students' rating in terms of their consistency with the experienced coders, and to find out whether there is any statistical significance among student scores in terms of student coders' gender in the actual setting of a physics classroom.

The conceptual area of NLM was selected on the basis of the researcher's previous teaching experiences during which students were observed to respond intuitively or struggle in solving problems in this area, as well as on the basis of the appropriateness of the topic in fulfilling the conditions of creating a rubric, i.e. establishing criteria providing descriptions of each level of student performance. A review of the literature shows, not surprisingly, that students' understandings of NLM ideas are not retained despite the researchers' attempts to incorporate constructivist ideas (Shymansky et al. 1997). There are even some cases in which students had problems with using newly acquired knowledge in problem situations and answered questions intuitively by using an Aristotelian-like model rather than the Newtonian model (Jimoyiannis and Komis 2003; Mildenhall and Williams 2001; Parker and Heywood 2000; Trumper and Gorsky 1997).

Andaloro et al. (1997) define students' intuitive ideas and concepts as 'common sense knowledge' in the analysis of motion. In this context, the authors find it necessary to implement different teaching/learning strategies to overcome conceptual and reasoning difficulties in solving problems. They used free-body diagrams (FBDs) to reveal students' difficulties in modelling the interactions between

the different parts of selected mechanical systems and considered each FBD as a step in problem solution. They reported genuine conceptual changes that occurred in many cases in which students drew Newtonian FBDs of interactions in the successive tasks from the beginning and felt the need to search for new forces in the systems. In another study, Rosengrant et al. (2005) examined whether students who used FBDs to solve mechanics and electrostatic problems were more successful than those who did not. They found that for over 12 problems on four exams, 85% of the students who drew an FBD correctly to solve a problem found the correct answer. In addition, they also emphasized that drawing incorrect FBDs led to more incorrect solutions than solving a problem without a diagram at all.

The literature on rubrics commonly agrees that implementing formative and self assessment requires the use of rubrics (Bednarski 2003; Jackson and Larkin 2002). A rubric helps students see the learning and performance goals, self-assess their work, which is found to be more powerful than instructor-provided feedback, and modify it to achieve goals and thus to develop better understandings of the task (McCollister 2002; Stoll 2003). Based on these considerations, the conceptual area of mechanics was thought to be suitable to construct such an assessment tool. During teaching in introductory mechanics course, the researcher paid attention to the seven steps of problem-solving strategy for NLM unit suggested by Serway and Beichner (2000). These seven steps of the strategy were in well accordance with the problem solving strategy employed in Gaigher et al.'s (2007) study. In their strategy, the diagram was the focus of attention in the study which was designed to improve students' problem-solving performance as well as to develop conceptual understanding. They reported that the structured problem-solving strategy yielded enhanced conceptual understanding of energy conservation principle in physics. Therefore, these steps were considered to help form the criteria to be assessed in a constructed rubric.

Aim of the Study

The main aim of this study is to develop and use a rubric to evaluate students' understandings by focusing on their performance in solving problems related to NLM. By analyzing the students' performance, the researcher sought answers to the following questions:

1. Do the students, who take part in the designing and application process of a rubric, perform better in solving problems than those who are instructed with the same teaching methods but never used a rubric?
2. To what extent do the students' and the researcher's assessment scores correlate?
3. Does a student's gender affect how she or he assesses their peers' solutions to problems using the rubric?

Methodology

Population and Sample

The target population for the study is the second-year university students in the primary science education departments of the education faculties in Turkey, while the accessible population has been the second-year university students in the primary science education department of a particular education faculty at Balikesir University. The sample of this study involves 153 second-year primary science university students in four intact classes of the education faculty at Balikesir University. The faculty where the researcher works was selected for the purposes of purposive sampling method. The faculty has a good reputation in training and has trained students for almost a century. The rubric used in this study (see Appendix 1) was designed to be both a supportive and an assessment tool to solve problems on NLM in the Introductory Physics course, a compulsory course for the second-year primary science university students. The student sample had been taught NLM during their secondary school years and while they were preparing for the university entrance examination.

Research Design and Instruments

This study followed a quasi-experimental control group design with a pretest and posttest. Students were first assigned to morning and evening classes by listing their scores in the university entrance examination organized by the Higher Education Council in a descending order while maintaining an even distribution of students as far as possible. The university entrance examination scores of the students in all classes of the faculty were compared by an independent samples one-way analysis of variance (ANOVA) test. The results indicated that there was not a significant difference among the groups in four classes ($F = 2.20$, $p = 0.09$). Considering that the university entrance examination does not involve only physics questions, it was decided to administer an achievement test to ensure an equal distribution among the classes in terms of their abilities in physics. The achievement test involved questions about the subjects taught at secondary schools, such as energy, force and motion, electricity, magnetism and waves.

The first version of the achievement test consisted of 36 multiple-choice questions. The questions were formulated by the researcher and revised on the basis of suggestions from three physics educators about face validity, clarity of language, and suitability for the age level concerned. To optimize the reliability and validity of the original test, it was first administered to a pilot group of 147 second-year primary science students during the academic term of 2006–2007 at the same university. After necessary revisions were made as based on the results of the item analyses of the pilot study in terms of item difficulty and discriminatory indices, a 23-question achievement test was formed. This final version of the test had a coefficient alpha (or KR-20) of 0.77 and an average item difficulty index of 0.75. The finalized achievement test was administered simultaneously to four classes of second-year primary science university students during the academic term of 2007–2008. ANOVA test was used for the statistical analysis of the achievement test. The results indicated that there were not any significant differences among the classes ($F = 1.35$, $p = 0.26$; Table 1).

After the four classes were determined to be equal in terms of their abilities in physics, the researcher decided that the experiment and control groups would be composed of two classes. Then, each class was randomly assigned to either the experimental or control group to address the research questions specified above. While the same content and teaching strategy and the same pre- and post-test questions were applied to both groups, the construction and use of the rubric only involved the experiment group. The same teacher (the researcher) instructed both groups by implementing faithfully the same teaching approach to minimize implementer effect. Both experiment and control group students were immersed in the same teaching conditions (i.e. solving problems by drawing traditional FBDs, using the same questions and simulations or applets downloaded etc.) and environment (i.e. using the same classroom) to minimize Hawthorne effect as suggested by Caleon and Subramaniam (2005).

It should be noted that the experiment group was not exposed to any additional activities or problem solving applications about the questions on NLM during rubric construction. The control group was equally occupied with activities concerning NLM. While the experiment students solved the sample questions by the help of the rubric, the same questions were also solved as exercises in the control group but without the aid of the rubric after the teaching of NLM was completed. Four questions (see Appendix 2) on NLM were administered to all students at the beginning and end of the course as pre- and post-tests, respectively.

Table 1 Comparison of the classes based on the results of the achievement test

Variance	Sum of squares	*df*	Mean square	*F*	*p*
Between groups	28.87	3	9.62	1.35	0.26
Within groups	1017.02	143	7.11		
Total	1045.89	146			

The questions were selected from among a collection of physics textbooks. The content validity of the questions were established by a panel of experts, including three physicists with more than 15 years of experience in teaching introductory physics at university level. They were asked to evaluate the appropriateness of the questions and their relevance to introductory physics course. In accordance with their comments and feedback, some minor changes were made in some figures and wording of the questions. The panel approved the final format of the questions and agreed that the questions covered the NLM topic in physics course, which implies content validity.

Procedure

Development of the Rubric

Having completed teaching the NLM unit within 3 weeks and 12 h of physics course, the experiment group students were briefed on the concept, types and design of rubrics used in a week's time, which consisted 4 h of two physics courses. Each 2 h of a physics course were offered on different weekdays. The students were asked to raise any questions they would like to be answered about rubrics. Once the rubrics were thoroughly introduced to the students, the researcher ensured that they had understood how to use the rubric by asking them to cooperate and to create their own rubric to assess the solutions provided for the problems about NLM in small groups (four students on an average in each group). The students were divided into 18 groups in two classes and asked to design their rubric in 2 weeks' time and to present their task as a group to the entire class by commenting on why the designed rubric was valuable and fruitful in different aspects. It was believed that the students could accomplish such a task since they were taught the topic and became familiar with the concept of rubrics. The students were free to select the type of the rubric they wanted; however, each group of students was supposed to justify the evaluation criteria of the rubric to ensure the construction of a good rubric which yields information about students' learning and appropriate assessment of problem solutions in their presentations. Thus, students' understanding of the grading criteria and their importance in assessing performance would be strengthened as they were involved in creating rubrics by reviewing some of the questions that had been solved while teaching.

All presentations were held in a fully equipped lecture room and the researcher chaired the presentations. Selection of the most practical and effective rubric was achieved by adding up the scores given by the researcher, two experts in science education area, and a spokesman who presented the consensus of the students in each group at two stages. First of all, the groups were asked to present their products to select the best rubric in each class. Secondly, students in both classes gathered together in an auditorium where the best rubrics of two classes were presented again and re-evaluated to select a final rubric to be used. During the presentations, scorings by the researcher, experts and the students were performed concurrently but independently from each other. Details pertaining to the selection of the best designed rubric were described before group presentations started and discussions were based on the originality of the rubric, its capacity to assess performance objectively, whether the assessment criteria were presented in a plausible order, whether the rubric was presented in an appropriate form, and the extent to which the rubric was successful in evaluating solutions to NLM problems.

