FORMATIVE ASSESSMENT AS A TOOL TO ENHANCE THE DEVELOPMENT OF INQUIRY SKILLS IN SCIENCE EDUCATION

Mária Ganajová, Ivana Sotáková, Stanislav Lukáč, Zuzana Ješková, Veronika Jurková, Renáta Orosová

Introduction

Assessment is the key process in assuring quality education (Lamanauskas & Vilkonienė, 2008). The problems of assessment – and formative assessment (FA), in particular – have been widely addressed by teachers, teachers’ educators, and researchers for several decades. Formative assessment has a variety of definitions. However, different researchers have come to a consensus regarding certain characteristics pertaining to this form of assessment. Formative assessment is based on regular and interactive evaluation of students’ work, it provides feedback on their learning in terms of the specified goals, and indicates what the next steps in teaching should be (OECD/CERI, 2008; OECD, 2005; Wiliam, 2011).

The efficacy of FA has been investigated by a number of researcher teams focused on education. Many of them have confirmed a positive influence on the students’ learning process such as an improvement in their academic results, i.e., FA helps improve students’ performance and the overall quality of education (Allal & Mottier Lopez, 2005; Bell & Cowie, 2001; Black & Wiliam, 1998a, 1998b; Brookhart, 2008; Floréz & Sammons, 2013; Fluckiger et al., 2010; Hattie & Timperley, 2007; OECD/CERI, 2008; Wiliam, 2010). Implementing FA enhances understanding of the educational contents (Schunk & Swartz, 1993). Black and Wiliam (1998a) state that lower-achieving students benefit from it the most. The efficacy of FA has also been emphasized by Brookhart (2001), Choi et al. (2001), Condie et al. (2005), Department for Education and Skills (2007), Furtak and Ruiz-Primo (2008), Kellard et al. (2008), Kingston and Nash (2011), Kirton et al. (2007), Office for Standards in Education, Children’s Services and Skills (2008), Ozan and Kincal (2018), Taras (2007), Torrance and Pryor (2001), Webb and Jones (2009) and many others.

FA develops students’ competences, mainly their ability to learn, i.e., students are able to learn from their mistakes, evaluate the results of their learning process, and propose how to improve their own performance (Heritage, 2007, 2010; Looney, 2011; Marshal, 2011; Myhill & Warren, 2005;
OECD, 2005). FA is also important in the process of developing social and personal competences (Starý et al., 2016).

There are studies that have not statistically confirmed any significant influence of FA on students' performance or academic achievement (e.g., Andrews, 2011; Collins, 2012; King, 2003; Tuominen, 2008; Yin et al., 2008). Bennett (2011) has pointed out there is little experimental and quantitative research verifying the impact of FA on students’ learning results, and the specialised literature addressing FA does not focus on the provability and interpretation of the research results. A similarly critical view of the research focused on FA efficacy can be found in Briggs et al. (2012), Dunn and Melvenon (2009), Filesecker and Kerres (2012), and McMillan et al. (2013). However, this kind of conclusion tends to occur when there are difficulties with effective implementation of FA.

Of course, implementing FA is accompanied by certain difficulties. Researchers have pointed out that the influence of FA on students’ learning results is difficult to prove because it only becomes visible in the long run (e.g., Starý, 2006), this kind of research is time-consuming (e.g., Cizek, 2010; Florez & Sammons, 2013; Stiggins, 2002), the requirements for educational contents are complex and determined by the curriculum – the amount of knowledge students are supposed to learn keeps increasing (e.g., Starý, 2006; Stiggins, 2002), the number of students in the classroom is high, teachers are not trained to use FA, and there is little access to mentoring (Clark, 2010; Novotná & Krabsová, 2013).

In the Slovak educational system, summative assessment prevails. The OECD evaluation team has brought attention to the need to improve students' understanding by using formative assessment (Shewbridge et al., 2014). This situation has also been addressed by Orosová et al. (2019) and Ganajová et al. (2019). Their research has shown that Slovak teachers use summative assessment more frequently than formative assessment as some of them are not familiar with formative assessment and they do not understand its methods or aim. Their ideas of certain FA methods such as self- or peer-assessment are often incorrect.

Formative Assessment Classroom Techniques

Formative assessment is based on “purposeful, planned and often spontaneous teacher-to-student, student-to-teacher and student-to-student verbal and written interactions” (Keeley, 2008). These interactions include a variety of assessment techniques known as “Classroom Assessment Techniques” (CATs) (Angelo & Cross, 1993) or “Formative Assessment Classroom Techniques” (FACTs) (Keeley, 2008).

Using FACTs is not time consuming, yet it can provide the teacher with a good overview and feedback on the lessons and at the same time, students get feedback on their own learning (Keeley, 2008). For example, FACTs can take the form of true or false statements (Hubbard et al., 2017; Keeley, 2008; Marzano, 2010), K-W-L chart (Bailey, 2017), Frayer Model (Buehl, 2001; Frayer et al., 1969; Keeley, 2013), self-assessment card (Keeley & Tobey, 2011), card mapping the learning process (Szarka, 2017), metacognition (Flavell, 1979; Livingston, 1997), exit card (He, 2019; Keeley, 2013), concept map (Pendley et al., 1994; Regis et al., 1996), check list (British Columbia Institute of Technology, 2010; Ma et al., 2012), task cards (Hattie, 2009), and others. These FACTs help students engage in deeper thinking about their ideas in science and identify the progress in the development of their scientific understanding.

