Scientia in educatione 8(Special Issue), 2017
p. 182–190
ISSN 1804-7106

Coding Scheme for Assessment of Students' Explanations and Predictions

Mihael Gojkošek, Gorazd Planinšič, Josip Sliško

Abstract

In the process of analyzing students' explanations and predictions for interaction between brightness enhancement film and beam of white light, a need for objective and reliable assessment instrument arose. Consequently, we developed a coding scheme that was mostly inspired by the rubrics for self-assessment of scientific abilities. In the paper we present the grading categories that were integrated in the coding scheme, and descriptions of criteria used for evaluation of students work. We report the results of reliability analysis of new assessment tool and present some examples of its application.

Key words: coding scheme, assessment, rubrics, explanation, prediction.

Introduction

Fundamental features of scientific work in physics are building explanations and on them based testable predictions (Giere, 1997). Therefore, in order to learn science by doing, students should be involved in authentic scientific tasks that include construction of explanations and predictions. Especially students, who are proficient in science, should be able to generate and evaluate scientific evidence and explanations (Duschl, Schweingruber & Shouse, 2007).

More than 600 high-school and university students from Slovenia and Czech Republic were tested during several phases of the extended research on students' ability to construct explanations and predictions for an unknown physics phenomenon. Consequently, the need for robust and reliable assessment tool arose. In this paper we present the process of development of the coding scheme that was used to evaluate the quality of students' explanations and predictions. The paper also addresses the reliability of the coding scheme and demonstrates some examples of its application.

Theoretical framework

In the process of development of the coding scheme we were mainly inspired by previous work of Eugenia Etkina and her co-workers. They have developed the tasks and rubrics for formative self-assessment in order to help students to perform better and thus develop scientific abilities (Etkina et al., 2006). Their rubrics are based on cognitive apprenticeship theory and address 7 areas of scientific abilities that scientists use when they construct knowledge and solve problems. These areas include the abilities (1) to represent information in multiple ways, (2) to design and conduct an observational experiment, (3) to design and conduct a testing experiment, (4) to design and conduct an application experiment, (5) to collect and analyze experimental data, (6) to engage in divergent thinking, and (7) to evaluate models, equations, solutions, and claims. Each of 7 rubrics consists of multiple categories that assess specific subabilities (e.g. "Is able to make a reasonable prediction based on a hypothesis."). Each category is further supplemented with detailed description of qualitative criteria that one should possess to be classified in one of four grading levels: "Missing", "Inadequate", "Needs some improvement" and "Adequate". Rubrics for assessment of scientific abilities were later used in several other studies (e.g. Etkina, Karelina & Ruibal-Villasenor, 2008; Etkina et al., 2009) and turned out to be a highly efficient tool. Although the purpose of our assessment differed from the Etkina's, we found the basic form of the rubrics very useful. We have re-designed the set of categories (subabilities) included in rubrics and adapted the criteria descriptions to best fit our needs.

Research instruments

Brightness enhancement film (BEF)

Brightness enhancement film is an interesting optical element that can be used in several demonstrational experiments suitable for introductory optics course (Planinšič & Gojkošek, 2011). It is one of the thin transparent foils from the backlight system in LCD monitors and can be easily obtained by dismounting any used monitor. The main advantages of using BEF in demonstrational experiment are a) it is an unknown element to vast majority of students and experts, and b) its structure cannot be seen with naked eye.

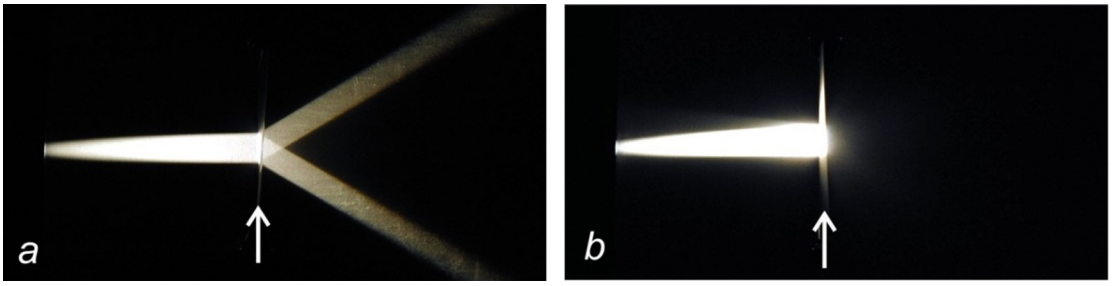

Figure 1: a) The split of the light beam incident perpendicularly to one side of the film, and b) the reflection of the light beam incident perpendicularly to the other side. The arrows show the position of the brightness enhancement film