While four groups constructed holistic rubrics, the remaining 14 groups worked on an analytic type to score the solutions to problems on NLM. The selected rubric was an analytical one with six criteria topics. The rubric shown in Appendix 1 was a slightly modified form of the selected rubric that was originally constructed as a performance task by the group who received the highest score among all groups in two classes. It is important to note that the rubric in its selected form and its modified form was constructed as a class activity with discussion and input from the students. Indeed, the six elements of the rubric, together with the evaluative criteria, the grading system and the overall percentage attributed to each criteria topic, represented a consensus of student opinion; the researcher only facilitated discussion about the correct order of the criteria topics and the correct proportion of criteria topics' percentage. The validity of the proportion of criteria topics (i.e. weights) along with criteria statements were verified by five experts in the field of physics education and two faculty members who were experts in both assessment and methodology.

It is believed that the use of these criteria may facilitate the solution of questions about NLM. This belief can be better explained in view of the solution of the second question in the test.

This question requires determining the movement direction of the system to be able to describe the objects' motion. Because the objects in a system may not move in the same dimension, drawing axes and defining the objects' motion according to those axes are also an important aspect of problem solution. In the next stage, the task of drawing FBDs for the system and each object in that system follows to better analyze the problem and to determine the direction and the value of forces acting on the object(s) of interest, as seen in Fig. 1. When the task of drawing FBDs is completed, the direction of the resultant force acting on the system can be found easily and the movement direction and type of motion (i.e. accelerating or moving with a constant speed) of the system can be described. This can be achieved by applying Newton's second law of motion equation ($F = ma$) to the question. As observed in the case of question two, it is found that the system accelerates in the direction of the gravitational force $m_2\vec{g}$ (i.e. clock-wise) and then the value of acceleration is calculated by writing the equation for the whole system. In some cases, it may be required to focus on a part of the system and the same equation can be used by examining the FBD of each object of the system to calculate tension in the string. It is calculated for both masses in Fig. 1. Finally, each physical quantity should be written in its own unit to complete the exact solution of the question.

Use of the Rubric

After collaborative modification of the rubric, 76 students in two experiment classes were told that they would use the rubric by viewing it as a helpful guide when solving problems about NLM and that their solutions would be evaluated fully by the researcher and their peers using the rubric. Each student was given the rubric and two post-test questions in the last 20 min of two physics courses to solve the problems with the aid of the rubric when teaching the topic and the construction of the rubric was completed. Following this activity, students were also asked to evaluate their peers' solutions to the pre- and post-test questions using the rubric. Prior to evaluation, the researcher asked the students to provide their candid and confidential peer-group assessment scores of the problem solutions using the rubric. Each student was circulated a one-page rating handout, a photocopied version of the rubric, and his/her peer's solution papers for the questions. In order to ensure maximum confidentiality, the researcher delivered evening class A's papers to morning class A students and vice versa. Moreover, the students' names on each question form were erased and a code assigned to each student on the top centre of the form for the sake of anonymity. Students' scoring started with the pre-test questions and two questions were rated in the last 20 min of each physics course and four pre-test questions were rated in a course-week. Students rated four post-test questions during the next course-week and peer rating lasted for 2 weeks in order to rate a total of eight pre- and post-test questions altogether. As a quick aid in scoring, the researcher solved each question on board by discussing the elements

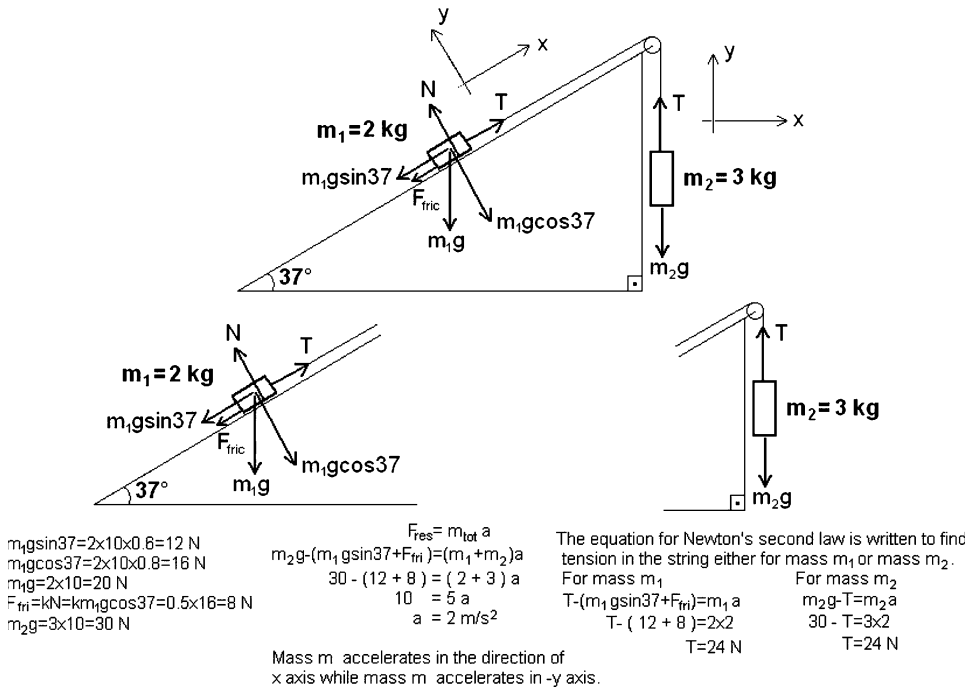


Fig. 1 Solution of the second question using the designed rubric

embedded in the criteria topics of the rubric and the students had to look for those on their peers' forms. Finally, the researcher projected the complete solution to the screen on the wall so that all the students could easily follow the steps. The students wrote down their names on the rating paper and they were asked to provide peer assessment of the problem solutions for each question using the rubric. The researcher indicated that he was there to answer if any questions were raised to ease the students' task without interrupting or harming the rating process and students were asked to provide the student's code to be assessed on the rating paper (provided on the question form), the scores on criteria topics and the total points for the solution to each question. The researcher also rated the students' responses to the pre- and post-test questions independently from the students' ratings using the rubric on later dates. Once the students' and the researcher's ratings were completed, the collected data were subjected to analysis.

Data Analysis

Individual scoring sheets completed by the students for peer assessment of the problem solutions to pre- and post-test questions in the experiment class and the instructor's and an independent coder's rubric-based scores for the same questions of both the experiment and control classes provided the main quantitative basis for this analysis. Descriptive statistics were employed to analyze the overall means and range of the instructor, coder and peer scores for all questions. Qualitative descriptions about the answers to the pre- and post-test questions were also presented to outline the attributes inherent in the answers and to express the similarities or differences in the students' answers between both classes.

The variables used in the analyses were examined for normality by using the Liliefors correction of the Kolmogorov–Smirnov test. For all questions in the pre- and post-tests, scores were found to be distributed normally for both groups of students; accordingly, all comparisons and analyses were performed using parametric statistical routines (i.e. *t*-tests, ANOVA, Pearson correlation, linear regression analysis).

To find out the validity of the students' ratings of the problem solutions using the rubric, the students' mean peer score for the questions were compared with the mean instructor score. It should be noted here that instructor's scores were independent from the students' scores although it employed the same rubric as used by the student coders. Traditional linear regression analyses were used to analyze the congruence between instructor's and peer scores and to investigate the correlation between them.

Several methods were employed to evaluate inter-coder reliability in this study. Firstly, an independent coder, who was an expert in the area of science education, was asked to score students' solutions for the experiment and control groups for four questions in the test. Having obtained the scores by the second coder, an analysis of inter-coder reliability was carried out by applying independent *t*-tests to examine the difference between the researcher's and the second coder's scores.

Secondly, an analysis of reliability was performed on peer assessments by different peer groups because peer

assessment was applied in two parallel sessions of morning and evening classes. This approach was in well accordance with the critique made by Magin and Helmore (2001) and Topping (1998) that peer assessments should be compared with those of other peers or the same peers over time for the reliability of peer scores. Furthermore, peer assessment was regarded as part of the course in which the rubric, its application, and the instructor (the researcher) were the same in both classes. The mean scores for the questions of each class were calculated to analyze the reliability of peer assessment by comparing scores given by different peer groups.

Thirdly, a reliability analysis was also carried out within peer groups. Since the students solved four problems using the rubric and the same peer assessed solutions across four questions of the same student, agreement between the assessment scores of the peers for four questions can be treated as an indicator of the reliability of peer assessment. Thus, one-way repeated measures ANOVA test was applied to analyze the difference between the peer scores for four questions.

Green et al. (2000) suggest controlling type I error-rate when more than two comparisons are conducted. Similarly, they define the probability of committing one or more type I errors for pairwise comparisons as experiment wise alpha (α_e). The preferred level of α_e is most frequently 0.05 (Hair et al. 1998). To minimize the chance of committing a type I error and keep α_e at the 0.05 level, the alpha level for each pairwise comparison among four questions had to be reduced to a value less than 0.05. Therefore, the Bonferroni correction method, which is recommended for use in repeated measures involving four or more groups and is the most conservative form of testing the likelihood of committing type I error compared to Sidak and LSD configurations, was used by adjusting the alpha level to 0.0125 for each pairwise comparison, as was done previously in the studies by Fowler et al. (2009) and Hohenshell and Hand (2006).