When assessing students’ performance in science education, students’ understanding of core concepts cannot be assessed separately from their abilities to use the practices of science. These two dimensions of learning should be assessed together to verify whether students can apply theoretical scientific concepts in practice – in other words, whether their level of understanding allows them to investigate the world through practices of scientific inquiry. The term “practices” indicates that engaging in scientific inquiry requires not only knowledge, but also practical skills (National Research Council, 2012, p. 30). These skills are often referred to as inquiry skills. The FACTs can be used to train these practices as suggested by Keeley (2015).

Research Problem

In the Slovak State Educational Programme for upper secondary education (ŠPU, 2014), specifically the “Man and nature” educational area, it has been stated that teaching science, i.e., physics, chemistry, and biology is supposed to develop scientific literacy in students who are subsequently able to apply the gained knowledge, formulate questions, and draw conclusions based on the evidence in order to understand the subject matter. Discovery and inquiry are basic approaches that allow students to acquire new knowledge and learn the basic scientific skills. These approaches also help students develop a positive attitude to the scientific approach in learn-
FORMATIVE ASSESSMENT AS A TOOL TO ENHANCE THE DEVELOPMENT OF INQUIRY SKILLS IN SCIENCE EDUCATION (pp. 204-222)

The main research aim was to determine the efficacy of teaching using FACTs with regard to inquiry skill development while using the currently prevalent teacher-centred teaching style in science education.

The research questions were defined as follows:
1. What is the level of the selected inquiry skills in the sample of students at the beginning of the research?
2. Is teaching using FACTs effective in the development of the selected inquiry skills?
3. How do different factors (gender and overall assessment at the end of the term) affect the efficacy of teaching using FACTs in terms of developing the selected inquiry skills?

Research Hypotheses

The following hypotheses were formulated and tested:

\( H_1: \) After FACTs are implemented into teaching, there is a significant difference in the overall level of inquiry skills in the experimental group in comparison to the control group.

\( H_2: \) After FACTs are implemented into teaching, there is a significant difference in the development of the individual inquiry skills in the experimental group in comparison to the control group.

\( H_3: \) In the experimental group, there is a significant difference in the inquiry skill development in students divided according to their gender.
H$_4$: In the experimental group, there is a significant difference in the inquiry skill development in students divided according to their overall assessment at the end of the term.

**Research Methodology**

**Research Design**

The pre-test and post-test two-group design was used in the research (Campbell & Stanley, 1963). In both experimental and control groups, science subjects (biology, chemistry, physics, mathematics, and informatics) were taught by the same teachers. Teaching in these groups did not differ in terms of content, methods, or duration. However, FACTs were used systematically to teach the experimental group. The number of lessons dedicated to each of these subjects in the students' weekly schedule, lesson contents and time available were identical in both groups. Figure 1 presents the stage of research design.

**Figure 1**

Stages of Research Design

Deliberate sampling was used to select the research sample. Three Slovak upper secondary schools were involved in the research conducted during the second term of the 2018/2019 school year. The selection of schools was deliberate and followed the following criteria. Firstly, schools inclined towards innovation whose management promoted active learning were selected. Secondly, schools whose teachers actively participated in the IT Academy – Education for the 21st Century project (http://itakademia.sk/) and expressed their interest to implement FA into their teaching were selected. In terms of this project, the teachers had an opportunity to get acquainted with a variety of FACTs. Based on these criteria, 11 teachers were selected for the project: 3 men and 8 women with 2 to 24 years of teaching practice.

At each school, teachers selected two 1st year forms. The form with the higher average academic achievement at the end of the term was selected as the control group, while the worse faring form was selected as the experimental group. This division did not affect the pre-test results; there was no statistically significant difference in the level of inquiry skills between the experimental and the control group. The research sample consisted of 121 students from the selected forms who took both tests (pre-test and post-test).

55 (45.5%) were males and 66 (54.5%) females. Students were aged 15–16 years. The experimental group consisted of 73 (60.3%) students, and the control group of 48 (39.7%) students. Table 1 summarises the number and percentage of students divided into groups based on their gender and overall assessment at the end of the term. Each term, the overall assessment of upper secondary school students in the Slovak Republic follows the rules specified in Guideline no. 21/2011 for the assessment and classification of upper secondary school students issued by the Ministry of Education, Science, Research and Sport of the Slovak Republic (MŠVVaŠ SR, 2011).
Table 1
Division of Students according to Gender and Overall Academic Performance at the End of the Term

<table>
<thead>
<tr>
<th>Variable</th>
<th>Experimental group</th>
<th>Control group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>34</td>
<td>46.6</td>
</tr>
<tr>
<td>Female</td>
<td>39</td>
<td>53.4</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>100</td>
</tr>
<tr>
<td>Overall assessment at the end of the term</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passed with distinction</td>
<td>35</td>
<td>48.0</td>
</tr>
<tr>
<td>Passed very well</td>
<td>25</td>
<td>34.2</td>
</tr>
<tr>
<td>Passed</td>
<td>13</td>
<td>17.8</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>100</td>
</tr>
</tbody>
</table>

The students passed with distinction (PD) if their average study results in all compulsory subjects were equal to or less than 1.5.
The students passed very well (PW) if their average study results in all compulsory subjects were equal to or less than 2.0.
The students passed (P) if their average study results in all compulsory subjects were more than 2.0.

Ethical Considerations

An initial meeting was held before the research started. The research team informed the teachers about the focus of the research and their role in it as well as about the contents of the database of FACTs created, and publishing of results. Both teachers and students participated in the research voluntarily. The teachers motivated the students to accept the FACTs positively. The students were informed that the test results concerning their inquiry skills would be statistically processed for research purposes. In terms of result evaluation, each student participant was assigned a code to maintain confidentiality of their personal information.

