

Proceedings Paper

Paweł Bernard¹ / Karol Dudek-Różycki¹

Influence of training in inquiry-based methods on in-service science teachers' reasoning skills

¹ Jagiellonian University, Department of Chemical Education, 2 Gronostajowa Str., 30-387 Kraków, Poland, E-mail: pawel.bernard@uj.edu.pl. <https://orcid.org/0000-0002-8618-3447>, <https://orcid.org/0000-0003-3518-2089>.

Abstract:

Reasoning skills can be described as abilities to use scientific knowledge to identify and solve problems, and to formulate conclusions based on empirical observations. The European Union promotes the development of reasoning skills as a way to grow knowledgeable and responsible citizens. In a school classroom, the development of reasoning skills can be stimulated by the use of inquiry-based methods. Students' reasoning skills are frequently measured and compared by many national and international programmes (e.g. PISA). It is obvious that teachers who help students to develop reasoning skills should also have a high level of those skills, however, the teachers who currently work in Central/Eastern European schools have had no opportunity to learn through or even experience inquiry, neither to participate in programmes measuring the level of reasoning skills. In this research, Lawson's tests were used to measure the level of reasoning skills among science teachers ($n = 71$) before and after training in using inquiry-based methods. The results showed that reasoning skills are the best developed by physics teachers. Training in using inquiry helps biology and chemistry teachers improve and close the gap between them and physics teachers.

Keywords: inquiry-based education, in-service science teachers, quantitative research, reasoning skills

DOI: 10.1515/cti-2018-0023


Introduction

The European Union policy promotes the development of a knowledge society, into which science education is expected to make a significant contribution. In line with this policy, the Rocard's report (Rocard et al., 2007) strongly advocates inquiry-based methods in education (Inquiry Based Science Education – IBSE) to increase students' interest in science and boost the number of graduates in science, as well as technical and mathematical studies. One of the main objectives of the IBSE is to improve students' thinking abilities, as inquiry-based methods require more intensive reasoning than traditional science education (Csapó et al., 2012). In years 2008–2015, “Supporting and coordinating actions on innovation in the classroom: Dissemination and use of inquiry-based teaching methods on a large scale in Europe” was one of the key elements in the “Science in Society” area in the European Union's Seventh Framework Programme (FP7). Therefore, many teachers had an opportunity to participate in pan-European projects focused on teacher training and the introduction of the inquiry-based methods into school practice. So far, unfortunately, there has been limited information on how those actions influenced teachers' skills and classroom practices.

Reasoning skills

Reasoning skills (RS) are in general thinking skills and can be described as abilities to use scientific knowledge to identify and solve problems, and to formulate conclusions based on empirical observations. They are the abilities to understand the character and features of science (Dylak, 2011) and the relationship between the cause and the effect. RS are integrally related to the educational system, both in lower and upper secondary schools, especially with respect to natural sciences and mathematics. In general, reasoning skills may be divided into three major groups due to their complexity (Csapó et al., 2012; Johnson-Laird, 2006), as presented in Table 1.

Paweł Bernard is the corresponding author.

 © 2019 IUPAC & De Gruyter.

This work is licensed under the Creative Commons Attribution 4.0 Public License.

Table 1: Classification of reasoning skills.

Basic reasoning skills – Piagetian reasoning (Caroll, 1993) or operational reasoning (Csapó, 1992).	Higher order thinking skills (Williams, 1999):	Scientific reasoning – the most advanced form of human thinking.
<p>Skills with a clear operational structure, often mathematically described, e.g.:</p> <ul style="list-style-type: none"> – Identification of the variables and exploration of the relationships between them (Kuhn, Pease, & Wirkala, 2009). – Combinatorial reasoning (Kishta, 1979; Lockwood, 2013; Schröder, Bödeker, Edelstein, & Teo, 2000) that covers controlling and manipulating variables and examining their dependencies. – Probabilistic reasoning (Giroto & Gonzalez, 2008; Jones, Langrall, Thornton, & Mogill, 1997) concerns risk estimation and might be considered an introduction to correlational reasoning (Kuhn, Phelps, & Walters, 1985; Lawson, Adi, & Karplus, 1979; Ross & Cousins, 1993; Schröder et al., 2000) and statistical thinking (Chance, 2002). 	<p>Complex thinking processes, e.g.:</p> <ul style="list-style-type: none"> – Analogical reasoning – the ability to apply knowledge mastered in one context to a new similar situation (Abdellatif, Cummings, & Maddux, 2008; Klauer, 1989; Polya, 1968). – Problem-solving – analysis of complex tasks in non-transparent situations, where the solution cannot be seen immediately, requires integrating information from different sources (Frensch & Funke, 1985). – Critical thinking – systematic use of several reasoning skills such as questioning and looking for proofs and evidences to organize an argument or to establish firm foundations for a judgment. 	<p>The use of abstraction and symbols to represent phenomena via variables and dimensions, e.g.:</p> <ul style="list-style-type: none"> – Deductive and inductive reasoning (Csapó, 1997; Hamers, de Koning, & Sijtsma, 1998; Polya, 1968; Watters & English, 1995). – Argumentation which requires organizing facts and figures, carrying out logical operations and establishing causal relationships between the observed changes. – Hypothesis generation and testing. – Designing experiments. – Analyzing the results. – Making generalizations (Adey, 1999; Howson & Urbach, 1996; Koerber, Sodian, Thoermer, & Nett, 2005; Venville, Adey, Larkin, & Robertson, 2003).