The ANOVA test was also required to assess the level of inter-coder agreement for the pattern of peer scores (Hafner and Hafner 2003). This test was used to determine two different reliability indices in a study by Magin and Helmore (2001) and the same procedure was applied in this study. The formulas for two indices were $r_{nm} = F - 1/F$ and $r_{11} = (F - 1)/(F + N - 1)$as used by Kilic and Cakan (2007) and originally developed by Ebel (1951). Here, the coefficient of r_{nn} defines inter-coder reliability of N coders' average scores; besides, the intra-class correlation coefficient r_{11} provides an idea of how reliable peer assessment can be if each problem solution is to be rated by one coder rather than N coders. In addition, Cronbach-alpha reliability coefficient was calculated to reveal single instructor reliability for both tests.

A generalizability study was performed for the scores of problem solutions to assess student inter-coder reliability. The inter-coder reliability coefficients for various numbers of students were calculated by using the Spearman-Brown prophecy formula (Ferguson 1971; Gillmore 2000; Litzinger et al. 2007; Magin and Helmore 2001). Information obtained from this study was used to estimate the number of peer ratings, which is needed to improve the reliability of the single-peer-scored rubric. Following this analysis, the same formula was also used to calculate student equivalent index N_{SE}, which provides the average number of students to reach the same reliability as calculated for single instructor ratings.

In addition to sample size estimation, performing power analysis was considered as another important aspect of experimental design to control experiment-wise alpha (α_e) level. This method has been reported not to be performed until recently in complicated research designs which use multivariate and repeated measures models (D'amico et al. 2001; Potvin and Schutz 2000). Gravetter and Wallnau (1996) and Hair et al. (1998) identify the factors that influence statistical power as alpha level, sample size, and effect size. In order to complete the power analysis, only the value of effect size had to be taken into consideration and partial eta-squared (η_p^2) values of the pretest and posttest were used to perform one-way repeated measures ANOVA test as an equivalent measure of the effect size (η^2) (Bakeman 2005).

Kinnear and Gray (2008) suggested to use η_p^2 values ranging between $0.01 \leq \eta_p^2 < 0.06$ as small, $0.06 \leq \eta_p^2 < 0.14$ as medium, and $\eta_p^2 \geq 0.14$ as large for repeated measures ANOVA. D'amico et al. (2001), Green et al. (2000) and Stevens (1996) also identified 0.01, 0.06 and 0.14 as traditional cut-offs deemed as small, medium and large η_p^2 values. Here, Kinnear and Gray's (2008) approach was used to interpret effect size values since it explicitly provides the values for SPSS application of repeated measures design, which was also used in this study. In the presentation of ANOVA test findings, the effect size (in terms of η_p^2) and the power of the test (symbolized as P, based on $\alpha = 0.05$) are reported.

Results

Profile of Instructor and Peer Assessment Scores

Using the rubric, each question was scored over a total point of 100. Class sizes for the control and experiment classes are 77 and 76 students, respectively. While the numbers of female and male students are close to each other (i.e. 38 females and 39 males) in the control group, gender ratio in the experiment class favours females (i.e.

Table 2 Descriptive statistics of the instructor, independent-coder and peer scores in both classes

	Instructor score				Independent-coder score				Peer score			
	Mean	SD	Skewness (p*)	Range	Mean	SD	Skewness (p*)	Range	Mean	SD	Skewness (p*)	Range
Experimental												
Pre	23.25	8.25	−0.03 (0.20)	4.25–41.25	23.11	8.24	−0.06 (0.20)	4.25–41.25	23.57	8.33	−0.00 (0.20)	6.00–43.25
Post	84.95	10.64	−1.00 (0.06)	50.50–100.00	84.85	10.78	−1.00(0.05)	50.50–100.00	84.99	10.64	−0.99 (0.06)	50.50–100.00
Control												
Pre	23.72	8.49	−0.05 (0.20)	5.25–43.25	23.34	8.23	−0.10(0.20)	5.25–42.50				
Post	33.86	10.16	0.14 (0.20)	9.75–57.50	33.51	10.21	0.16 (0.18)	9.00–57.50				

* Shows the value of significance obtained from the Kolmogorov–Smirnov test for normality

1.375:1 of 44 females and 32 males). Table 2 shows the descriptive statistics of peer, instructor and independent coder scores for four questions. In the control group, the mean score given by the instructor was 23.72 with an SD of 8.49 before the instruction and it slightly increased to 33.86 ± 10.16 following the instruction. In the experiment group, the mean score awarded by the instructor was 23.25 ± 8.25 before the instruction, whereas it considerably rose to 84.95 ± 10.64 following the instruction.

Temporal variation in the groups' achievement was examined by using paired samples t-test. While the means for the pre- and post-tests for the experimental group are significantly different ($t = 54.09$; $p = 0.000 < 0.05$), there is a considerable change in the means for the control group, too ($t = 11.69$, $p = 0.000 < 0.05$). The mean pre-test scores of the experiment ($\bar{X}_e = 23.25$) and control groups ($\bar{X}_c = 23.72$) were compared by using independent samples t-test, which found no significant difference among the pre-test scores ($t = 0.35$, $p = 0.73$); however, the post-test results show a significant heterogeneity among the mean scores ($\bar{X}_e = 84.95$ and $\bar{X}_c = 33.86$, respectively and $t = 30.46$, $p = 0.000 < 0.05$).

The difference between the mean scores of the instructor and peers was examined by using an independent samples t-test. The mean of instructor scores ($\bar{X}_i = 23.25$) was found to be distributed uniformly with the mean peer score ($\bar{X}_p = 23.57$, $t = 0.24$, $p = 0.81 > 0.05$) in the pre-test. In the post-test, there was no significant difference among the mean values of the instructor ($\bar{X}_i = 84.95$) and the peer scores, either ($\bar{X}_p = 84.99$, $t = 0.03$, $p = 0.98 > 0.05$).

Qualitative Descriptions About Problem Solutions

Having explained the difference between the experiment and control groups' scores quantitatively, qualitative descriptions about the responses to the questions are presented below to portray the difference and to provide a more comprehensible explanation of the types of both groups' answers. Once general characteristics of the problem solutions from both groups of students in the pre-test are described, the solutions of one sample student selected from each group to one question of the test are presented. Furthermore, qualitative descriptions about the general features of problem solutions in both groups and the same two students' solutions in the post-test are outlined.

Both the experiment and control group students mostly drew incorrect or incomplete figures of forces for related questions in their solutions before teaching. This resulted in miscalculation of the resultant force on the system and incorrect definition of the direction of frictional force acting on the objects. Consequently, they expressed the direction and/or type of the system's motion incorrectly. Additionally, it is interesting that the lack of FBD for each object of the system in general caused incorrect motion equations written for the related object and miscalculations in finding the value of tension force. Moreover, the students appeared not to have used in their responses several physical concepts (i.e. force, acceleration, mass, etc.) with their units.

Figure 2 presents the students' solutions to the second question before teaching. It seems that students from both groups tried to indicate the forces acting on the objects in their solutions. However, it is evident that they miscalculated the value of frictional force (both indicated as $F_s = 10$ N) and the experiment group student also drew the frictional force in an incorrect direction (upwards on the inclined plane). As a result of the miscalculations about the values of forces, the resultant force ($F_{\text{net}} = 24$ N for the experiment and $F_{\text{net}} = 8$ N for control group students) and acceleration of the system were computed incorrectly ($a = 4.8$ m/s^2 for the experiment and $a = 1.6$ N/g for control group students). Additionally, the students were asked to calculate the value of tension in this question. Unfortunately, students from both groups tried to calculate the tension in the string without drawing an FBD of the related object and by considering the forces acting not only on the related object but also on other objects in the system, rather than only concentrating on the object of interest.