Research Instrument

A test of inquiry skills was used as the research instrument. This test was updated from the previous version presented and described in Ješková et al. (2016); it was developed based on existing tests analyses, e.g., ScInqLiT (Wenning, 2007), TISP (Burns et al., 1985), TOSLS (Gormally et al., 2012), TGSA (Zion et al., 2005) and inquiry skills taxonomies (Fradd et al., 2001; Tamin & Lunetta, 1981; Van den Berg, 2013). The test consisted of 16 items, it was administered in the paper-and-pencil form, and took 45 minutes. The items focused on the following inquiry skills: formulating hypothesis, designing experiments, transforming data into tables and graphs, determining relationships between variables based on tables and graphs, and identifying possible sources of errors. At least two items were assigned to each skill. The context of the tasks drew from the subject matter addressed by the main science subjects (physics, chemistry, biology). Therefore, the test items related to the concepts in different subjects (Table 2). To understand the relationships described in the tasks, basic primary school knowledge was needed, but the students were required to apply their knowledge and skills in new situations. The tasks were selected deliberately – each of them required application of a single specific inquiry skill.

The majority were multiple-choice items with one or two correct answers. In four items, students were offered a set of variables among which they were supposed to identify the appropriate variables and the relationships between them. Test scores were calculated based on the following criteria. In the four aforementioned tasks, only the maximum of .5 point could be achieved; multiple-choice test items (12) were assigned the values of 0 or 1, therefore students could achieve a maximum of 14 points. Most multiple-choice test items included one correct answer. If the student selected the correct answer only, they received 1 point. In all other cases, even when the correct answer was combined with another option, students got 0 points. Multiple-choice items with two correct answers were assigned 1 and .5 point if the student selected both the two, or just one of the correct answers respectively. In all other cases, 0 points were assigned (Ješková et al., 2016).
The reliability of the selected inquiry skill test was determined using the Cronbach Alpha; $C_\alpha = .63$. Several researchers (e.g., Cohen et al., 2007; Marx et al., 2004) have stated that lower $C_\alpha$ values are acceptable for tests with fewer test items.

Table 2 includes all test characteristics. The last column of Table 2 presents scientific practices supported by FACTs as proposed by Keeley (2015). They correspond with the tested/measured inquiry skills presented in the first column.

<table>
<thead>
<tr>
<th>Tested/measured inquiry skills</th>
<th>Number of test items</th>
<th>Subject</th>
<th>Scientific practices as proposed by Keeley (2015)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formulate a hypothesis to test</td>
<td>2 (2 p)</td>
<td>Chemistry, physics</td>
<td></td>
</tr>
<tr>
<td>Designing experiments (identify variables and their relationships)</td>
<td>6 (4 p)</td>
<td>Physics, informatics, chemistry</td>
<td>Planning and carrying out investigations</td>
</tr>
<tr>
<td>Transform data to standard forms (i.e., tables or graphs)</td>
<td>2 (2 p)</td>
<td>Physics, informatics, mathematics</td>
<td>Analysing and interpreting data</td>
</tr>
<tr>
<td>Determining the relationship between variables (based on tables)</td>
<td>2 (2 p)</td>
<td>Physics, informatics, mathematics</td>
<td>Using mathematics and computational thinking</td>
</tr>
<tr>
<td>Determining the relationship between variables (based on graphs)</td>
<td>2 (2 p)</td>
<td>Physics, mathematics</td>
<td></td>
</tr>
<tr>
<td>Determining the accuracy (identify possible sources of errors)</td>
<td>2 (2 p)</td>
<td>Physics, biology</td>
<td></td>
</tr>
</tbody>
</table>

Procedures

In the preparatory stage of the research (December – October 2018), a group of experts in subject didactics at the Faculty of Science, Pavol Jozef Šafárik University in Košice prepared FACTs for thematic units in the following science subjects – chemistry, biology, physics, mathematics, informatics. In accordance with the respective State Educational Programme for upper secondary education (ISCED 3), these thematic units are taught in the second term of the 1st year at upper secondary schools (ŠPÚ, 2014).

In January 2019, teacher research participants were invited to attend a specialised seminar to learn about the focus and goals of the research. These teachers were given access to the database of FACTs and became conversant with their use in teaching.

At the end of January 2019, all student participants were administered the pre-test to identify their initial level of inquiry skills. Nor teachers, neither students were informed of the pre-test results to avoid influencing the post-test. The pre-test results showed that the experimental group students and the control group students were at a statistically similar level ($p > .05$).

In both control and experimental groups, teaching took place in the same period of the school year, i.e., February to June 2019; In terms of content and teaching methods, the teaching process in both control and experimental groups was comparable. In the control group, the teachers used the teacher-centred teaching style i.e., they did not use FACTs. In the experimental group, teachers used FACTs at their own discretion in different phases of teaching/learning (Table 3). They chose the specific FACTs from the aforementioned database. Their work was coordinated to ensure that at least three subjects were taught using FACTs, and they were used at least once a week in at least one of the subjects.
Table 3
FACT Type Used in the Teaching-Learning Process

<table>
<thead>
<tr>
<th>FACT type</th>
<th>Average number of uses per student</th>
</tr>
</thead>
<tbody>
<tr>
<td>True or false statements</td>
<td>2</td>
</tr>
<tr>
<td>K-W-L chart</td>
<td>.5</td>
</tr>
<tr>
<td>Frayer Model</td>
<td>1</td>
</tr>
<tr>
<td>Self-assessment card</td>
<td>1</td>
</tr>
<tr>
<td>Card mapping the learning process</td>
<td>.5</td>
</tr>
<tr>
<td>Metacognition</td>
<td>1</td>
</tr>
<tr>
<td>Exit card</td>
<td>1.5</td>
</tr>
<tr>
<td>Concept map</td>
<td>1.5</td>
</tr>
<tr>
<td>Check list</td>
<td>1</td>
</tr>
<tr>
<td>Task cards</td>
<td>2</td>
</tr>
</tbody>
</table>

Every FACT addressed the content of the respective lesson.