We integrated two demonstrational experiments with BEF in our testing procedure. Both experiments include a beam of white light (produced by a flashlight) incident perpendicularly to the sides of the film. On one side, the beam of light is split into two symmetrical beams (Figure 1a), while the beam incident perpendicularly to the other side of the film is mostly reflected into the direction of origin (Figure 1b).

The structure of the film can be easily revealed using the school microscope. A magnified cross-section shows that BEF is flat on one side and has microscopic prismatic ridges with the apex angle of approx. 90° on the other side (Figure 2).

Now we can also explain observed outcomes of both demonstration experiments. Light incident perpendicularly to the prismatic side of BEF is refracted in two directions — depending on which side of the prisms the beam strikes (Figure 3a). The light beam incident perpendicularly to the flat side of BEF undergoes double total internal reflection and returns back into the original direction (Figure 3b).

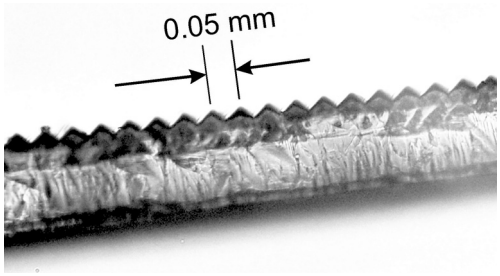

Figure 2: Cross-section of the brightness enhancement film under the microscope reveals prismatic structure

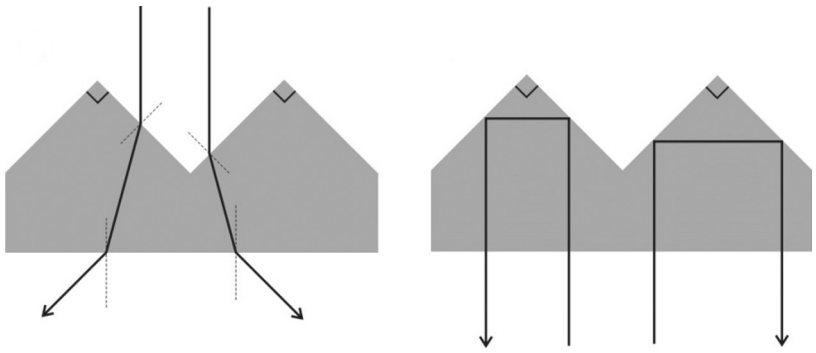

Figure 3: a) Double refraction of the light beam incident on the prismatic ridges, and b) double total internal reflection of the light beam incident perpendicularly to the flat side of BEF

Note that these demonstrational experiments can be combined into two different two step sequences, depending on which experiment is first shown to students. We named them split-reflection (or shorter SR) task sequence (when first the split of light beam was shown to students and then the reflection) and reflection-split (RS) task sequence (when first experiment demonstrated the reflection and second the split).

Foil test

Students were tested with foil test, which was designed by our research group. One part of the foil test was two demonstrational experiments with the BEF described above. First, a teacher showed students one of both experiment (split in SR and reflection in RS task sequence). Then they were asked to construct one or more explanations for interaction of light beam and the BEF on the basis of observed outcome. We encouraged them to present their explanations verbally (text description) and graphically (sketch). Additionally, students had to name optical phenomenon/a, that is on their opinion involved in observed experiment.

Next, students were informed about the second experiment, in which light beam will be incident perpendicularly to the other side of the BEF. They were asked to construct a prediction for experimental outcome on the basis of their previously proposed explanation. Again, their prediction should consist of verbal and graphical part. Teacher later performed second demonstrational experiment (reflection in SR and split in RS task sequence) and asked students, weather their prediction agrees with observed outcome. Finally, students had one more opportunity to construct the improved explanation compatible with the outcomes of both demonstrational experiments.