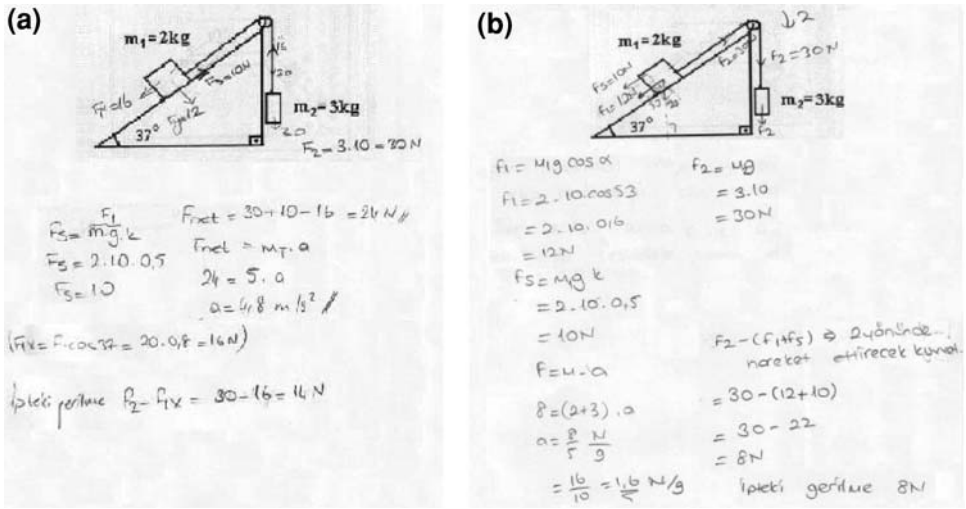


Fig. 2 Solutions of the selected students in the **a** experiment and **b** control groups to the second question in the pre-test

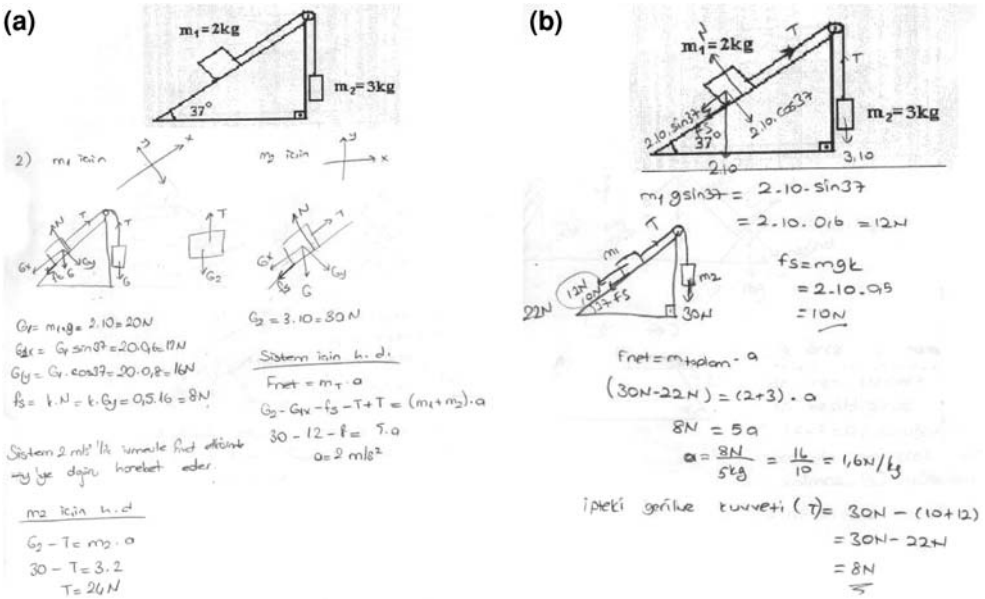


Fig. 3 Solutions of the selected students in the **a** experiment and **b** control groups to the second question in the post-test

Figure 3 shows the same students' solutions to question two after teaching. When the solution of the student in the experiment group is examined, it is clear that she started solving the problem by indicating axes and then showed the directions of forces by drawing FBDs for the system and each object and calculated the values of those forces correctly. After she wrote the equations for motion of the system and one of the objects with a mass of 3 kg correctly, the direction and type of motion were defined well and physical quantities were provided with their units. However, the student in the control group continued to make the same mistakes as in the pretest. Although she drew the FBD of the system, it was not drawn for each object and this shortcoming caused an incapability to compose the motion equation for a single object in the system correctly. Therefore, the value of tension was miscalculated, even though she wrote the equation for the system's motion correctly with incorrect frictional values and hence the resultant forces acting on the system.

Dependence of Achievement on Gender

Further analyses were performed to investigate possible gender-based differences in relation to student achievement in problem solving. From the ratings of the researcher, the analysis results show that mean values for total scores for male and female students for pre- and post-tests in the experiment group indicate slightly greater means for females than for males ($\bar{X}_F = 24.36 \pm 7.46$ and $\bar{X}_M = 21.73 \pm 9.13$ for the pre-test, $\bar{X}_F = 86.64 \pm 9.35$ and

Table 3 Differences between genders for mean values of peer scores in both tests

Sex	Pre-test						Post-test					
	n	$\bar{X}$	SD	*df*	*t*	*p*	*n*	$\bar{X}$	SD	*df*	*t*	*p*
Female	44	24.69	7.39	74	1.37	0.17	44	86.60	9.45	74	1.55	0.12
Male	32	22.05	9.39				32	82.79	11.89			

Table 4 Regression analyses of the instructor's rating on the students' peer group ratings for each question in the pre- and post-tests

Question number	Pre-test					Post-test				
	b	*R*	R^2	*F*	*p*	*b*	*R*	R^2	*F*	*p*
1	1.005	0.987	0.975	2883.422	0.000	0.956	0.974	0.949	1375.000	0.000
2	0.953	0.973	0.947	1321.273	0.000	0.964	0.975	0.951	1440.583	0.000
3	0.979	0.985	0.971	2489.793	0.000	0.917	0.937	0.878	533.412	0.000
4	0.946	0.988	0.976	3016.872	0.000	0.950	0.977	0.954	1536.605	0.000

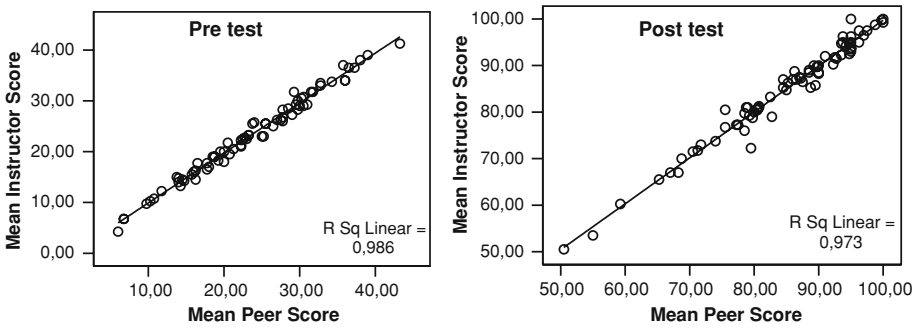


Fig. 4 The scatter plots of the mean peer scores against mean instructor scores for the pre and post-test questions, respectively

$\bar{X}_M = 82.64 \pm 11.97$ for the post-test), but such values are not significantly different ($t = 1.38$, $p = 0.17$ and $t = 1.63$, $p = 0.11$ for the pre- and post-tests, respectively). Possible differences between genders in their relative leniency in ratings of a student's peers were also examined by evaluating whether there is an association between genders and mean peer scores. The results of the analysis demonstrate that there are no statistically significant differences between genders for the mean values of peer scores in both tests, as can be seen in Table 3.

Validity

The validity of the rubric in the hands of the peer-group student coders was assessed on the basis of the instructor's (researcher's) rubric scores. First of all, question-by-question regression analyses were performed to provide a check on the accuracy with which the students used the rubric, as well as an overall assessment of the agreement level between the students' and instructor's employment of the rubric for four questions. As a result of these analyses, significant positive functional relations were detected between the instructor and peer scores in both tests, as observed in Table 4. Secondly, a linear regression analysis was performed to analyze mean peer and mean instructor scores for the questions. Figure 4 shows scatter plots of the students' mean peer scores against the mean instructor scores for all questions in the pre- and post-tests.

For the pre-test, the regression analysis yielded a significant slope of $b = 0.983$ ($t = 71.075$, $p = 0.000$) and an intercept of $a = 0.080$ (Fig. 4). This analysis also provided an R value of 0.993 and R^2 value of 0.986, which were statistically significant at the level of 0.001 ($F = 5051.675$, $p = 0.000$). For the post-test, a linear regression analysis also yielded a significant slope of $b = 0.986$ ($t = 51.374$, $p = 0.000$) and an intercept of $a = 1.111$ (Fig. 4). The R and R^2 values for this part of the analysis were 0.986 and 0.973, respectively, which were also significant at the level of 0.001 ($F = 2639.323$, $p = 0.000$).

Question-by-question independent t-test comparisons also showed that there was no statistically significant difference between the mean instructor and peer scores, as seen in Table 5. This was also evident during the regression analyses in which peer scores significantly correlated (see values R in Table 4) with the instructor scores for both tests, indicating a high level of agreement between the instructor and peer scores.