Assessment of reasoning skills

The assessment of students' scientific reasoning competences is the objective of many projects and studies, for example, the Programme for International Student Assessment (PISA). It was established in 1997 and is being coordinated by the Organization for Economic Co-operation and Development (OECD 2003a; 2003b; 2006; 2009; 2013). In general, the PISA investigates the level of students' (age over 15) skills regarding problem-solving, reading comprehension, mathematical reasoning, reasoning in the natural sciences in different countries. PISA tests are connected to the RS in many aspects. The tests consisting of multiple-choice questions investigate students' abilities to solve problems using the scientific and logical thinking. The PISA's interest is not limited to the skills, as it also covers students' knowledge. The PISA assessment in 2015 aimed also at measuring collaborative problem solving (OECD, 2012). Evidence-Based Reasoning Assessment System (EBRAS) is another programme. It focuses on how students can use the evidence to support arguments (Brown, Nagashima, Fu, Timms, & Wilson, 2010). Also, the cognitive domain of the TIMSS (Trends in International Mathematics and Science Study) measures knowing, applying, and reasoning among school students all over the world (Mullis & Martin, 2013). Moreover, starting from 2019, the digitally-based eTIMSS will include extended problem solving and inquiry tasks, designed to simulate real world and laboratory situations, where students can integrate and apply process skills and content knowledge to solve mathematical problems and carry out scientific experiments or investigations (Centurino & Jones, 2017).

The assessment of the reasoning skills is a complex task, but there are several dedicated frameworks and instruments (Amsel et al., 2008; Drummond & Fischhoff, 2015; Gormally, Brickman, & Lutz, 2012; Kind, 2013; Lovelace & Brickman, 2013; Zhou et al., 2016). The best-known tools for measuring RS include the Lawson's Classroom Test of Scientific Reasoning (Lawson, 2000a). The set of tests was first published in 1978, and their second (revised) version in 2000. The tests were developed for secondary school and college age students (Lawson, 1978). Each test consists of multiple choice questions, and it is specially designed for biology, chemistry, and physics. All the tests include from 24 to 32 questions, but the first 6 questions are the same for all subjects. Those common questions are based on general science knowledge, and the following questions are subject-specified. When answering the questions, the students have to imagine a specific situation and deal with certain problems – for example, they must decide what the probability of drawing a particular colour of a randomly selected piece of wood is and justify their choice marking the correct answer (Lawson 2000a; 2000b; 2003; 2005; Pyper, 2011). The validity of Lawson's test was the subject of several studies which showed that the instrument has a good overall reliability, checked with both the individual and pair scoring methods (Cronbach $\alpha > .8$) (Bao, Xiao, Koenig, & Han, 2018; Pratt & Hacker, 1984; Stefanich, Unruh, Perry, & Phillips, 1983).

The science teachers' responsibility is to help students to develop the reasoning skills as much as possible to minimize students' errors in solving problems requiring higher order thinking processes (Utomo, Narulita, & Shimizu, 2018). It should be remembered that teachers who currently work at schools in many European countries (e.g. in Poland) finished studies that were not oriented on RS development, and as the students they didn't participate in RS measuring programs, practice solving reasoning tests, and had no opportunity to experience or learn through inquiry (Bernard, Maciejowska, Odrowąż, Dudek, & Geoghegan, 2012; Dudek, Bernard, & Odrowąż, 2015). The aim of this research was to answer the question: *What is the influence of short-term training in the IBSE on science teachers' reasoning skills?*. Even though the teachers didn't participate in inquiry-oriented classes, they finished master studies in a chosen science subject (that is obligatory in Poland as well as in many European countries), and it can be assumed that had many opportunities to develop various kinds of RS. Therefore, participation in short-term IBSE training may fill the possible "gap in learning through inquiry", and help teachers to unlock their potential by reorganizing and delivering missing elements of scientific inquiry to the reasoning process. To answer the research question teachers' RS had to be measured before and after the training in the IBSE. Since the influence of the training could be determined by teacher's background – the subject of specialization – the various science subjects' groups were compared separately.

Methodology of research

Background of the research

The research comprised the assessment of science teachers' reasoning skills before and after the training in the IBSE organized as a part of the FP7 project – Strategies for Assessment of Inquiry Learning in Science (SAILS). The training was carried out in the form of a 5-day-long summer/winter school (33 contact hours of training) and it included lectures, seminars, and laboratory classes. The teachers were introduced into the basics of the IBSE methodology, practised formulating research questions, developing hypotheses, designing and carrying out experiments, collecting data, making conclusions and forming coherent arguments. They experienced inquiry from the students' perspective, but also from the teachers' point of view, practising various ways of assessing the inquiry process. The philosophy, framework and specific programs of the training were elaborated and published earlier (Bernard, Maciejowska, Krzeczowska, & Odrowąż, 2015; Jönsson, Lundström, Finlayson, McLaughlin, & McCabe, 2014; Orwat, Bernard, & Dudek, 2016; Bernard, Dudek, & Orwat, 2019).

Participants

The research group consisted of Polish science teachers ($n = 106$), who participated in the IBSE training in 2 years: 2014 ($n = 60$) and 2015 ($n = 46$). The group included 93 females and 13 males, average age 44.6 years old. The teachers participated in the training voluntarily. The training was free of charge, and the teachers had full participation costs paid, together with accommodation and board. Admission to the training programme was decided on a first-come, first-served basis. Before the registration date, information about the recruitment, dates and participation terms was publicly available for a month. Participation in the training did not translate to participation in a study automatically; the latter was voluntary and anonymous. Every participant was aware of the data to be collected, the goal of their collection, and the mode of their processing, according to the Jagiellonian University ethical standards. The participants were able to renounce their participation in the study at any stage. During the practical part of the training, the teachers were divided into groups according to the declared subject taught at school, as workshops and laboratory classes were adapted to the national curricula of the individual subjects.

Most of the training participants finished uniform master's degree studies or the combination of BA/BSc and supplementary MA/MSc studies in chosen area of science, but some of them finished postgraduate science/science teaching studies only, or they finished related studies (e.g. food technology) and then undertook preparatory courses to teach a given subject. In order to unify the groups, it was assumed that the analysis would only include the tests completed by the teachers with a master's degree in the subject they taught at school. Moreover, incomplete questionnaires were excluded from the analysis. These facts had a significant impact on the size of the studied groups and eventually, it allowed for the comparison of the performance of 71 teachers; the details are presented in Table 2.