Lawson's Classroom Test of Scientific Reasoning

As a reference test, Lawson's Classroom Test of Scientific Reasoning (CTSR) was used. A 24-item multiple-choice version of the test was translated into Slovene and used to classify students as concrete-operational, transitional and formal-operational reasoners according to their scores.

Development of coding scheme

Purpose

Previous research showed that majority of students is not able to reveal the actual structure of the BEF on the basis of two demonstrational experiments. Even more, the proportion of those who manage to do so remains low (less than 5 %) even if students are previously involved in pedagogical activity with macroscopic prism and laser ray-box (Gojkošek, Sliško & Planinšič, 2013). Therefore, we wanted to construct a reliable and objective tool for assessment of the quality of students' explanations regardless of their (mis)match with the actual structure of the BEF. Note that observed experimental outcomes can also be explained e.g. with suitable arrangement of reflecting surfaces. Our goal was to develop a set of categories, with which students' explanations and predictions could be easily assessed, and would allow obtaining overall quality grade and further calculation of students' average success.

Grading categories

Our coding scheme consists of three main parts that are formulated for assessment of initial explanation, prediction and improved explanation, respectively. Each part further consists of 4 or 5 categories that assess students' abilities that are needed to solve the task successfully. Assessment categories are presented in Table 1.

Table 1: Categories for assessment of initial/improved explanation and prediction

Initial explanation
Graphical representations
Verbal representations
Correct use of physics
Consistency between outcomes predicted by explanation and observed outcomes
Number of different models
Prediction
Graphical representations
Verbal representations
Consistency with initial explanation
Ability to evaluate agreement of prediction and observed outcome
Improved explanation
Graphical representations
Verbal representations
Correct use of physics
Addressing asymmetry
Consistency between outcomes predicted by explanation and observed outcomes

Devising code descriptions

After selection of grading categories included in our coding scheme, we devised detailed descriptions of codes. We decided to keep 4-level coding scale used by Etkina et al. as well as descriptive names of grading levels: 0-Missing, 1-Inadequate, 2-Needs some improvement and 3-Adequate. Descriptions of students' work that merit a particular grading level can be found below.

Grading categories for initial explanation

In category "graphical representations", basic drawing elements of the sketch were assessed. We were looking for the structure of the foil (its cross-section), light rays and majority of labels. If these were present, sketch was coded with 3, while the sketch without labels was coded with 2. Any other sketch was coded with 1 and no sketch with 0.

Also in the category "verbal representations", we expected from students to describe foil structure and name involved optical phenomenon. When both included, code 3 was assigned, while for one of them code 2 was used. Other verbal descriptions were considered as "inadequate" and no text was coded with "missing".

When assessing correct use of physics, both graphical and verbal parts of explanation were considered. When optical phenomenon was applied without mistakes, code 3 was used. Misapplication of the phenomenon was coded with 2. Typical

students' mistakes include split of the light beam by diverging lens or diffraction grating and total internal reflection of the light incident perpendicularly to the inner surface of a medium. Confusing, contradictory or incomprehensible application of optical phenomenon (e.g. "lens reflects light") were coded with 1 and when no optical phenomenon was included in explanation code 0 was assigned.

We also assessed the consistency between outcomes predicted by explanation and observed outcomes. Particular attention was devoted to the direction of incident and outgoing light rays. If explanation and observed result were consistent, code 3 was assigned, while discrepancy between them was coded with 2. When student's explanation failed to reproduce the main experimental result (split or reflection) code 1 was used, while code 0 was given to explanations that had nothing in common with observed experimental result.

In the grading category "number of different models", two or more explanations that employed different optical phenomenon merit code 3. When the same phenomenon was applied in several explanations, code 2 was assigned. One explanation was coded with 1 and no explanation with 0.