Table 5 *T*-test results showing a comparison of instructor and peer scores in detail

Question number	Coder	Pre-test						Post-test					
		n	$\bar{X}$	SD	*df*	*t*	*p*	*n*	$\bar{X}$	SD	*df*	*t*	*p*
1	Instructor	76	27.09	12.68	150	0.34	0.73	76	84.97	14.52	150	0.06	0.95
	Peer	76	27.79	12.46				76	85.12	14.79			
2	Instructor	76	27.13	10.29	150	0.08	0.93	76	82.41	15.45	150	0.17	0.87
	Peer	76	26.98	10.51				76	81.99	15.63			
3	Instructor	76	13.92	9.12	150	0.03	0.98	76	81.22	13.68	150	0.04	0.97
	Peer	76	13.96	9.18				76	81.31	13.98			
4	Instructor	76	24.87	10.71	150	0.39	0.69	76	91.21	12.26	150	0.18	0.86
	Peer	76	25.56	11.19				76	91.57	12.60			

Table 6 *T*-test results showing a comparison of instructor and independent-coder scores in detail

Question number	Coder[a]	Pre-test						Post-test					
		Control			Experimental			Control			Experimental		
		$\bar{X}$	SD	*t* (*p*)	$\bar{X}$	SD	*t* (*p*)	$\bar{X}$	SD	*t* (*p*)	$\bar{X}$	SD	*t* (*p*)
1	A	27.67	11.14	0.13 (0.90)	27.09	12.68	0.01 (0.99)	39.87	12.97	0.09 (0.93)	84.97	14.52	0.01 (0.99)
	B	27.91	10.97		27.08	12.71		39.69	13.06		85.00	14.58	
2	A	24.88	11.47	0.05 (0.96)	27.13	10.29	0.07 (0.94)	37.40	15.83	0.04 (0.97)	82.41	15.45	0.03 (0.98)
	B	24.79	11.40		27.01	10.27		37.31	15.92		82.34	15.58	
3	A	16.91	9.84	0.74 (0.46)	13.92	9.12	0.24 (0.81)	22.99	9.65	0.51 (0.61)	81.22	13.68	0.11 (0.92)
	B	15.75	9.55		13.56	9.04		22.19	9.62		80.99	13.92	
4	A	25.43	12.10	0.28 (0.78)	24.87	10.71	0.05 (0.96)	35.16	13.89	0.15 (0.88)	91.21	12.26	0.07 (0.94)
	B	24.90	11.93		24.79	10.87		34.83	13.98		91.08	12.45	

[a] A denotes the instructor and B denotes the independent-coder

Inter-coder Reliability

An independent samples *t*-test was performed to reveal the level of agreement between two independent coders in assessing the scores for the experiment and control group students' problem solutions. The test did not yield a significant difference across four questions in the pre- and post-tests, as clear from Table 6. Furthermore, correlations between two coders, a more traditional measure of inter-coder reliability, give extra evidence of agreement in coder scores. Therefore, the Pearson correlation analysis was performed between the scores of coder pairs and significant inter-coder reliability was also detected at the level of 0.01. The least value of bivariate correlations, r_p, across coders is 0.98 for the third question of the pre-test for both groups and for the same question of the post-test for the control group, while the rest of the questions yielded a coefficient of $r_p = 0.99$ ($p < <0.001$ in all cases).

In order to analyze the reliability of peer assessment by different peer groups, the mean scores of morning and evening classes of the experiment group were compared by using independent samples *t*-test for four questions. In that way, it was possible to demonstrate the agreement between the scores of peers for each question in detail. The results of *t*-tests are presented in Table 7. There is no statistically significant difference between the means of four questions ($p > 0.05$) in the pre and post-tests. These results indicate that peer scores are reliable.

The reliability of peer assessment within peer group was analyzed by using one-way ANOVA test for repeated measures. According to the ANOVA results (see Table 8), there was a significant difference between the means of the questions for both tests. Because a significant difference was found between the means of four questions in the pre- ($F = 48.40$, $p = 0.000$, $\eta^2 = 0.39$, $P = 1.00$) and post- ($F = 13.72$, $p = 0.000$, $\eta^2 = 0.16$, $P = 1.00$) tests, a post hoc analysis was performed to determine the question(s) that contributed to the significance. The mean peer scores (SDs in parentheses) were 27.79 (12.46), 26.99 (10.51), 13.96 (9.18) and 25.57 (11.20) in the pre-test and 85.12 (14.80), 81.99 (15.63), 81.32 (13.98) and 91.57 (12.60) in the post-test for questions one, two, three and four, respectively. A Bonferroni test conducted to compare the means indicated that a significant difference ($p = 0.000$) stemming from the difference between the means of the third and the other three questions for the pre-test and

Table 7 *T*-test results for all questions to show the level of agreement between peer coders in two different classes

Question number	Class type	Pre-test						Post-test					
		n	$\bar{X}$	SD	*df*	*t*	*p*	*n*	$\bar{X}$	SD	*df*	*t*	*p*
1	Morning	40	28.15	12.55	74	0.26	0.79	40	86.02	15.77	74	0.56	0.58
	Evening	36	27.39	12.52				36	84.11	13.79			
2	Morning	40	27.55	10.53	74	0.49	0.63	40	83.32	15.25	74	0.78	0.43
	Evening	36	26.36	10.59				36	80.50	16.12			
3	Morning	40	15.57	9.13	74	1.63	0.11	40	80.37	14.34	74	0.62	0.54
	Evening	36	12.17	9.03				36	82.36	13.69			
4	Morning	40	26.97	10.16	74	1.16	0.25	40	92.97	12.29	74	1.03	0.31
	Evening	36	24.00	12.20				36	90.00	12.93			

Table 8 One-way ANOVA results for four questions of the pre- and post-tests

Variance	Pre-test					Post-test				
	Sum of squares	*df*	Mean square	*F*	*p*	Sum of squares	*df*	Mean square	*F*	*p*
Between groups	20832.51	75	277.77			33976.75	75	453.02		
Measure	9561.09	3	3187.03	48.40	0.000	4998.98	3	1666.33	13.72	0.000
Error	14816.66	225	65.85			27329.27	225	121.46		
Total	45210.26	303								

Table 9 Reliability coefficients for the peer assessments

Coder	Number of students assessed (*N*)	Pre-test		Post-test	
		Inter-coder reliability index (r_{nn})	Single coder reliability (r_{11})	Inter-coder reliability index (r_{nn})	Single coder reliability (r_{11})
Peer	76	0.98	0.38	0.93	0.14

between the means of the fourth and the other three questions for the post-test in relation to pairwise comparisons. An overall alpha of 0.05 was maintained using the Bonferroni adjustment.

The reliability of the scores (r_{nn}), and an estimate of the intra-class correlation coefficient for individual peers (r_{11}) for two tests were calculated using the *F*-ratios as displayed in Table 9.

In the pre test, inter-coder reliability index (r_{nn}) was 0.98 while the intra-class correlation coefficient (r_{11}) was 0.38, which gave an estimate of the overall reliability of scores if only one student had rated the solution of each question. The analysis was repeated on the peer-rated data for post-test questions. The inter-coder reliability index (r_{nn}) was 0.93 and the estimated single coder reliability of peers (r_{11}) was 0.14. The reliability analyses indicated that although the reliability was high (0.98 and 0.93 in the pre- and post-tests, respectively) with 76 peer coders, reliability values were low with a single coder in both tests (0.38 and 0.14 in the pre- and post-tests, respectively). This result suggested that the number of peer coders should be increased to improve the reliability since an individual student could not be a reliable assessor of problem solutions to NLM unit.

Figure 5 shows the variation in inter-coder reliability coefficients for various numbers of peer coders. In general, reliability estimate is higher as more students rate the solutions of the questions. The score generalizability of the rubric as a function of the number of coders in Fig. 5 illustrates that as few as 15–25 coders yield reasonably high reliability coefficients for both tests. There is a rapid gain in score reliability from a single coder to about 15 coders but that the incremental increase in reliability levels off with approximately 30 coders. This information from the generalizability study can be used in the context of a decision study in reconsidering the improvement of the reliability of the peer-rated rubric and it can be reported that reasonably high reliability coefficients ($r = 0.80$) are predicted with peer-group coders as small as only about 20 students (Fig. 5) when using this rubric.

Table 10 provides comparisons of the single coder reliability for the instructor and peers, as well as the student equivalent index. It is estimated that an average of five peer scorings would be required to reach the same reliability as

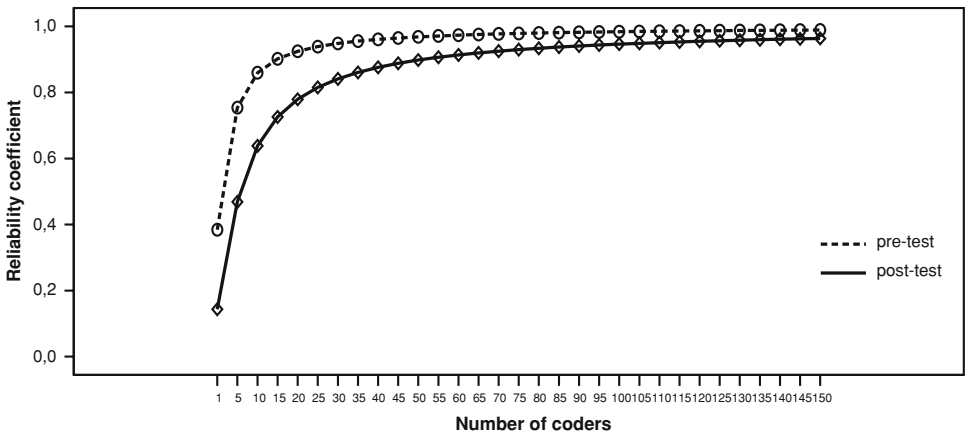


Fig. 5 Inter-coder reliability as a function of the number of students taking part in the rating process of the questions

Table 10 Comparison of the reliabilities for the single instructor and peer ratings

Test type	Instructor single coder reliability (r_i)	Peer single coder reliability (r_p)	Student equivalent index (N_{SE})
Pre	0.77[a]	0.38	5.43
Post	0.75[a]	0.14	16.13

[a] Shows the reliability of single instructor coding using Cronbach-alpha statistic

resulting from single-instructor scorings in the pre-test. However, single-instructor assessments are equivalent to approximately 16 peer assessments in the post-test. The same estimates can also be checked from Fig. 5 to indicate that reliability values of 0.77 and 0.75 for the pre- and post-tests, respectively, give the approximate intercept values of 5 and 16 in the horizontal axis.