Table 2: Description of the research group.

Group	Number of teachers			
	Biology	Physics	Chemistry	All subjects
2014	12	10	10	32
2015	11	12	16	39
Total	23	22	26	71

Instrument and procedures

The teachers completed Lawson tests (2000a) according to the subject they taught before – pre-test (test 1), and after – post-test (test 2) the training. The form of the test and arrangement of the questions were presented in the original form, but the content was translated into Polish. The test was preceded by a questionnaire concerning the respondents' personal information, such as age, gender, type of university they graduated from, and their field of specialization. The test was anonymous, but to enable the comparison of the results achieved by the teachers before and after the workshop, each participant was asked to sign both parts of the test with the same nickname (Dudek, Bernard, & Migdał-Mikuli, 2014).

Teachers' tests were coded into a spreadsheet using the notation: 0 – wrong answer, 1 – correct answer. Then, the percentage values of correct answers were calculated for each teacher concerning the entire test, and separately for questions common for all subjects (Q 1–6), and for subject-specific questions. Further analysis was run in Statistica 13.3 software. The main method used was the analysis of variance (ANOVA) (Howell, 2002). Every analysis of variance was preceded by an examination of the necessary conditions of its application (King, Rosopa, & Minium, 2011). This analysis determined whether the studied samples were characterized by a normal distribution (to this end, a Shapiro-Wilk test was used (Shapiro & Wilk, 1965)) and whether the variances in the tested groups were homogeneous (this parameter was checked using the Levene's test (Levene, 1960)). In cases when the necessary conditions were not met, an analysis with a nonparametric test was carried out, using H test (Kruskal & Wallis, 1952). For *post hoc* tests, the HSD Tukey's test was applied. The significance level was defined as $\alpha = .05$ for this research.

Results

The first part of the analysis was the teachers' RS before the training, and comparison of various subject groups. The results are presented in Table 3 and Figure 1.

Table 3: Pre-test results for entire test (all questions) divided by subject.

Teachers' group	N	Average score [%]	Standard deviation	Confidence interval	
				–95%	95%
Biology	23	69	11	64	74
Physics	22	81	11	76	87
Chemistry	26	70	15	65	75
All together	71	73	14	70	76

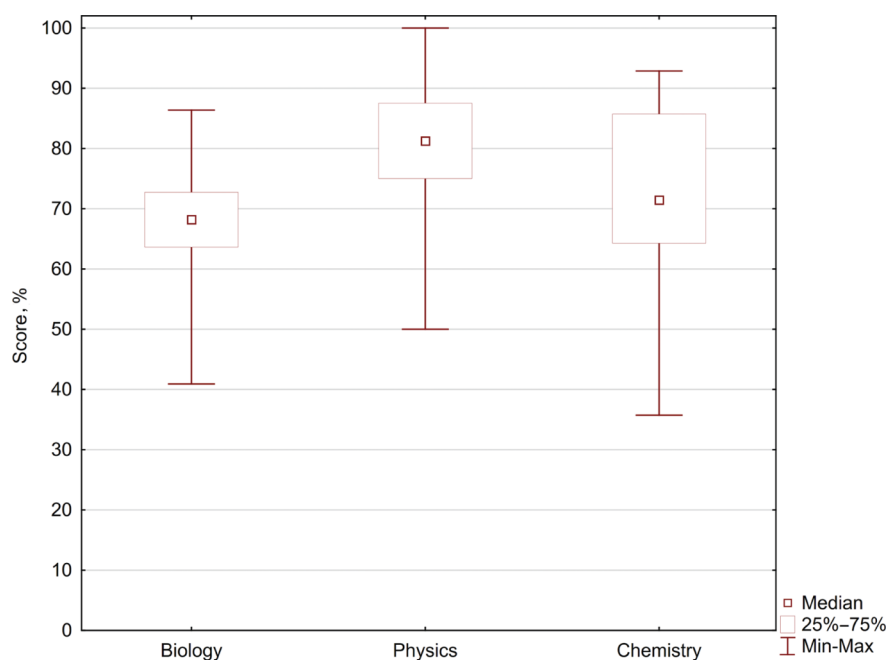


Figure 1: Results of the pre-test for the entire test (all questions) divided by subject.

For the entire test, the distribution of results was described with a normal distribution, and variances in the tested groups were homogeneous, so ANOVA was used. The ANOVA test returned a statistical significance of $p = .0239$, so it could be assumed that there was a statistically significant difference between groups. The plot in Figure 1 suggests that the result for biology teachers was significantly different than that of physics teachers, which was confirmed by the Tukey's *post hoc* test (see Table 4 for its results). Moreover, the *post hoc* results indicated that the score of chemistry teachers was also statistically different, comparing to physics teachers.

Table 4: Tukey's HSD *post hoc* test results for the entire test divided by subject. Bold values indicate statistical significance with a 95% level of confidence.

	Biology	Physics
Biology		
Physics	.0046	
Chemistry	.9079	.0146

The first part of the analysis provided information that the level of RS among physics teachers was significantly higher than in the chemistry and biology teachers' groups. The following tests were to check whether the difference was caused by the score in the common questions (based on general science knowledge), or by the score in subject-specific part of the test, or by both. The results for questions 1–6 are gathered in Table 5 and presented in Figure 2.

Table 5: Pre-test results for common questions (Q 1–6) divided by subject.