Grading categories for prediction

In assessment categories "graphical and verbal representations", evaluation criteria for prediction were the same as for initial explanation coding. Next grading category assessed consistency between prediction and initial explanation. Prediction that was consistently derived from previously proposed explanation was coded with 3. Inconsistent derivation from initial explanation merit code 2, while any other prediction was coded with 1 and no prediction with 0.

Grading category "ability to evaluate agreement of prediction and observed outcome" assessed students' report about (mis)match of predicted and observed outcome of second experiment. Reasonable decision about agreement/disagreement was coded with 3, while code 2 was assigned when one made a decision about agreement/disagreement that evaluator was unable to judge due to imprecise prediction. When this decision was clearly incorrect, code 1 was assigned, while no agreement assessment was coded with 0.

Grading categories for improved explanation

Similar to previous grading, in assessing graphical representations we were looking for structure of the film, light rays describing both experimental results and majority of labels. Sketch that included all these elements was coded with 3. Film's structure and light rays for both experiments were enough for code 2, while the sketch without one of these elements was coded with 1. For no sketch code 0 was assigned.

Category "verbal representations" addressed presence of verbal description of film's structure and optical phenomena involved in both experiments. When all these elements were present, explanation was coded with 3. If only description of the structure or only optical phenomena was present, or there were both for explanation of just one experiment, code 2 was assigned. Every other verbal explanation was coded with 1, and code 0 was used when no text was present.

For assessment category "correctness of physics" we used the same criteria as for initial explanation coding. With category "addressing asymmetry" we assessed the way in which asymmetrical behavior of the BEF was explained. Code 3 was assigned when film's asymmetrical properties were explained in consistent way. If

asymmetry was provided through mechanical composition of two optical elements, explanation was coded with 2. Code 1 was used when asymmetry was granted but not explained, and code 0 was assigned when asymmetry was not addressed.

In improved explanation, we also assessed consistency between outcomes predicted by explanation and observed outcomes. Similar to coding of initial explanation, code 3 was assigned when explanation and observed results were consistent. Code 2 was used when direction of incident/outgoing light beams were misinterpreted, while code 1 was assigned to explanatory models that failed to reproduce main experimental outcomes — split and reflection of incident light beams. If incident or outgoing light beams were not drawn, code 0 was assigned.

Analysis of reliability

Tests of 197 students from Slovenian high-schools were assessed with described coding scheme. Approximately 20 % of all tests were independently evaluated by two researchers. Their coding matched in 90 % of all cases. Also inter-rater agreement coefficients like Cohen's kappa ($\kappa = 0.87$) and Pearson's correlation coefficient ($r = 0.92$) indicate high reliability of this assessment tool.

Combined grades

As mentioned, one of our goals was to obtain combined grades for overall quality of students' explanations and predictions. Before that, some assumptions needed to be taken into account. First, we assumed scale nature of grading levels. As a consequence of that assumption, one can summarize and calculate average grades for different categories. And secondly, weights suitable to importance of each grading category needed to be set. Since in our opinion all addressed categories play similarly important role in overall quality of explanations and predictions, all weights have been set to 1. Combined grade for the quality of initial explanation is consequently calculated as a sum of grades of all five categories that assess this explanation. Similarly combined grades for the quality of prediction and improved explanation are calculated by summarizing grades of individual categories.

Some examples of application and obtained results

Using grades achieved in single grading category and combined grades, we were able to compare different groups of students according to scientific reasoning ability level (concrete/transitional/formal) and task sequence they were involved in (SR/RS). Our results suggest that difference between concrete-operational and formal-operational reasoners is statistically significant for some categories and insignificant for others. An example of grading category in which this difference was among highest is correct use of physics in improved explanation. Average grades achieved in this category can be found in table 2. Mann-Whitney nonparametric U-test revealed that difference between concrete- and formal-operational groups are highly statistically significant in both, SR and RS task sequences ($U = 50$, $p = 0.002$, and $U = 137$, $p = 0.001$, respectively). On the other hand, in the category "number of different models" no significant difference between these groups was observed ($U = 126$, $p = 0.31$ in SR, and $U = 276$, $p = 0.75$ in RS task sequence).