During the 5-month implementation of FACTs, teachers and experts in subject didactics met three times to share and discuss their new knowledge and experience. At the end of the school year in June 2019, students in both the control and experimental groups were administered, the post-test and the efficacy of FACTs on the development of their inquiry skills was evaluated.

Data Analysis

The qualitative and quantitative analyses were divided into several parts according to the respective research question. The overall test score was analysed as well as specific scores for different inquiry skills. Two-sample t-test was used to compare the overall scores of experimental and control groups in the pre-test and post-test. A paired sample t-test was used to compare the differences in the level of students’ inquiry skills (post-test vs. pre-test) in the experimental and control group. Regarding the effect of the given variables (gender, overall assessment at the end of the term) on the test scores, conditions of the two-way analysis of variance (ANOVA) were complied with. The Shapiro-Wilk test was used to verify the normality of test scores or differences in all parts of the data analysis. In the data analysis, the significance level was $\alpha=.05$, $p<.05$, therefore the difference was considered significant. The SPSS software (SPSS Inc., 2009) was used to analyse the data.

Research Results

The basic descriptive statistics pertaining to the pre-test and post-test results in both groups can be seen in Table 4.

Table 4
Pre-test and Post-test Results

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experimental group</td>
<td>Control group</td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>.273</td>
<td>.228</td>
</tr>
<tr>
<td>SD</td>
<td>.138</td>
<td>.102</td>
</tr>
</tbody>
</table>

The results in Table 4 show that the initial level of inquiry skills failed to match the relative score of .3. In the pre-test, no statistically significant difference was found between the experimental and control groups ($p=.068$, therefore $p>\alpha$).
However, the comparison of the post-test results showed a statistically significant difference between the experimental and the control group ($p=.001$, therefore $p<\alpha$). Therefore, the H1 hypothesis was confirmed.

After the overall test results were evaluated, the results of students in the control and experimental groups respectively were grouped according to the inquiry skills measured (see Table 5).

### Table 5
Results of Students in the Control and Experimental Groups according to the Inquiry Skills Measured

<table>
<thead>
<tr>
<th>Inquiry skills</th>
<th>Pre-test</th>
<th></th>
<th></th>
<th>Post-test</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experimental group</td>
<td>Control group</td>
<td>Experimental group</td>
<td>Control group</td>
<td>Experimental group</td>
<td>Control group</td>
</tr>
<tr>
<td></td>
<td>$\bar{x}$</td>
<td>$SD$</td>
<td>$\bar{x}$</td>
<td>$SD$</td>
<td>$\bar{x}$</td>
<td>$SD$</td>
</tr>
<tr>
<td>Formulate a hypothesis</td>
<td>.223</td>
<td>.248</td>
<td>.154</td>
<td>.151</td>
<td>.233</td>
<td>.229</td>
</tr>
<tr>
<td>Design experiments (identify variables and their relationships)</td>
<td>.312</td>
<td>.193</td>
<td>.269</td>
<td>.172</td>
<td>.412</td>
<td>.225</td>
</tr>
<tr>
<td>Transform data to standard forms (i.e. tables or graphs)</td>
<td>.151</td>
<td>.297</td>
<td>.167</td>
<td>.260</td>
<td>.219</td>
<td>.289</td>
</tr>
<tr>
<td>Determine the relationship between variables (based on tables)</td>
<td>.219</td>
<td>.250</td>
<td>.250</td>
<td>.292</td>
<td>.288</td>
<td>.322</td>
</tr>
<tr>
<td>Determine the relationship between variables (based on graphs)</td>
<td>.370</td>
<td>.315</td>
<td>.292</td>
<td>.289</td>
<td>.449</td>
<td>.361</td>
</tr>
<tr>
<td>Determine accuracy (identify possible sources of errors)</td>
<td>.322</td>
<td>.305</td>
<td>.207</td>
<td>.260</td>
<td>.274</td>
<td>.344</td>
</tr>
</tbody>
</table>

Table 5 shows the scores obtained by the experimental and the control groups in the pre-test and post-test with regard to the measured inquiry skills. The results indicate low level of the monitored skills.

As can be seen in Table 2, in most cases, two tasks were used to measure the level of one inquiry skill. Therefore, students could gain 0, 1 or 2 points in total (or multiples of .5 point in tasks with two correct answers). The average success rate shows that students frequently scored 0 points. On the other hand, some students were able to score full 2 points. This range of results resulted in rather high $SD$ values in some inquiry skills.

Testing the H2 hypothesis, the differences in the inquiry skill development measured comparing pre-test and post-test results were analysed using the paired $t$-test. The paired $t$-test values in the experimental and control groups are listed in Table 6.