Teachers' group	N	Average score [%]	Standard deviation	Confidence interval	
				–95%	95%
Biology	23	67	22	58	77
Physics	22	77	22	68	87
Chemistry	26	69	30	56	81
All together	71	71	26	65	77

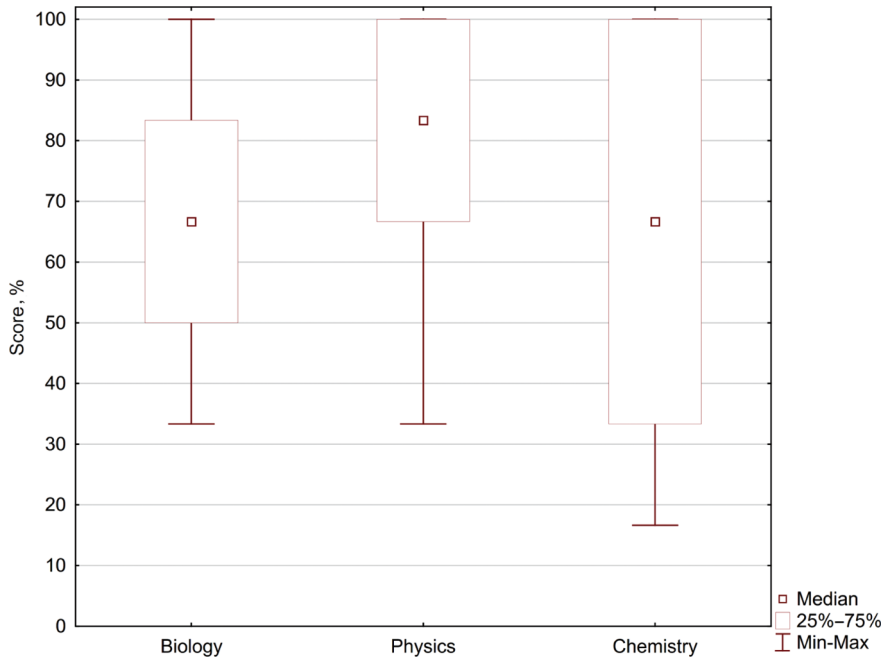
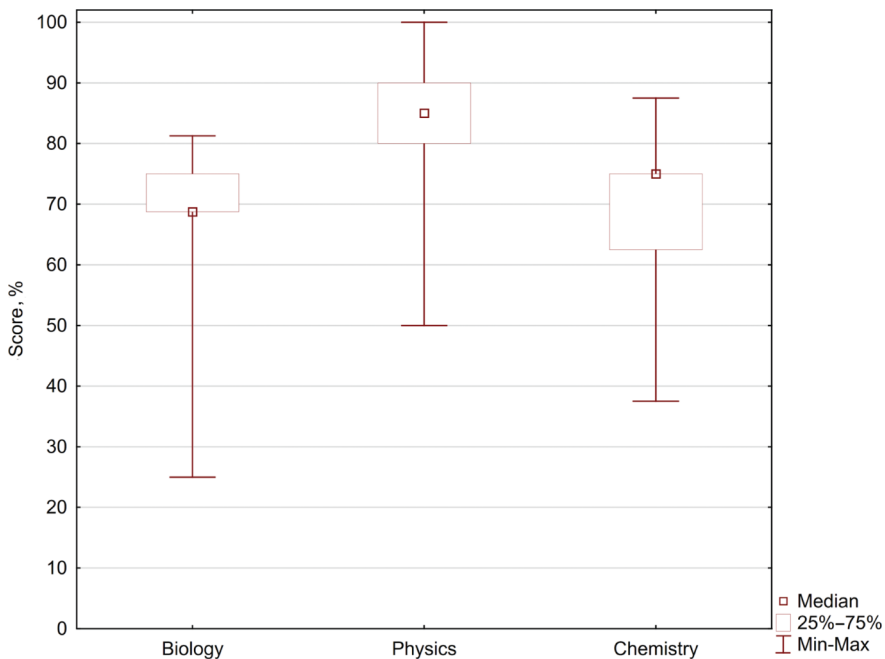


Figure 2: Results of the pre-test for common questions (Q 1–6) divided by subject.

In this case, the results were not normally distributed, so the nonparametric test was used. The Kruskal-Wallis’ test returned $p = .3675$, so it could be assumed that the differences between the results noticeable in Figure 2 were not statistically significant. Therefore, the difference in the results of the entire test should be caused by the difference in the subject-specific questions. The results of this analysis are presented in Table 6 and Figure 3.

Table 6: Pre-test results for subject specific questions divided by subject.

Teachers’ group	N	Average score [%]	Standard deviation	Confidence interval	
				–95%	95%
Biology	23	69	12	64	74
Physics	22	84	12	78	89
Chemistry	26	72	13	67	77
All together	71	75	14	71	78



Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

Figure 3: Results of the pre-test for subject-specific questions divided by subject.

Also, in this case, the condition of the results' normal distribution was not met for all groups, so the nonparametric test was used. The Kruskal-Wallis's test returned a statistical significance of $p = .0001$, so the hypothesis about the difference between groups could be accepted. The *post hoc* tests results are gathered in Table 7; they confirmed the significant difference between the physics group and the chemistry and biology groups.

Table 7: Tukey's HSD *post hoc* test results for subject-specific questions by subject. Bold values indicate statistical significance with a 95% level of confidence.

	Biology	Physics
Biology		
Physics	.0002	
Chemistry	1.0000	.0010

The next part of the analysis was to check if there was an influence of the training in the IBSE on the teachers' RS. Consequently, the analysis was run three times: for the entire test, and for common and subject-specific questions' parts separately. In all those cases, the condition of normal distribution of differences was met so ANOVA test for repeated measurements was used. The results for the entire test are presented in Figure 4.