Table 2: Average grades achieved in the category "correct use of physics" in improved explanation and "number of different models" in initial explanation

	split-reflection (SR)		reflection-split (RS)	
	concrete thinkers	formal thinkers	concrete thinkers	formal thinkers
improved explanation: verbal representations	1.4	1.8	1.2	1.5
initial explanation: number of different models	1.3	1.2	1.2	1.2

Table 3: Average combined grades for the quality of improved explanation

	split-reflection (SR)		reflection-split (RS)	
	concrete thinkers	formal thinkers	concrete thinkers	formal thinkers
improved explanation: combined grade for quality	5.7	8.5	4.1	7.4

Significant difference between concrete-operational and formal-operational thinkers was found also by comparison of combined grades for the quality of improved students' explanations (Table 3). Again, Mann-Whitney U-test was used to calculate the significance of these differences in SR ($U = 55.5$, $p = 0.010$) and RS task sequences ($U = 115.5$, $p = 0.000$).

Conclusions

In our study, high-school students' ability to construct explanations and on them based predictions was taken under examination. For that purpose students were involved in testing procedure with two demonstrational experiments, in which interaction between brightness enhancement film (BEF) and beam of white light was presented. Students were asked to propose possible explanations for observed interaction and to predict the outcome of the second experiment. During the analysis of students' tests the need for objective assessment tool arose. We decided to develop a coding scheme based on the rubrics for assessment of scientific abilities (Etkina et al., 2006, 2009; Etkina, Karelina & Ruibal-Villasenor, 2008) that would allow obtaining reliable grades for the quality of students' explanations and predictions.

Developed coding scheme consists of three separate rubrics that assess students' initial explanation, prediction and improved explanation, respectively. Each rubric further consists of grading categories that assess students' work in explanation and prediction formation. Four-level grading scale is used to evaluate each grading category. Categories are equipped with detailed descriptions of essential elements that need to be present to merit a particular level. Combined grades for the quality of students' explanations and predictions are obtained by summarizing grades of categories in one rubric.

We conclude that rubric-like coding scheme is an effective tool for assessment of students' explanations and predictions. Developed coding scheme shows high level of reliability assessed through inter-rater agreement coefficients. Under some assumptions, grading categories of the coding scheme can be used to evaluate overall quality of students' explanations/predictions and their average performance.

Acknowledgement

Authors would like to thank Bor Gregorčič for the help with evaluation of students' tests, and Eugenia Etkina for helpful discussion in the process of development of coding scheme.

References

Duschl, R. A., Schweingruber, H. A. & Shouse, A. W. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.

Etkina, E., Van Heuvelen, A., White-Brahmia, S., Brookes, D. T., Gentile, M., Murthy, S., Rosengrant, D. & Warren, A. (2006). Scientific abilities and their assessment. *Phys. Rev. Spec. Top.*, 2, 020103.

Etkina, E., Karelina, A. & Ruibal-Villasenor, M. (2008). How long does it take? A study of student acquisition of scientific abilities. *Phys. Rev. Spec. Top.*, 4, 020108.

Etkina, E., Karelina, A., Murthy, S. & Ruibal-Villasenor, M. (2009). Using action research to improve learning and formative assessment to conduct research. *Phys. Rev. Spec Top.*, 5, 010109.

Giere, R. N. (1997). *Understanding scientific reasoning.* 4th edition. Orlando: Harcourt Brace College Publishers.

Gojkošek, M., Sliško, J. & Planinšič, G. (2013). Do learning activities improve students' ability to construct explanatory models with a prism foil problem? *CEPS Journal*, 3(3), 9–28. Available at http://www.cepsj.si/pdfs/cepsj_3_3/cepsj_pp_9-28_Gojkosek.pdf

Planinšič, G. & Gojkošek, M. (2011). Prism foil from an LCD monitor as a tool for teaching introductory optics. *Eur. J. Phys.*, 32(2), 601–613.

Mihael Gojkošek
Gorazd Planinšič
Faculty of Mathematics and Physics, University of Ljubljana, Slovenia

Josip Sliško
Facultad de Ciencias Fisico Matematicas, Benemerita Universidad Autonoma de Puebla, Mexico