### Table 6
H2 Hypothesis Testing: Paired $t$-test Results

<table>
<thead>
<tr>
<th>Paired Differences</th>
<th>Paired samples test</th>
<th>$95%$ Confidence Interval of the Difference</th>
<th>$\bar{x}$</th>
<th>$SD$</th>
<th>$SD \bar{x}$</th>
<th>Lower</th>
<th>Upper</th>
<th>$t$</th>
<th>df</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Post-test – Mean Pre-test</td>
<td></td>
<td>.054</td>
<td>.156</td>
<td>.018</td>
<td>.018</td>
<td>.090</td>
<td>2.952</td>
<td>72</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>Mean Post-test – Mean Pre-test</td>
<td></td>
<td>.011</td>
<td>.152</td>
<td>.022</td>
<td>-.033</td>
<td>.055</td>
<td>.509</td>
<td>47</td>
<td>.613</td>
</tr>
</tbody>
</table>
The paired \( t \)-test value in the experimental group (Pair E) was \( p = .004 \), therefore \( p < \alpha \). In the experimental group, a statistically significant difference in the inquiry skill development, was confirmed. In the control group (Pair C), the paired \( t \)-test value was \( p = .613 \), therefore \( p > \alpha \), so that significant difference was not identified in this case. Therefore, the \( H_2 \) hypothesis was confirmed.

An overview of the changes in the students’ inquiry skills measured in the experimental group can be seen in the histogram (Figure 2). The histogram shows that the majority of students in the experimental group achieved better results in the post-test in comparison with the pre-test (most students manifested an approximately 10% improvement).

**Figure 2**

*The Inquiry Skill Development in the Experimental Group*

![Histogram showing inquiry skill development](https://doi.org/10.33225/jbse/21.20.204)

To compare the level of individual inquiry skills in the experimental group (pre-test and post-test), the paired \( t \)-test was used (Table 7).

**Table 7**

*Differences in the Development of the Individual Inquiry Skills in the Experimental Group: Pre-test and Post-test*

<table>
<thead>
<tr>
<th>Pair</th>
<th>Skill Description</th>
<th>( \bar{x} )</th>
<th>SD</th>
<th>( SD \bar{x} )</th>
<th>Lower</th>
<th>Upper</th>
<th>( t )</th>
<th>df</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pair 1</td>
<td>Formulate a hypothesis</td>
<td>.010</td>
<td>.310</td>
<td>.036</td>
<td>-.062</td>
<td>.083</td>
<td>2.83</td>
<td>72</td>
<td>.778</td>
</tr>
<tr>
<td>Pair 2</td>
<td>Design experiments (identify variables and their relationships)</td>
<td>.100</td>
<td>.244</td>
<td>.029</td>
<td>.043</td>
<td>.157</td>
<td>3.507</td>
<td>72</td>
<td>.001</td>
</tr>
<tr>
<td>Pair 3</td>
<td>Transform data to standard forms (i.e., tables or graphs)</td>
<td>.068</td>
<td>.366</td>
<td>.043</td>
<td>-.017</td>
<td>.154</td>
<td>1.598</td>
<td>72</td>
<td>.114</td>
</tr>
<tr>
<td>Pair 4</td>
<td>Determine the relationship between variables (based on tables)</td>
<td>.079</td>
<td>.454</td>
<td>.053</td>
<td>-.027</td>
<td>.185</td>
<td>1.481</td>
<td>72</td>
<td>.143</td>
</tr>
<tr>
<td>Pair 5</td>
<td>Determine the relationship between variables (based on graphs)</td>
<td>.068</td>
<td>.366</td>
<td>.043</td>
<td>-.017</td>
<td>.154</td>
<td>1.598</td>
<td>72</td>
<td>.114</td>
</tr>
<tr>
<td>Pair 6</td>
<td>Determine accuracy (identify possible sources of errors)</td>
<td>-.048</td>
<td>.401</td>
<td>.047</td>
<td>-.142</td>
<td>.046</td>
<td>-1.021</td>
<td>72</td>
<td>.311</td>
</tr>
</tbody>
</table>
Table 7 shows how the individual inquiry skills improved in students after the experimental intervention. A statistically significant difference in the pre-test and post-test was found in the “designing experiments” skill level ($p=.001$, therefore $p<\alpha$). In other inquiry skills, no statistically significant difference was found ($p>\alpha$).

The $H_3$ and $H_4$ hypotheses were tested by comparing the scores of students in the experimental group divided according to their gender and their overall assessment at the end of the term in pre-test and post-test (see Table 8).

Table 8
Scores of Students in the Experimental Group Divided According to their Gender and their Overall Assessment at the End of the Term in Pre-Test and Post-Test

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{x}$</td>
<td>$SD$</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>.278</td>
<td>.140</td>
</tr>
<tr>
<td>F</td>
<td>.267</td>
<td>.138</td>
</tr>
<tr>
<td>Overall assessment at the end of the term</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>.207</td>
<td>.126</td>
</tr>
<tr>
<td>PD</td>
<td>.297</td>
<td>.137</td>
</tr>
<tr>
<td>PW</td>
<td>.271</td>
<td>.140</td>
</tr>
<tr>
<td>Total</td>
<td>.273</td>
<td>.138</td>
</tr>
</tbody>
</table>

$PD$ – Passed with distinction, $PW$ – Passed very well, $P$ – Passed

Table 8 shows that a bigger difference in the pre-test and post-test results was identified in males. Based on the overall assessment at the end of the term, it can be stated that the biggest difference in the pre-test and post-test results was observed in average students ($P$), while the least impact could be observed in their excellent counterparts.