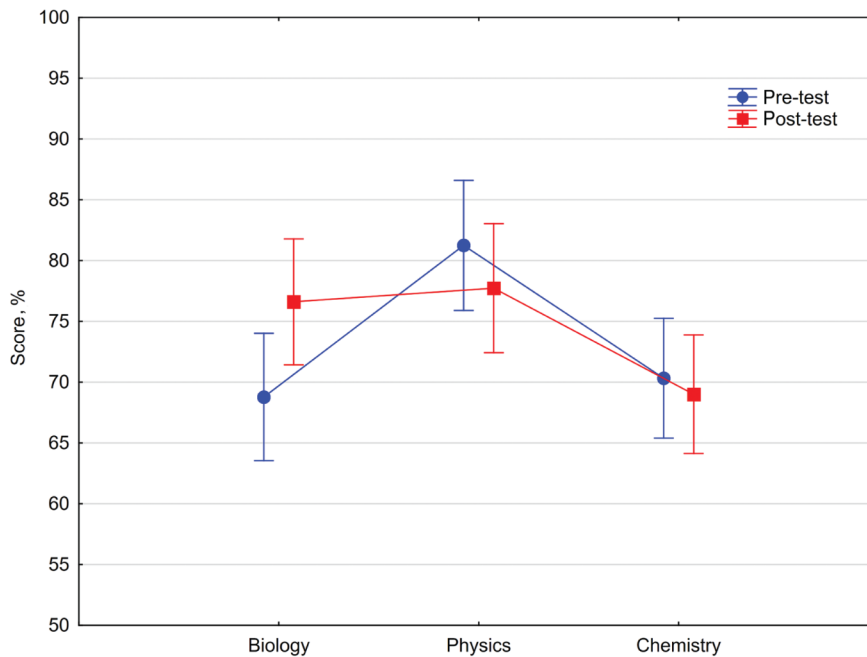


Figure 4: Average scores with 95% confidence intervals of pre- and post-tests for the entire test divided by subject.

The ANOVA test suggested statistically significant differences between groups ($p = .0138$), but the *post hoc* tests clarified that the differences occurred between various subject groups, but not between the pre- and post-test scores within the subject groups. However, analyzing the plots in Figure 4, one could notice that for the chemistry teachers, the pre- and post-test scores were almost equal, while for the physics teachers, there was a small decrease in results in post-test results, and the largest change (increase) is visible for the biology teachers. Next analysis covered only questions common for all subjects (Q 1–6), results presented in Figure 5.

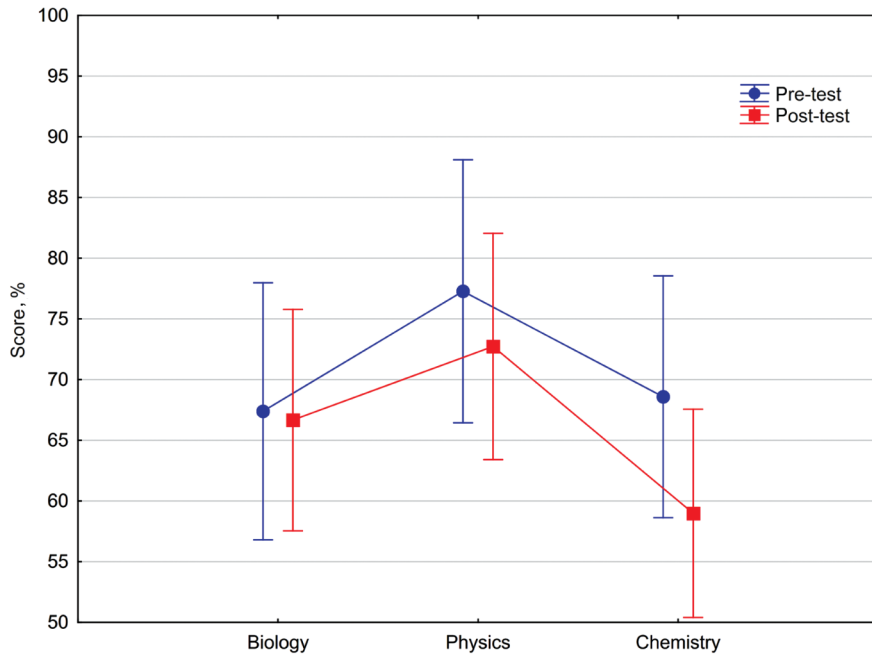


Figure 5: Average scores with 95% confidence intervals of pre- and post-tests for common questions (Q 1–6) divided by subject.

It can be noticed that in this group of questions, the results for all subject groups were lower after the training, but the differences were very small, and the analysis of variance suggested that they were statistically not significant ($p = .5247$). These results meant that not only the differences between the pre- and post-test results were statistically not significant, but also the differences between the scores of various groups after training remain statistically similar. The results for subject-specific questions are presented in Figure 6.

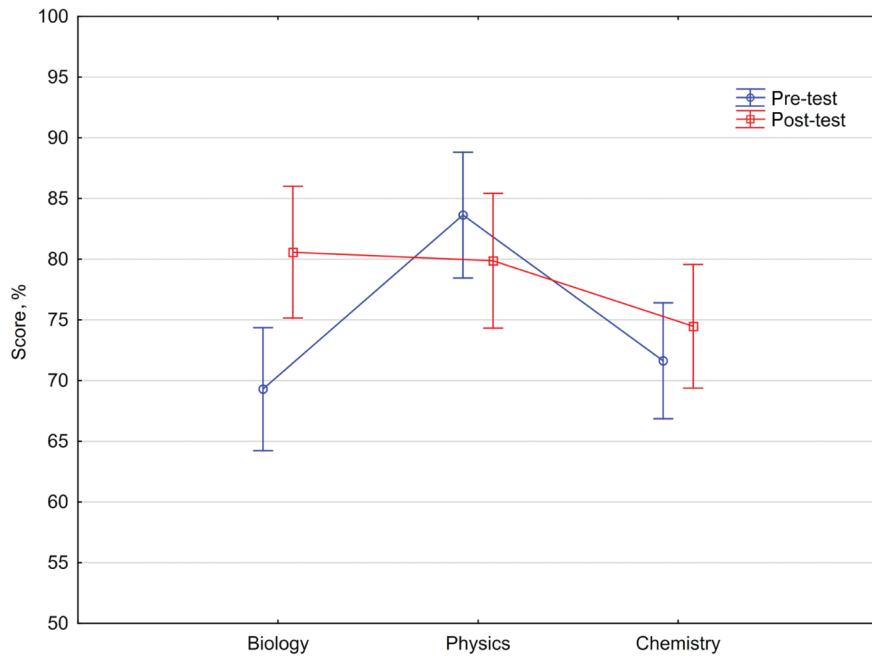


Figure 6: Average scores with 95% confidence intervals of pre- and post-tests for subject-specific questions divided by subject.