Discussion

In this study, which investigated the effect of the use of a designed rubric on students' achievement in solving problems about NLM, the *t*-test results demonstrated that the mean of post-test scores of the experiment group students was significantly different from the mean of the control group students who received the same instruction but solved the questions without using the rubric. It is believed that such a result supports the idea that the use of a rubric makes students aware of what they are asked for and what they ought to focus on and this awareness has a positive impact on their problem solving ability. Similar results were obtained from the rubric designed by students to analyze laboratory flow diagrams in a study by Davidowitz et al. (2005). More than one-half of the students were able to differentiate accurately between various pieces of apparatus and their use of symbolic language as indicated in students' flow diagrams.

Another interesting outcome was that the students managed to use the rubric for peer-group assessment with a high degree of accuracy and comparability. There was a consistency in ratings between the instructor and the peers. The ranges of scores given by the instructor and peers were close to each other, as evident from the similar values of standard deviations and the mean values. This indicates that the students rated their peers with scores similar to the instructor's in both tests, which might be due to the fact that the rating dimensions of each category of the rubric were clear and students were aware of the assessment criteria in the rubric. These results were in accordance with other studies in the literature conducted by Falchikov (1995), Freeman (1995), Kwan and Leung (1996) and Stefani (1994), who found no significant difference between the mean instructor and peer scores. Group discussions were held and the students were involved in the construction and use of the rubric throughout this study since it is believed to have a considerable effect on students' understanding of the grading criteria and their importance in the performance, as Jackson and Larkin (2002) pointed out. Shaka and Bitner (1996) also found inter-correlations of the scores on the diversity of concept maps as statistically significant, a result which they attributed to a common understanding about the topic and a consistent interpretation of the attributes in the rubric.

Although the mean scores for females are higher than those for males in both tests when student achievement in solving problems are compared for genders based on the instructor scores, there is no statistically significant difference between genders. Moreover, the difference between genders was also investigated using the peer scores for both tests but this analysis also revealed a non-significant difference. Such findings support the previous research by Hafner and Hafner (2003) and show that although males and females perform differently when being evaluated with traditional assessment methods (for instance, Pomplun and Capps 1999; Pomplun and Sundbye

1999), the use of rubrics does not cause possible gender-based differences with regard to student achievement and students evaluate others with equal rigor using the rubric; in other words, neither gender is significantly more or less lenient than the other in peer scoring. Given that Hafner and Hafner (2003) concentrated on oral presentation skills, it is worth emphasizing that gender neutrality of the rubrics in assessing students' problem solutions was also examined in this study with a selected physics topic.

The validity of the rubric was tested by applying a linear regression analysis to the peer and instructor scores for each question. High correlations (the least value is $r = 0.937$, $p < 0.01$ for the scores of the third question of the post-test) were found between the instructor and peer scores of both tests. These significant correlations demonstrate that the students employ the rubric in the same way as does the instructor. Thus, peer scoring is a valid assessment of performance and a predictor of instructor scores with high accuracy in this study.

Although there is a lack of investigation in the literature on the achievement of students' problem solving using a rubric, the results of the regression analysis obtained from this study have a common approach with the studies mostly based on the assessment of students' presentation skills. For instance, MacAlpine (1999) found a correlation coefficient of 0.80 between the instructor and peer scores in assessing student presentations. Hughes and Large (1993) also detected a significant correlation ($r = 0.83$, $p < 0.001$) between the peer and tutor assessment scores in assessing communication skills. These findings reveal that correlations obtained in this study have higher values than other studies. It is also clear that instructor and peer scores correlated significantly for four different questions in both the pre- and post-tests, a finding which supports Kilic and Cakan's (2007) research that found high correlations in the first and second peer assessment applications.

Several methods were used to analyze the reliability of students' scores obtained from problem solutions. Firstly, the relationships between the instructor and an independent coder scores were examined using independent samples *t*-test and Pearson correlation analyses. The results of both analyses indicated that there was no difference between two coders' scores and high level of agreement existed between those coders. Secondly, the reliability of peer assessment was analyzed between different peer groups for four questions and no significant difference was found. This result was also in accordance with the result of the study in which Kilic and Cakan (2007) analyzed the elementary science teaching skills of three different peer groups and did not find a significant difference between the means of three groups in two applications. Since the students were briefed on the criteria to be searched and the steps to be followed before starting rating in this study, it was thought that training students in rating their peers had a positive impact on obtaining a high reliability of peer assessment. Thirdly, reliability of peer assessment within the peer group was examined by performing the ANOVA statistic. While the inter-coder reliability indices were considerably high in both tests (0.98 and 0.93 in the pre and post-tests respectively), it was found that if a single student assessed his/her peers' solutions, the reliability would be lowered to 0.38 and 0.14 for the pre- and post-tests, respectively. This finding suggests that it would be unsafe to rely on the scores given by a single peer coder alone as the basis to determine students' achievement in solving problems concerning NLM. In addition, the most reasonable line of explanation for the low inter-coder and single-peer coder reliability estimates in the post-test (0.93 and 0.14) when compared to the pre-test (0.98 and 0.38) is that students are able to provide elaborate problem solutions in the post-test, when compared to simple or obvious types of reasoning in the pre-test. Indeed, this was evident during the normality analysis of the questions for pre- and post-tests. Although the significance of the distribution of peer scores in the pre-test was 0.20, it was 0.07 in the post-test, which indicated that a valid normal distribution of data cannot be assumed below the value of 0.05. These results suggest that the quality of learning of students is also a factor. The higher scores obtained from post-test questions and the consequent bunching of scores towards the top of the range (Fig. 4) may result in decreased reliability.

The results of the generalizability study, which was conducted to define the mean number of students required to reach comparable inter-coder reliabilities, showed that 20 students on average were needed to reach a correlation coefficient of 0.80. As a review of previous research demonstrated, Magin and Helmore (2001) found five students on average to be used to reach inter-coder reliability ranging between 0.53 and 0.70, while Kilic and Cakan (2007) reported this value as about 30 students. A close examination of the variation of inter-coder reliability coefficients (Fig. 5) in this study shows that comparable inter-coder reliabilities (0.75 and 0.47 for the pre- and post-tests, respectively) can be reached with five students on average as parallel with the finding of Magin and Helmore (2001). It is interesting that the numbers of mean coders in both studies are close to reach comparable reliabilities despite the different contexts of the studies.

While the reliability of ratings by a single instructor is higher than the reliability of a single peer coder, the reliability of instructor scores is unlikely to be higher than averaged peer scores, which are based on more than 16 peer ratings for both tests. Therefore, the reliability of assessments concerning students' problem solutions can be improved by combining instructor scores with the average multiple peer scores.

In this study, sample size was analyzed to find an alpha of 0.05 and a power level of 0.80 for both the pre- and post-tests and three factors (i.e. alpha, sample size and effect size) were considered simultaneously as suggested by Hair et al. (1998). When the effect size was considered, a large effect size value ($\eta^2 = 0.39$) in the pre-test required a sample size of approximately five students (Fig. 5) to reach an acceptable power level of 0.80; however, approximately 25 students (Fig. 5) were needed to have the same power level with the effect size of $\eta^2 = 0.16$, which was large enough but close to upper limit of medium effect size in the post-test for an alpha level of 0.05. This shows us the impact of the effect size on the sample size with a constant alpha level. In other words, power becomes acceptable with lower sample sizes in cases with a large effect size at alpha level of 0.05 as in the pre-test. Yet, in the post-test, a sample of five students with an alpha of 0.05 produced the approximate power level of 0.50 (Fig. 5), since the effect size was smaller than in the pre-test. This implies that we should design the study with larger sample sizes if the effect size is found to be small to achieve the desired power level in planning the research and we might assume that 76 students involved in the experiment group is an acceptable number of sample size, which is also evident from the power value of 1.00 obtained from the ANOVA test in both the pre- and post-tests.

Fraenkel and Wallen (1996) outline that researchers may use their knowledge of population to judge whether or not a particular sample will be a representative. They warn the researchers that their judgement may not be correct in estimating the representativeness of a sample regarding the data they needed with their expertise or prior information. In fact, this study also considered the rankings of all faculties determined by the final year or graduate students' scores in a nation-wide exam (KPSS), which was conducted by the Higher Education Council to appoint graduates as teachers at schools. Since the faculty studied had a middle ranking based on the KPSS results, the probability of selecting a representative sample is considered to be high. This consideration was also supported by Fraenkel and Wallen (1996, p. 107), who argue that ''whenever purposive or convenience samples are used, generalization is made more plausible if … the sample is representative of the intended population on at least some relevant variables''. However, the researcher is aware of the fact that the sample must be sufficiently large to generalize the outcomes of the population and hence, the extent to which the results of a study can be generalized determines the external validity of the study. Therefore, to increase the external validity of this study, the researcher suggests conducting the selected design by other researchers with similar student groups.