Figure 3 shows that in males, divided according to their overall assessment at the end of the term, the development in inquiry skills was less significant than in females. In females, divided according to their overall assessment at the end of the term, the differences were significant; the biggest development could be compared in females assessed as “Passed” at the end of the term. This group of females achieved the biggest development in comparison with males in terms of their assessment at the end of the term.
The results of the experimental group divided according to the students’ gender and their overall assessment at the end of the term in pre-test and post-test were statistically evaluated using the two-factor analysis of variance (Tables 9 and 10). The test aimed to identify whether the difference in the average test scores between the groups was merely accidental or statistically significant. The interaction between the “gender” and “overall assessment at the end of the term” factors determined whether the influence of gender on the average success rate in the pre-test and post-test depended on the “overall assessment at the end of the term” factor.

**Table 9**

Two-factor Analysis of Variance in the Pre-test

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>.128</td>
<td>5</td>
<td>.026</td>
<td>1.382</td>
<td>.242</td>
</tr>
<tr>
<td>Intercept</td>
<td>3.113</td>
<td>1</td>
<td>3.113</td>
<td>167.481</td>
<td>.001</td>
</tr>
<tr>
<td>Gender</td>
<td>.037</td>
<td>1</td>
<td>.037</td>
<td>1.975</td>
<td>.165</td>
</tr>
<tr>
<td>Overall assessment at the end of the term</td>
<td>.104</td>
<td>2</td>
<td>.052</td>
<td>2.784</td>
<td>.069</td>
</tr>
<tr>
<td>Gender * overall assessment at the end of the term</td>
<td>.034</td>
<td>2</td>
<td>.017</td>
<td>.920</td>
<td>.404</td>
</tr>
<tr>
<td>Error</td>
<td>1.245</td>
<td>67</td>
<td>.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6.795</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1.374</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the pre-test, no statistically significant difference was found between males and females (p=.165; p>α). There is no statistically significant difference between the groups of students with different overall assessment at the end of the term either (p=.069; p>α). The interaction (gender * overall assessment at the end of the term) was not confirmed: p=.404; p>α.

**Table 10**

Two-factor Analysis of Variance in the Post-test

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>.106</td>
<td>5</td>
<td>.021</td>
<td>.798</td>
<td>.555</td>
</tr>
<tr>
<td>Intercept</td>
<td>5.229</td>
<td>1</td>
<td>5.229</td>
<td>196.087</td>
<td>.001</td>
</tr>
<tr>
<td>Gender</td>
<td>.080</td>
<td>1</td>
<td>.080</td>
<td>3.001</td>
<td>.088</td>
</tr>
<tr>
<td>Overall assessment at the end of the term</td>
<td>.020</td>
<td>2</td>
<td>.010</td>
<td>.368</td>
<td>.694</td>
</tr>
<tr>
<td>Gender * overall assessment at the end of the term</td>
<td>.004</td>
<td>2</td>
<td>.002</td>
<td>.084</td>
<td>.920</td>
</tr>
<tr>
<td>Error</td>
<td>1.787</td>
<td>67</td>
<td>.027</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9.678</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1.893</td>
<td>72</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similarly, the post-test did not show statistically significant difference between males and females (p=.088; p>α) or groups of students with different overall assessment at the end of the term (p=.694; p>α). Again, the interaction (gender * overall assessment at the end of the term) was not confirmed: p=.920; p>α.

The H₃ and H₄ hypotheses were tested using the paired t-test (see Tables 11 and 12). The pre-test and post-test success rates were compared separately for males and females, and the groups of students with PD, PW, and P assessment at the end of the term. Each student was assigned one pair (success rate in pre-test and post-test) of mutually independent values.
Table 11
Comparison of the Inquiry Skill Level in Males and Females in the Pre-test and Post-test

| Paired samples test | Mean (SD) | 95% CI | Lower | Upper | t | df | p  
|---------------------|-----------|--------|-------|-------|---|----|-----
| Pair 1 Post-test_M - Pre-test_M | .084 (.163) | .028 (.09) | .022 (.141) | 3.032 | 33 | .005 |
| Pair 2 Post-test_F - Pre-test_F | .027 (.148) | .024 (.075) | -.020 (.75) | 1.160 | 38 | .253 |

In the experimental group, a statistically significant difference in the level of inquiry skills between the pre-test and post-test was identified in males ($p = .005; p < \alpha$). In females, the development of inquiry skills between the post-test and pre-test was not statistically significant ($p = .253; p > \alpha$). Therefore, the H3 hypothesis was confirmed.

Table 12
Comparison of the Inquiry Skill Level in Groups of Students with Different Overall Assessment at the End of the Term in the Pre-test and Post-test

| Paired samples test | Mean (SD) | 95% CI | Lower | Upper | t | df | p  
|---------------------|-----------|--------|-------|-------|---|----|-----
| Pair 1 P_Post-test - P_Pre-test | .114 (.126) | .035 (.038) | .190 (.258) | 12 | .007 |
| Pair 2 PW_Post-test - PW_Pre-test | .071 (.166) | .033 (.003) | .140 (.2149) | 24 | .042 |
| Pair 3 PD_Post-test - PD_Pre-test | .020 (.155) | .026 (-.034) | .073 (.742) | 34 | .463 |

PD – Passed with distinction, PW – Passed very well, P – Passed

A statistically significant difference in the inquiry skill level between the pre-test and post-test was identified in average students (P) ($p = .007; p < \alpha$) and very good students (PW) ($p = .042; p < \alpha$). However, in excellent students (PD), the difference in the inquiry skill level between the pre-test and post-test was not statistically significant. Therefore, the H4 hypothesis was confirmed.

