Science Education Outcome Monitoring Systems Globally

Alhumaidi Alderaan*, Amjad Alderaan,

Trainers, Technical and Vocational Training Corporation (TVTC), Saudi Arabia

*Email: a.alderaan@outlook.com

Abstract

The motivation for this paper was the inadequacy of international competitions to provide continuous monitoring of science education outcomes. Many countries have developed their models, and systems and implemented specific projects for regular monitoring of science education outcomes. Google Scholar was used for identifying papers and PRISMA was used for screening and selection of papers. After all screening, 25 papers were selected for this review. The reviewed papers reflected the global dimensions of the problem and its solutions. Many of the papers were discussion and review papers using a wide variety of findings from different countries. Overall, this review demonstrated the wide range of models, frameworks, techniques and tools usable for regular monitoring of science education outcomes in different countries and thus globally. Some limitations of this review are mentioned at the end of this paper.

Keywords: Science, Education Monitoring Systems, Review

Introduction

Monitoring and evaluation is a management function for decision-making based on progressive performance. If expected results are not achieved, the causes for them are investigated and rectified to the extent possible, so that future progress is nearer to the goals. The purposes of monitoring educational outcomes are the same. In science education monitoring additional variables like critical thinking, problem-solving, and project management may also be included. Scores obtained in targeted tests are the main parameters of science education performance outcomes.

Three trends are important in this respect: STEM education, PISA and TIMSS international assessment. Science, Technology, Engineering, and Mathematics (STEM) professionals generate a stream of scientific discoveries and technological innovations that fuel job creation and national economic growth. Undergraduate STEM education prepares graduates for today's STEM professions and those of tomorrow, while also helping all students develop knowledge and skills they can draw on in a variety of occupations and as citizens (National Academies of Sciences, 2018).

The Programme for International Student Assessment (PISA) is a triennial global-level assessment of the Organisation for Economic Co-operation and Development (OECD) in member and non-member nations aimed to evaluate educational systems by measuring the scholastic performance of 15-year-old school students in mathematics, science, and reading. In the case of science, PISA measures scientific competencies and knowledge in physical, living, earth and space systems. It assesses these in the contexts of personal, local/national, and global levels. It assesses the extent to which the students have acquired knowledge and skills for participating in economic and social life. It does not test how well they reproduce what they had learned by the end of their term. PISA is unique in its policy orientation, innovative concepts on literacy, relevance to lifelong learning, regularity, and coverage breadth. In its 2018 tests, 66

countries participated. Students from 27 countries participated in global competence tests and the module questionnaire. Students from 39 countries participated in the module questionnaire only (OECD, 2018).

Trends in International Mathematics and Science Study (TIMSS) is a specific method of assessing the competence of 4th and 8th-grade school students in mathematics and science at the global level. It is the largest and most comprehensive assessment of mathematics and science for primary and secondary education. It is a global enterprise consisting of more than 70 educational systems participating in the assessment. TIMSS tests were introduced in 1995. Since then, the TIMSS tests are conducted every four years. Hence, it has the longest trend in mathematics and science achievement. TIMSS also collects rich background information from the assessed students, their mathematics and science teachers, school principals, and parents of the grade four students, along with the system-level data. This helps to provide a holistic perspective of education in the participating countries. Since 2019, attempts to increasingly digitalise the system is in progress (Mullis, Martin, Foy, L, & Fishbein, 2020).

Both STEM and TIMSS are international systems of assessing mathematics and science education systems. Many countries do not participate in these two assessments. They may have very effective methods of assessing science education outcomes. Thus, the scope of the topic is very wide. However, international monitoring occurs every 3-5 years only. For more frequent, short-term monitoring, nations need to devise their methods. The question, here, is whether and how they monitor science education outcomes and determine methods of improving the performance of students in science subjects.

This paper reviews the various monitoring systems of science education outcomes practised in different countries and compares their relative effectiveness using TIMSS and STEM as the benchmarks.

Method and Results

Method

Being a topic of public interest also, many useful articles available on Google have been cited above. Now research articles from Google Scholar are considered. The search terms derived from the title of the topic were used. Only papers available in English were selected. Abstracts were included if they contained useful information. The identified papers were screened and selected using the PRISMA flow process. The process yielded 25 papers finally for this review. The selected papers are discussed and tabulated in various ways to achieve the aim of this review.

Results

Keeves (2004) pointed out the issues of shortage of science and mathematics teachers as most students opted for computer science and internet-related education. PISA was introduced to enhance and evaluate the outcomes of science education. The curriculum implementation theory and the related model (Fig 1) were developed from a workshop conducted by the IEA in Sweden in 1971.