In this case, the increase in the post-test results was visible for biology and chemistry groups, however, the scale of the effect is clearly larger for biologists. Physics teachers' group score was lower after training, and the scale of the effect was similar to the change for the chemistry group. The ANOVA results confirm that there were statistically significant differences between groups, and what's important, the results based on the Tukey's

test presented in Table 8 indicated that there was a statistically significant increase in the result of biologists' group after the training.

Table 8: Tukey's HSD *post hoc* test results pre- and post-tests for subject-specific questions divided by subject. Bold values indicate statistical significance with a 95% level of confidence.

		Biology Pre-test	Biology Post-test	Physics Pre-test	Physics Post-test	Chemistry Pre-test
Biology	Post-test	.0132				
Physics	Pre-test	.0022	.9668			
Physics	Post-test	.0602	1.0000	.8732		
Chemistry	Pre-test	.9888	.1539	.0198	.2534	
Chemistry	Post-test	.7307	.5704	.1526	.7153	.9414

Discussion

The average result of 73% (see Table 3) for all the respondents was not low, but considering the fact that the tests are suitable for upper secondary school students and all study participants completed a master's degree and were authorized to teach, it was not a very good one either. Results at that level were reachable for university and upper secondary school students (Coletta & Phillips, 2005; Maloney, 1981). It was generally expected that during the studies in the science fields, the skills should become more developed, and the persons working as teachers, responsible for the development of those skills in students, should demonstrate a very high level of such skills. However, the obtained result was not surprising in light of progression trend of scientific reasoning from elementary school to university described by Ding (2018) which showed that the RS are mainly developed in grades 8–12, so at the time when students were taking general science courses and at the academic level, they remained constant.

Considering the distribution of results according to the subject taught (Table 3, Figure 1), it could be observed that reasoning skills within a group of physics teachers were better developed, comparing to chemistry and biology teachers. In the results for the common questions (Table 5 and Figure 2), the higher score of physics teachers was also noticeable, but the difference was lower and not statistically significant. It should be also noticed that the scores for those questions covered a wide range of values, and the distribution of results could not be considered as normal, which affected the high value of standard deviation. Analysis of results for the subject-specific questions (Table 6, Figure 3) showed a smaller range of results, but normal distribution was still not observed. For those questions, physics teachers' scores were once again higher than those of the chemistry and biology groups, and in this case, the difference was statistically significant. Thinking stereotypically, the highest scores of physics teachers are no surprise. Physics, as an experiment-based subject, heavily dependent on mathematical skills, required its teacher's to have highly developed reasoning skills. Using the same stereotype, one would expect that the results of chemistry teachers should be significantly higher than those of biology teachers, but the study outcomes showed no significant differences between the chemistry and biology groups.

The obtained results suggest that training in the IBSE methodology and exercises in formulating research questions, hypotheses, arguments, conclusions etc. may affect the teachers' level of reasoning skills. The highest increase (statistically significant) was observed for biology teachers, the group with the lowest RS level before the training (Figure 4). The change resulted from an increase in score for the subject-specific questions (Figure 6, Table 8), while none of the groups improved their skills in the area of common questions (Figure 5). The increase in the subject-specific questions area was observed also for chemistry teachers, but in this case, the change was smaller and statistically not significant. Finally, it should be noted that after training differences between scores of biology, chemistry and physics teachers became statistically not significant.

Conclusion

The obtained results indicate that the level of RS among science teachers in Poland is not higher than that of upper secondary (class 11–12) and university students. This fact may have a significant influence on their teaching efficiency. The teachers need to be confident in solving complex problems with students, as well as assisting them and assessing tests based on the RS level verification. Therefore, it is necessary to introduce

courses which enable raising the level of RS among teachers. Unfortunately, trainings in the IBSE, which were very popular within the FP7 programme and are still quite common in many countries, have a limited effect on the level of the teachers' reasoning skills. In the described study, the increase was noticeable only for the biology teachers – the group with the lowest RS level before the training, and after the training, their results equalized with other subject groups. It can be concluded that such training gives a positive effect only for specific groups, with significantly lower RS level. Therefore, the training in inquiry does not solve the problem of the generally low level of science teachers' reasoning skills, and helping them to reach a higher RS level requires a longer and dedicated training.

Acknowledgements

The research was based on the teacher training within the SAILS FP7 project (Strategies for Assessment of Inquiry Learning in Science). The SAILS project has received funding from the European Commission Seventh Framework Programme [FP7/2007-2013] under grant agreement number 289085 and support from the Polish Ministry of Science and Higher Education [2887/7PR/2013/2], funds for science in the years 2013–2015 for co-financed international projects.