Discussion

Considering the first research question, the research showed that the level of inquiry skills identified in the pre-test was low (see Table 4). This result may be connected with the fact that the teacher-centred teaching style dominates in many schools in Slovakia. This teaching style is based mainly on transmissive methods with a small number of active-learning elements (OECD, 2009). The teacher-centred teaching style and traditional assessment has no significant impact on the development of these skills (OECD, 2018a, 2018b).

In the pre-test, no statistically significant difference was found between the control and the experimental group ($p = .068$). This fact was important for the experimental intervention.

In terms of the research, the inquiry skill levels in students were compared after teaching experimental and control groups with and without using FACTs respectively. The identified inquiry skill levels in both groups were measured using a post-test to provide data for evaluating the efficacy of teaching using FACTs referred to in the second research question.
The post-test results revealed that the experimental group scored higher in the total average than the control group, and the difference was statistically significant (p<.05). By comparing the results of the experimental group in the pre-test and post-test, a statistically significant difference in the students' inquiry skill level was identified (p=.004) while the difference in the control group was not statistically significant (p=.613). The research findings indicate that teaching using FACTs is more efficient than teaching without them.

The research results presented in this study supplement the results of Bell and Cowie (2001), Grog et al. (2017), and Harlen (2003) who have confirmed the positive impact of FA on the inquiry skill development in the context of IBSE. The results of this research clearly show that using FACTs – even without special intervention concerning systematic implementation of IBSE – is effective in the development of inquiry skills.

It relates to the fact that using FA in teaching helps students achieve a higher level of understanding of the subject matter (Schunk & Swartz, 1993) and develops their higher-order cognitive skills (Brookhart, 2010; Shute, 2008; Kluger & DeNisi, 1996).

In terms of the individual inquiry skill development, the experimental group achieved a statistically significant difference only in the “designing experiments” skill. Although there was some development in other inquiry skills (except the ability to determine the accuracy of data), it was not statistically significant (see Tables 5 and 7). It may relate to the fact that multiple high-frequency FACTs were used to develop the “designing experiments” skill in the learning process (see Table 3). Other skills were developed using a lesser frequency of FACTs. Based on these findings, it can be stated that it is necessary to implement multiple types of FACTs with higher frequency during the given time period to develop the respective skills.

Another research question was: “How do different factors (gender and overall assessment at the end of the term) affect the efficacy of FACTs in terms of developing the selected inquiry skill?”

For this purpose, the results of students in the experimental group were divided according to their gender and overall assessment at the end of the term, and compared (see Tables 8, 9, and 10). The pre-test and post-test showed no statistically significant difference between the inquiry skill level in males and females (see Tables 9 and 10). The research findings indicate that both genders profit from learning using FACTs almost equally. Even the Slovak students’ results in scientific literacy according to the PISA 2018 national report have pointed out that scientific literacy remains the only area in which no statistically significant difference was measured across different reporting cycles in males and females in the Slovak Republic (Miklovičová & Valovič, 2019).

However, by comparing the differences between the pre-test and post-test results in the experimental group (see Table 11), a statistically significant difference in the males’ inquiry skill level was identified (p=.005). However, there was no statistically significant difference in the females (p=.253). As can be seen in Figure 3, in the best performing females (PD – passed with distinction), no improvement of inquiry skills could be observed. It may have resulted from anxiety induced by too much responsibility for academic performance or the fact that their learning style focused on memorising. However, the development in inquiry skills between females and males with PW (passed very well) and P (passed) overall assessment at the end of the term did not differ significantly. The results achieved by females with PD assessment at the end of the term affected the overall improvement in inquiry skills, but it was not found significant after all. These findings are similar to the results of the research focused on motivation and engagement in male students compared to female students with regard to inquiry-based teaching (Kuo et al., 2020). However, most researchers investigating the differences between males and females state that gender is not a determining factor in effective implementation of FA (e.g., Ajobgeje, 2013; Asadifard & Afghari, 2019; Moyosore, 2015).

By comparing the pre-test and post-test results (see Table 12), a statistically significant difference in the average (P) (p=.007) and very good students (PW) (p=.042) was found in terms of their inquiry skill level. However, in excellent students (PD), the difference was not statistically significant (p=.463). The aforementioned indicates that using FACTs in teaching improves the inquiry skill level mainly in average-performing students. The PISA 2018 international comparative measurement of scientific literacy has identified a high number of low-performing students (31.4%) referred to as the risk group (Miklovičová & Valovič, 2019). These students are falling behind their peers due to lack of support at school or at home from parents. These students show weak internal motivation, low self-confidence, inability to endure when a complex problem needs to be solved, and also lack of confidence in their own abilities. However, using FACTs has helped them improve their confidence in their own skills and become proud of their achievements (Black & Wiliam, 1998b; Flórez & Sammons, 2013). They have also become more active in learning, more ready to cooperate and independent when provided
formative assessment (Boston, 2002; McMillan, 2007). When students become more self-confident, they learn to attribute their failure and success to the factors they can control and believe in their ability to succeed if they try hard enough. These findings are in line with the results of the study that has confirmed that the greatest influence of FA is on the lower-achieving students (Black & William, 1998a).

Research Limitations

The compared results may have been influenced by the following.

The teachers who participated in the research had access to a database of FACTs for individual subjects. They selected and used individual FACTs at their own discretion. If a teacher did not have access to pre-existing FACTs or if they had a negative attitude to FACTs, the statistical difference could be less significant.

The limitations of FA implementation in this research include a relatively small sample of upper secondary school students and teachers. However, the FACTs were applied during the same time period and in the same topics, therefore the results allow for drawing conclusions that could help implement FACTs into the teaching-learning process with regard to the development of skills in upper secondary school students.