The rubric developed appeared to be successful and relatively easy to use in a variety of questions concerning NLM unit. The rubric was able to highlight aspects of the problem solutions overall, such as drawing FBDs, representing the forces and calculating the resultant force, and showing the movement direction of the system. At the same time, the rubric used gave insight into individual students' understanding of the key concepts of the related topic and allowed access to the mental pictures formed by students of the types and causes of motion.

During the analysis of problem solutions, it was evident that students, who drew FBDs and consistently showed evidence of deep processing of the ideas in their diagrams in the experiment group, were successful in explaining the type of the motion as well as indicating the movement direction of the system. It could be argued that the use of the rubric caused the experiment students to develop the habit of drawing FBDs and to be able to analyze the components of the mechanical system examined to achieve the correct solution of the problem. Similar findings were reported in Rosengrant et al.'s (2005) study, in which the role of FBDs in successful problem solving was examined. They found that students who drew FBDs correctly were more likely to solve the problem correctly and drawing an incorrect FBD led students to more incorrect solutions than having no diagram at all.

Conclusion

On visiting the literature, there appears to be lack of studies reporting peer assessment of problem solving in a selected science topic. The study reported in this paper utilized a data set consisting of peer and instructor ratings of problem solutions on NLM. What struck the researcher was the absence of any studies examining the effects of the use of a designed rubric on student achievement in a certain area of science. Given the limitations of external validity that a study like this has, the data obtained in this study have been analyzed to provide, for the first time, information on the effectiveness of a rubric on students' achievement and on the comparative reliabilities of peer and the instructor assessments of problem solving skills. Nonetheless, the researcher believes that the usefulness of the rubric developed in problem solving and assessment has been demonstrated. Its usefulness should be examined and tested on larger and more diverse samples of students.

The constructivist approach recommends ensuring student participation in the learning process and helping students understand how to construct their own learning. Moreover, it is reported in the literature that group work and peer assessments have resulted in increased student understanding and achievement (Gatfield 1999). The students in the experiment group participated in the construction of the rubric and in the assessment of their peers'

problem solutions. The results of this study suggest that rubric use in combination with peer assessment provides an effective teaching and learning strategy for conceptual understanding of the area without learning the concepts by rote. Indeed, it was evident while rating the students' solutions that the main difficulty for the students in the control group was their inability to analyze the forces acting on the system and thereby miscalculating the resultant force to explain the behaviour of the system after teaching. Most students either did not draw an FBD or drew it incorrectly in their solutions. Although this approach resulted in writing correct mathematical equations, physical explanations and representations were unacceptable throughout the questions. This shows us that the students did not assimilate the ideas provided during teaching to understand and solve the questions in different contexts.

In order to further increase student mastery of both conceptual reasoning and quantitative problem solving and reduce the gender gap, PI can be incorporated into teaching of NLM, as Crouch and Mazur (2001) discussed in their study. Implemented effectively at different institutions and in upper-secondary courses with small and large classes (Crouch 1998; Fagen et al. 2002; Nicol and Boyle 2003), reported to be suitable to a wide range of context and instructor styles (Crouch and Mazur 2001; Green 2003) and accompanied by further increases in student understanding (Fagen 2003; Lasry et al. 2008), such an approach may involve three aspects. First, pre-class reading with free response web-based assignments due before each class can be a good opportunity for both students and the instructor to prepare for class more effectively. Secondly, a research-based textbook can be introduced for parts of the course to increase student contribution during discussions. Finally, cooperative learning activities, which are described as conceptual reasoning, hands-on and quantitative problem solving activities by Crouch and Mazur (2001), can be incorporated into teaching of the NLM to require students to be more actively involved and independent in learning throughout a mechanics course. The complementation of PI with other strategies that increase student engagement was found to yield high learning gains (Crouch et al. 2007) and the development and use of a rubric as shown in this study may serve for this purpose. Thus, the combination of PI methodology and rubric use might be employed in future research to examine further their effects on gender gap, conceptual understanding and quantitative problem solving.

The findings obtained from this study may also be used for replication with different groups taking the same course in following years, as Magin and Helmore (2001) and Hafner and Hafner (2003) also used this approach to examine the consistency of outcomes for the classes over 4 years. This might allow us to examine the trend in the peer and instructor assessment distributions over the years in terms of homogeneity of scores. In addition, students can be asked to evaluate their own problem solutions using the rubric in future studies. It may be interesting to examine students' self-assessment scores and validity of their assessments. Etkina et al. (2006) and Jackson and Larkin (2002) also consider this kind of self-assessment strategy as supporting student learning and enabling students to evaluate their own learning. Obviously, whatever research design is selected, the use of rubrics is an ideal way of implementing formative or self assessment and requires knowledge of the related content area.

While the reliability of scores given by the instructor were indeed superior to the reliability of single-peer scores and closer to the acceptable level of 0.80, instructor assessment was found to have almost equal reliability with a group of 20 peer assessments in this study. It is clear from the results of this study that by involving peer groups in the task of assessment we can achieve a better learning of the topic to be taught and assessment of problem solutions with highly satisfactory reliability. In future studies of the same context with this study, analyses of reliability and other measures could be compared.

Appendix 1

See Table 11.

Table 11 The rubric for the assessment of problem solutions to Newton's Laws of Motion questions

1. Axes	Weight: 15%
15	Axis for each object in the system was drawn completely and correctly
10	Axes for some objects in the system were drawn completely and correctly
5	Axis for each object in the system was drawn incompletely or mistakenly
3	Axes for some objects in the system were drawn incompletely or mistakenly
0	No work done

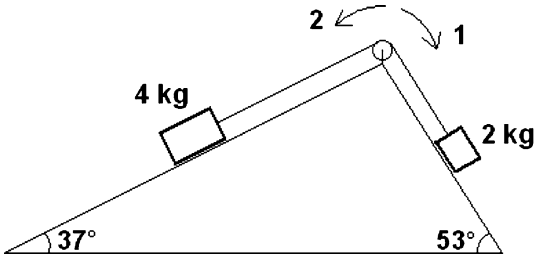
Table 11 continued

2. Drawing free-body diagrams	Weight: 35%
35	Both the system's and each object's free-body diagrams were drawn
30	Minor incompleteness in both the system's and objects' free-body diagrams existed
25	Each object's free-body diagram was drawn but the system's free-body diagram was not
15	The system's free-body diagram was drawn but objects' free-body diagram were not
10	Major incompleteness in both the system's and objects' free-body diagrams existed
5	Objects' free-body diagrams were incomplete and the system's free-body diagram was missing
3	The system's free-body diagram was incomplete and objects' free-body diagrams were missing
0	No work done
3. Representation of forces	Weight: 20%
20	Directions and magnitudes of the forces on the system were drawn completely and correctly
15	Directions and magnitudes of the forces on the system were drawn incompletely
10	Directions and magnitudes of some forces on the system were drawn mistakenly
5	Directions and magnitudes of the forces on the system were drawn mistakenly
0	No work done
4. Type and direction of motion	Weight: 20%
20	Resultant force was found correctly with a correct notation of type and direction of motion
15	Resultant force was found correctly with a correct notation of type or direction of motion
10	Resultant force was found mistakenly with a correct notation of type and/or direction of motion
5	Resultant force was found mistakenly with an incorrect notation of type and direction of motion
0	No work done
5. Solutions for equations	Weight: 5%
5	Equations for unknown variables were written and a correct result was obtained
4	Equations for unknown variables were written but an incorrect result was obtained
2	Equations for all unknown variables were not written but a correct result was obtained
1	Written equations and the result were incorrect
0	No work done
6. Units	Weight: 5%
5	Each term was used in the same and its own system of unit
4	Some terms were used in the same and their own system of unit
3	Terms were used in the different but their own system of unit
2	Some terms were used in the correct system of unit but shown with incorrect symbols
0	No work done

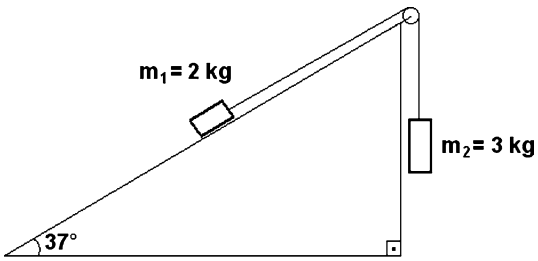
Appendix 2

Pre- and post-test questions

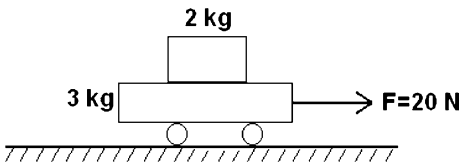
1. Find the movement direction and acceleration value of the frictionless system shown above if it is released? (sin37 = 0.6, cos37 = 0.8, sin53 = 0.8, cos53 = 0.6, $g = 10$ m/s^2)



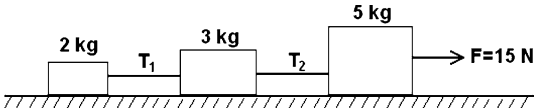
2. The values of masses m_1 and m_2 are 2 and 3 kg, respectively, in the system shown above. Friction coefficient between the inclined plane and mass m_1 is 0.5. If the system is released, find the values of acceleration and tension in the string. (sin37 = 0.6, cos37 = 0.8, $g = 10$ m/s^2)



3. Suppose the friction between the car and the horizontal surface can be ignored in the figure shown above. When the force of 20 N acts on the car, find the minimum value of the friction coefficient between the car of a mass of 3 kg and the block of a mass of 2 kg so as to keep the block at rest on the car.