References

- Abdellatif, H. R., Cummings, R., & Maddux, C. D. (2008). Factors affecting the development of analogical reasoning in young children: A review of literature. *Education*, 129(2), 239–249.
- Adey, P. (1999). Thinking Science: Science as a gateway to general thinking ability. In: J. M. Hamers, J. E. Van Luit, & B. Csapó (Eds.), *Teaching and learning thinking skills* (pp. 63–80). Lisse: Swets and Zeitlinger.
- Amsel, E., Klaczynski, P. A., Johnston, A., Bench, S., Close, J., Sadler, E., & Walker, R. (2008). A dual-process account of the development of scientific reasoning: The nature and development of metacognitive intercession skills. *Cognitive Development*, 23(4), 452–471.
- Bao, L., Xiao, Y., Koenig, K., & Han, J. (2018). Validity evaluation of the Lawson classroom test of scientific reasoning. *Physical Review Physics Education Research*, 14(2), 020106, 1–19. DOI: 10.1103/PhysRevPhysEducRes.14.020106.
- Bernard, P., Dudek, K., & Orwat, K. (2019). Integration of inquiry-based instruction with formative assessment: The case of experienced chemistry teachers. *Journal of Baltic Science Education*, 18(2), 184–196. DOI: 10.33225/jbse/19.18.184.
- Bernard, P., Maciejowska, I., Krzeczowska, M., & Odrowąż, E. (2015). Influence of in-service teacher training on their opinions about IBSE. *Procedia – Social and Behavioral Sciences*, 177, 88–99. 10.1016/j.sbspro.2015.02.343.
- Bernard, P., Maciejowska, I., Odrowąż, E., Dudek, K., & Geoghegan, R. (2012). Introduction of inquiry based science education into Polish science curriculum – general findings of teachers' attitude. *Chemistry-Didactics-Ecology-Metrology*, 17(1–2), 49–59. DOI: 10.2478/cdem-2013-0004.
- Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment*, 15(3–4), 142–174.
- Caroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. Cambridge: Cambridge University Press.
- Centurino, V., & Jones, L. R. (2017). *Chapter 2. TIMSS 2019 Science Framework*. Chesnut Hill, MA: TIMSS, & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA). Retrieved from <http://timss2019.org/wp-content/uploads/frameworks/T19-Assessment-Frameworks-Chapter-2.pdf>.
- Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3). Retrieved from www.amstat.org/publications/jse/v10n3/chance.html.
- Coletta, V. P., & Phillips, J. A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics*, 73(1172), 1–11. DOI: 10.1119/1.2117109.
- Csapó, B. (1992). Improving operational abilities in children. In: A. Demetriou, M. Shayer, & A. Efklides (Eds.), *Neo-Piagetian theories of cognitive development. Implications and applications for education* (pp. 144–159). London: Routledge and Kegan.
- Csapó, B. (1997). Development of inductive reasoning: Cross-sectional measurements in an educational context. *International Journal of Behavioral Development*, 20(4), 609–626.
- Csapó, B., Csíkos, C., Korom, E., Németh, M. B., Black, P., Harrison, C., van Kampen, P., & Finlayson, O. (2012). *Report on the strategy for the assessment of skills and competencies suitable for IBSE. Strategies for the Assessment of Inquiry Learning in Science (SAILS FP7 project)*. Retrieved from <http://sails-project.eu/sites/default/files/outcomes/d2-1.pdf>.
- Ding, L. (2018). Progression Trend of Scientific Reasoning from Elementary School to University: a Large-Scale Cross-Grade Survey among Chinese Students. *International Journal of Science and Mathematics Education*, 16(8), 1479–1498. DOI: 10.1007/s10763-017-9844-0.
- Drummond, C., & Fischhoff, B. (2015). Development and validation of the scientific reasoning scale. *Journal of Behavioral Decision Making*, 30(1), 26–38.
- Dudek, K., Bernard, P., & Migdał-Mikuli, A. (2014). Reasoning skills of Polish science teachers. In: *Science and Technology Education for the 21st Century. Research and Research Oriented Studies. Proceedings of the 9th IOSTE Symposium for Central and Eastern Europe* (pp. 78–90). Hradec Králové: Gaudeamus.