Other limitations of the FACTs implementation include the fact that their preparation is time-consuming and can be difficult, teachers lack training in FA, and their effect only becomes visible in the long run. These particular limitations were eliminated in this research as the FACTs were prepared in advance to make it easier for the teachers; material equipment and methodological guidance were also provided.

Conclusions and Implications

The results of this research confirmed the efficacy of teaching using FACTs in terms of the inquiry skill development while using the currently prevalent teacher-centred teaching style in science education. This finding adds to the available evidence supporting the importance of FA and its role in education.

A detailed analysis showed that using FACTs in teaching and learning developed the selected inquiry skills (design experiments, transform data to tables or graphs, determine relationship between variables) in students. The statistical analysis revealed a statistically significant difference only in the “designing experiments” skill, in which FACTs were used with the highest frequency. The efficacy with which the selected skills were developed correlates with the frequency of using FACTs.

Moreover, it was found that teaching using FACTs stimulated learning mainly in the lower performing students, in this case, the highest efficacy in terms of inquiry skill development could be observed. The research indicates that the problems faced by the low-performing students (whose number is very high in Slovakia according to OECD PISA 2018) can be tackled by using FACTs in teaching.

Gender was not a determining factor in effective implementation of FA. The fact that inquiry skills improved more significantly in males than in females after the experimental intervention may result from the female’s anxiety about maintaining outstanding academic performance reflected in their overall assessment at the end of the term as well as their learning style. This finding needs to be verified by future research.

Based on this research, it is recommended to include FA in (future) teacher training. If FA is to be successfully introduced in Slovak schools, the deeply rooted beliefs of Slovak teachers regarding assessment need to change first. It is also necessary to provide them with proper support in implementing FA into their day-to-day practice on the level of educational policy (investing into teachers’ professional development, mentoring and peer-support, FA research). In the initial stages of the process, ready-made FACTs databases for individual subjects could also be provided.

However, these requirements are not addressed systematically. They have been partially addressed by the authors of this paper who are involved in VEGA (Scientific Grant Agency, Ministry of Education & Slovak Academy of Sciences) and KEGA (Cultural and Educational Grant Agency, Ministry of Education) research projects focused on building the support system for implementing FA by creating teacher materials addressing the goals and ways to implement FA into education; digital collections of FACTs; methodological materials addressing the implementation of FA into the educational process based on research and verification; systems of contact and online teacher education focused on new assessment forms.

Last but not least, methodological and professional guidance must be continually provided to (future)
FORMATIVE ASSESSMENT AS A TOOL TO ENHANCE THE DEVELOPMENT OF INQUIRY SKILLS IN SCIENCE EDUCATION (pp. 204-222)

teachers regarding their selection and adaptation of the learning content, and creation of the school educational programmes with integrated FA. In this way, we can be certain that students are exposed to the long-term and complete benefits of the influence of FA.

Currently, the methodology centres in Slovakia are starting to provide this kind of training focused on further teacher education under the “Teachers” (professional development) project supported by the European Social Fund. The knowledge and experience gained during this research will be applied in the creation of the “Formative assessment for promoting learning in students” educational project and its implementation.

Acknowledgements

This research was supported by the VEGA No. 1/0265/17 “Formative Assessment in Natural Sciences, Mathematics and Informatics” and KEGA No. 004UPJS-4/2020 “Creation, Implementation, and Verification of the Effectiveness of Digital Library with the Formative Assessment Tools for the Natural Sciences, Mathematics and Informatics at the Elementary School” grants.

References

Brookhart, S. M. (2010). How to assess higher-order thinking skills in your classroom. ASCD.


DfES. *Department for Education and Skills. Assessment for learning: Eight schools project report*. DfES. https://dera.ioe.ac.uk/7600/1/1f1ab286269a7e24df53c86372da971.pdf


Received: September 28, 2020  
Accepted: March 05, 2021


Mária Ganajová  
(.corresponding author)  
PhD, Associate Professor, Pavol Jozef Šafárik University in Košice, Faculty of Science, Department of Didactics of Chemistry, Šrobarová 2, Košice, Slovakia.  
E-mail: maria.ganajova@upjs.sk

Ivana Sotáková  
PhD, Lecturer, Pavol Jozef Šafárik University in Košice, Faculty of Science, Department of Didactics of Chemistry, Šrobarová 2, Košice, Slovakia.  
E-mail: ivana.sotakova@upjs.sk  
ORCID: https://orcid.org/0000-0002-8910-5786

Stanislav Lukáč  
PhD, Associate Professor, Pavol Jozef Šafárik University in Košice, Faculty of Science, Department of Didactics of Mathematics, Jesenná 5, Košice, Slovakia.  
E-mail: stanislav.lukac@upjs.sk

Zuzana Ješková  
PhD, Associate Professor, Pavol Jozef Šafárik University in Košice, Faculty of Science, Department of Didactics of Physics, Park Angelinum 9, Košice, Slovakia.  
E-mail: zuzana.jeskova@upjs.sk

Veronika Jurková  
PhD, Lecturer, Pavol Jozef Šafárik University in Košice, Lifelong Learning Centre and Projects Support, Šrobarova 2, Košice, Slovakia.  
E-mail: veronika.jurkova@upjs.sk

Renáta Orosová  
PhD, Lecturer, Pavol Jozef Šafárik University in Košice, Faculty of Arts, Department of Education, Moyzesova 9, Košice, Slovakia.  
E-mail: renata.orosova@upjs.sk