4. A constant force of 15 N is applied to one of the blocks of masses 2, 3 and 5 kilograms as shown above. If the coefficient of friction between the blocks and the horizontal surface is 0.1, find the values of acceleration and tensions in the strings.



References

Andaloro G, Bellomonte L, Sperandeo-Mineo RM (1997) A computer-based learning environment in the field of Newtonian mechanics. Int J Sci Educ 19(6):661–680

Bakeman R (2005) Recommended effect size statistics for repeated measures designs. Behav Res Methods 37(3):379–384

Bednarski M (2003) Assessing performance tasks. Sci Teach 70(4):34

Caleon I, Subramaniam R (2005) The impact of a cryogenics-based enrichment programme on attitude towards science and the learning of science concepts. Int J Sci Educ 27(6):679–704

Crouch C (1998) Peer instruction: an interactive approach for large classes. Opt Photonics News 9(9):37–41

Crouch CH, Mazur E (2001) Peer instruction: ten years of experience and results. Am J Phys 69(9):970–977

Crouch CH, Watkins J, Fagen AP, Mazur E (2007) Peer instruction: engaging students one-on-one, all at once. In: Redish EF, Cooney PJ (eds) Research-based reform of university physics: reviews in PER, vol. 1. Retrieved August 21, 2009 from American Association of Physics Teachers, College Park, MD, http://www.per-central.org/document/ServeFile.cfm?ID=4990

D'amico EJ, Neilands TB, Zambarano R (2001) Power analysis for multivariate and repeated measures designs: a flexible approach using the SPSS MANOVA procedure. Behav Res Methods Instrum Comput 33(4):479–484

Davidowitz B, Rollnick M, Fakudze C (2005) Development and application of a rubric for analysis of novice students' laboratory flow diagrams. Int J Sci Educ 27(1):43–59

Ebel RL (1951) Estimation of the reliability of ratings. Psychometrika 16(4):407–424

Etkina E, Van Heuvelen A, White-Brahmia S, Brookes DT, Gentile M, Murthy S, Rosengrant D, Warren A (2006) Scientific abilities and their assessment. Phys Rev Spec Top Phys Educ Res 2(2): 1–15

Fagen AP (2003) Assessing and enhancing the introductory science course in physics and biology: peer instruction, classroom demonstrations and genetics vocabulary. PhD thesis, Harvard University

Fagen AP, Crouch CH, Mazur E (2002) Peer instruction: results from a range of classrooms. Phys Teach 40(4):206–209

Falchikov N (1995) Peer feedback marking: developing peer assessment. Innov Educ Train Int 32(2):175–187

Ferguson GA (1971) Statistical analysis in psychology and education. McGraw-Hill, New York

Fowler SR, Zeidler DL, Sadler TD (2009) Moral sensitivity in the context of socioscientific issues in high school science students. Int J Sci Educ 31(2):279–296

Fraenkel JR, Wallen NE (1996) How to design and evaluate research in education, 3rd edn. McGraw-Hill, New York

Freeman M (1995) Peer assessment by groups of group work. Assess Eval High Educ 20:289–299

Gaigher E, Rogan JM, Braun MWH (2007) Exploring the development of conceptual understanding through structured problem-solving in physics. Int J Sci Educ 29(9):1089–1110
Gatfield T (1999) Examining student satisfaction with group projects and peer assessment. Assess Eval High Educ 24:365–377
Gillmore GM (2000) Drawing inferences about instructors: the inter-class reliability of student ratings of instruction. Office of Educational Assessment, report no. 00-02, University of Washington, US
Gravetter FJ, Wallnau LB (1996) Statistics for the behavioral sciences: a first course for students of psychology and education, 4th edn. West Publishing Company, Minneapolis
Green PJ (2003) Peer instruction for astronomy. Prentice Hall, Upper Saddle River
Green SB, Salkind NJ, Akey TM (2000) Using SPSS for Windows: analyzing and understanding data. Prentice Hall, New Jersey
Hafner JC, Hafner PM (2003) Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating. Int J Sci Educ 25(12):1509–1528
Hair JF, Anderson RE, Tatham RL, Black WC (1998) Multivariate data analysis. Prentice Hall, New Jersey
Herman JL, Aschbacher PR, Winters L (1992) A practical guide to alternative assessment. Association for Supervision and Curriculum Development, Alexandria
Hohenshell LM, Hand B (2006) Writing-to-learn strategies in secondary school cell biology: a mixed method study. Int J Sci Educ 28(2&3):261–289
Hughes IE, Large BJ (1993) Staff and peer group assessment of oral communication skills. Stud High Educ 18:379–385
Jackson CW, Larkin MJ (2002) Teaching students to use grading rubrics. Teach Except Child 35(1):40–45
Jimoyiannis A, Komis V (2003) Investigating Greek students' ideas about forces and motion. Res Sci Educ 33:375–392
Kilic GB, Cakan M (2007) Peer assessment of elementary science teaching skills. J Sci Teacher Educ 18(1):91–107
Kinnear PR, Gray CD (2008) SPSS 16 made simple. Psychology Press, Hove
Kwan K, Leung R (1996) Tutor versus peer group assessment of student performance in a simulation training exercise. Assess Eval High Educ 21:205–215
Lasry N, Mazur E, Watkins J (2008) Peer instruction: from Harvard to the two-year college. Am J Phys 76(11):1066–1069
Litzinger TA, Lee SH, Wise JC, Felder RM (2007) A psychometric study of the index of learning styles. J Eng Educ 96(4):309–319
Lorenzo M, Crouch CH, Mazur E (2006) Reducing the gender gap in the physics classroom. Am J Phys 74(2):118–122
Luft JA (1999) Rubrics: design and use in science teacher education. J Sci Teacher Educ 10(2):107–121
MacAlpine JMK (1999) Improving and encouraging peer assessment of student presentations. Assess Eval High Educ 24:15–25
Magin D, Helmore P (2001) Peer and teacher assessments of oral presentation skills: how reliable are they? Stud High Educ 26(3):287–298
Mazur E (1997) Peer instruction: A user's manual. Prentice Hall, Upper Saddle River, NJ
McClure R, Johnson B, Jackson D, Hoff J (2000) Assessing the constructivist classroom. (ERIC No. ED443459) Retrieved September 18, 2007 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/55/50.pdf
McCollister S (2002) Developing criteria rubrics in the art classroom. Art Educ 55(6):46
Mildenhall PT, Williams JS (2001) Instability in students' use of intuitive and Newtonian models to predict motion: the critical effect of the parameters involved. Int J Sci Educ 23(6):643–660
Nicol DJ, Boyle JT (2003) Peer instruction versus class-wide discussion in the large classes: a comparison of two interaction methods in the wired classroom. Stud High Educ 28(4):458–473
Parker J, Heywood D (2000) Exploring the relationship between subject knowledge and pedagogic content knowledge in primary teachers' learning about forces. Int J Sci Educ 22(1):89–111
Pomplun M, Capps L (1999) Gender differences for constructed response mathematics items. Educ Psychol Measur 59:597–614
Pomplun M, Sundbye N (1999) Gender differences in constructed response reading items. Appl Measur Educ 12:95–109
Popham WJ (1997) What's wrong- and what's right-with rubrics. Educ Leadership 55(2):72–75
Potvin PJ, Schutz RW (2000) Statistical power for the two-factor repeated measures ANOVA. Behav Res Method Instrum Comput 32(2):347–356
Rosengrant D, Van Heuvelen A, Etkina E (2005) Free-body diagrams: necessary or sufficient? Am Inst Phys Conf Proc 790(1):177–180
Rutherford S (2007) Using a laboratory conclusion rubric. Sci Act Classroom Projects Curr Ideas 43(4):9–14
Serway RA, Beichner RJ (2000) Physics for scientists and engineers. Saunders College Publishing, Orlando
Shaka FL, Bitner BL (1996) Construction and validation of a rubric for scoring concept maps. In: AETS conference papers and summaries of presentations, pp 650–669. Retrieved October 31, 2007 from http://www.ed.psu.edu/CI/journals/96pap43.htm
Shymansky JA, Yore LD, Treagust DF, Thiele RB, Harrison A, Waldrip BG, Stocklmayer SM, Venville G (1997) Examining the construction process: a study of changes in level 10 students' understanding of classical mechanics. J Res Sci Teach 34(6):571–593
Stefani LAJ (1994) Peer, self and tutor assessment: relative reliabilities. Stud High Educ 19:69–75
Stevens J (1996) Applied multivariate statistics for the social sciences. Erlbaum, Mahwah
Stoll SA (2003) Assessing elementary students. Strategies 16(1):33
Topping K (1998) Peer assessment between students in colleges and universities. Rev Educ Res 68(3):249–276
Trumper R, Gorsky P (1997) A survey of biology students' conceptions of force in pre-service training for high school teachers. Res Sci Technol Educ 15(2):133–149