- Dudek, K., Bernard, P., & Odrowąż, E. (2015). First steps in Assessment of students' inquiry: A case study of non-experienced chemistry teacher. In: *State-of-the-art and future perspectives. Proceedings of the 1st International Baltic Symposium on Science and Technology Education* (pp. 42–44). Sialuliai: Scientia Educologica.
- Dylak, S. (Ed.). (2011). *Strategia kształcenia wyprzedzającego*. Poznań: Ogólnopolska Fundacja Edukacji Komputerowej. Retrieved from https://edustore.eu/download/Strategia_Kształcenia_Wyprzedzajacego.pdf.
- Frensch, P. A., & Funke, J. (Eds.). (1985). *Complex problem solving: The European perspective*. Hillsdale, NJ: Lawrence Erlbaum.
- Giroto, V., & Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition*, 106(1), 325–344.
- Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE – Life Sciences Education*, 11(4), 364–377.
- Hamers, J. H. M., de Koning, E., & Sijtsma, K. (1998). Inductive reasoning in third grade: Intervention promises and constraints. *Contemporary Educational Psychology*, 23(2), 132–148.
- Howell, D. (2002). *Statistical methods for psychology* (5th ed.). Duxbury: Thomson Learning.
- Howson, C., & Urbach, P. (1996). *Scientific reasoning. The Bayesian approach. 2nd ed.* Chicago: Open Court Publishing Company.
- Johnson-Laird, P. N. (2006). *How we reason*. Oxford: Oxford University Press.
- Jones, G. A., Langrall, C. W., Thornton, C. A., & Mogill, A. T. (1997). A framework for assessing and nurturing young children's thinking in probability. *Educational Studies in Mathematics*, 32, 101–125.
- Jönsson, A., Lundström, M., Finlayson, O., McLaughlin, E., & McCabe, D. (2014). *Report on IBSE Teacher Education and Assessment programme – Stage 1. Strategies for the Assessment of Inquiry Learning in Science (SAILS FP7 project)*. Retrieved from <http://sails-project.eu/sites/default/files/outcomes/d4-2.pdf>.
- Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching*, 50(5), 530–560. DOI: 10.1002/tea.21086.
- King, B. M., Rosopa, P. J., & Minium, E. W. (2011). *Statistical reasoning in the behavioral sciences* (6th ed.). Danvers: John Wiley, & Sons, Inc.
- Kishta, M. A. (1979). Proportional and combinatorial reasoning in two cultures. *Journal of Research in Science Teaching*, 16(5), 439–443.
- Klauer, K. J. (1989). Teaching for analogical transfer as a means of improving problem solving, thinking, and learning. *Instructional Science*, 18, 179–192.
- Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology*, 64(3), 141–152.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621. DOI: 10.1080/01621459.1952.10483441.
- Kuhn, D., Pease, M., & Wirkala, C. (2009). Coordinating the effects of multiple variables: A skill fundamental to scientific thinking. *Journal of Experimental Child Psychology*, 103(3), 268–284.
- Kuhn, D., Phelps, E., & Walters, J. (1985). Correlational reasoning in an everyday context. *Journal of Applied Developmental Psychology*, 6(1), 85–97.
- Lawson, A. E., Adi, H., & Karplus, R. (1979). Development of correlational reasoning in secondary schools: Do biology courses make a difference? *American Biology Teacher*, 41(7), 420–425.
- Lawson, E. A. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1), 11–24.
- Lawson, E. A. (2000a). *Classroom test of scientific reasoning*. Arizona: Test revision, Arizona State University. Retrieved from <http://www.public.asu.edu/~anton1/LawsonAssessments.htm>
- Lawson, E. A. (2000b). The generality of hypothetico – deductive reasoning: making scientific thinking explicit. *The American Biology Teacher*, 62(7), 482–495.
- Lawson, E. A. (2003). The nature and development of hypothetico – predictive argumentation with implications for science teaching. *International Journal of Science Education*, 25(11), 1387–1408.
- Lawson, E. A. (2005). What is the role of induction and deduction in reasoning and scientific inquiry? *Journal of Research in Science Teaching*, 42(6), 716–740.
- Levene, H. (1960). Robust tests for equality of variances. In: *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278–292). Palo Alto: Stanford University Press.
- Lockwood, E. (2013). A model of students' combinatorial thinking. *Journal of Mathematical Behavior*, 32(2), 251–265.
- Lovelace, M., & Brickman, P. (2013). Best practices for measuring students' attitudes toward learning science. *CBE – Life Sciences Education*, 12, 606–617.
- Maloney, D. P. (1981). Comparative reasoning abilities of college students. *American Journal of Physics*, 49, 784–785. DOI: 10.1119/1.12676.
- Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 Assessment frameworks*. Chesnut Hill, MA: TIMSS, & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
- OECD (2003a). *Assessment framework. Mathematics, reading, science and problem solving*. Paris: OECD.
- OECD (2003b). *Learners for life. Student approaches to learning. Results from PISA 2000*. Paris: OECD.
- OECD (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris: OECD.
- OECD (2009). *Assessment Framework. Key competencies in reading, mathematics and science*. Paris: OECD.
- OECD (2012). *Draft PISA 2015 Collaborative problem solving assessment framework*. Paris: OECD.
- OECD (2013). *PISA 2012 assessment and analytical framework. Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD.
- Orwat, K., Bernard, P., & Dudek, K. (2016). Starego drzewa nie zegniesz, czyli jak szkolić nauczycieli z zakresu stosowania metod samodzielnego dociekania wiedzy przez uczniów (IBSE) [Old trees do not bend, how to train teachers in the application of inquiry-based methods (IBSE)]. *Aktualne problemy dydaktyki przedmiotów przyrodniczych* (pp. 165–179). Kraków: Wydzał Chemii UJ.
- Polya, G. (1968). *Induction and analogy in mathematics. Volume 1: Mathematics and plausible reasoning*. Princeton: Princeton University Press.
- Pratt, C., & Hacker, G. R. (1984). Is Lawson's classroom test of formal reasoning valid? *Educational and Psychological Measurement*, 44(2), 441–448. DOI: 10.1177/0013164484442025.

- Pyper, A. B. (2011). Changing scientific reasoning and conceptual understanding in college students. *AIP Conference Proceedings*, 1413, 63–65. DOI: 10.1063/1.3679994.
- Rocard, M., Csermely, P., Jorde, D., Lenzen, D., Walberg-Henriksson, H., & Hemmo, V. (2007). *Science education now: A renewed pedagogy for the future of Europe*. Brussels: European Communities.
- Ross, J. A., & Cousins, B. (1993). Enhancing secondary school students' acquisition of correlational reasoning skills. *Research in Science, & Technological Education*, 11(2), 191–205.
- Schröder, E., Bödeker, K., Edelstein, W., & Teo, T. (2000). Proportional, combinatorial, and correlational reasoning. A manual including measurement procedures and descriptive analyses. *Study "Individual Development and Social Structure"*. *Data Handbooks Part 4*. Berlin: Max Planck Institute for Human Development.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4), 591–611. DOI: 10.1093/biomet/52.3-4.591.
- Stefanich, G. P., Unruh, R. D., Perry, B., & Phillips, G. (1983). Convergent validity of group tests of cognitive development. *Journal of Research in Science Teaching*, 20(6), 557–563. DOI: 10.1002/tea.3660200607.
- Utomo, A. P., Narulita, E., & Shimizu, K. (2018). Diversification of reasoning science test items of TIMSS grade 8 based on higher order thinking skills: A case study of Indonesian students. *Journal of Baltic Science Education*, 17(1), 152–161.
- Venville, G., Adey, P., Larkin, S., & Robertson, A. (2003). Fostering thinking through science in the early years of schooling. *International Journal of Science Education*, 25(11), 1313–1332.
- Watters, J. J., & English, L. D. (1995). Children's application of simultaneous and successive processing in inductive and deductive reasoning problems: Implications for developing scientific reasoning skills. *Journal of Research in Science Teaching*, 32(7), 699–714.
- Williams, R. L. (1999). Operational definitions and assessment of higher-order cognitive constructs. *Educational Psychology Review*, 11(4), 411–427.
- Zhou, S., Han, J., Koenig, K., Raplinger, A., Pi, Y., Li, D., Xiao, H., Fu, Z., & Bao, L. (2016). Assessment of scientific reasoning: The effects of task context, data, and design on student reasoning in control of variables. *Thinking Skills and Creativity*, 19, 175–187. DOI: 10.1016/j.tsc.2015.11.004.