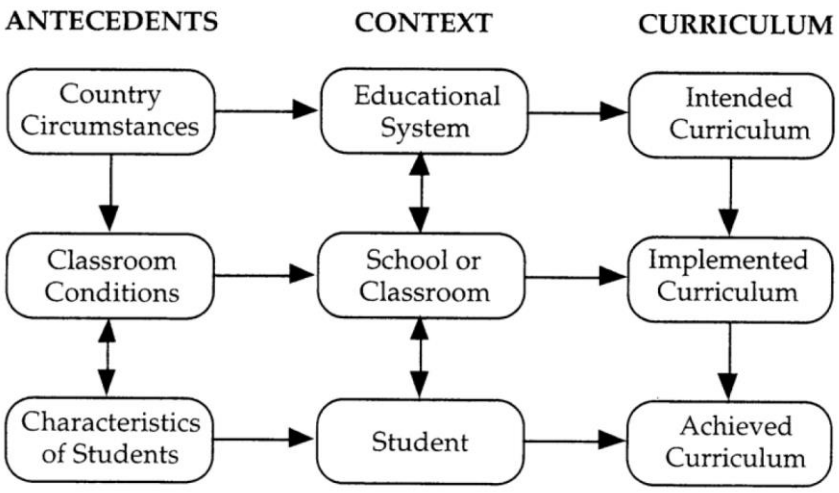

Figure 1 Curriculum Implementation Model (Keeves, 2004).

The theory stipulates that curriculum exists at three levels: intended, implemented, and achieved levels. These are influenced by antecedents and contexts. The intended curriculum is set by the policies of the government or the institution. What is implemented depends on the teacher and school. Achievement is the result of the gap between intended and implemented curricula. It reflects the extent of the student's learning from what was planned and offered to them. The contexts of the three are the educational system, the school, and the student respectively. In another model, Carroll developed a model for school learning to predict the success of complex learning tasks. The components of the model are three variables in terms of time (aptitude, perseverance and opportunity), and instructional quality. This model was further developed into a causal model of student performance, as given in Fig 2.

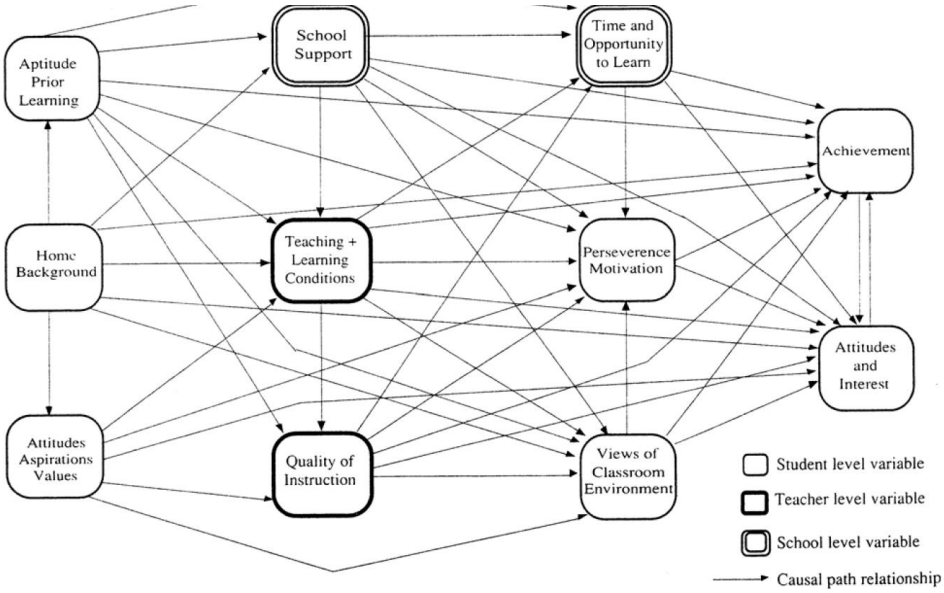

Figure 2 Causal model of student performance (Keeves, 2004).

Dahoff (1967) developed a third model for IEA in the 1967 Lake Mohonk Conference as a cross-national model of educational achievement in a national economy. The most important aspect of

this model (Fig 3) is the focus given to three policy-making frame variables: (a) the environment and economy, (b) demand for manpower, (c) curriculum content, and (d) the objectives of education. However, all nations do not consider all these points in their educational systems.

Of particular interest for policy making in education are the frame variables. However, they depend on (a) the environment and economy, (b) demand for manpower, (c) curriculum content, and (d) the objectives of education.

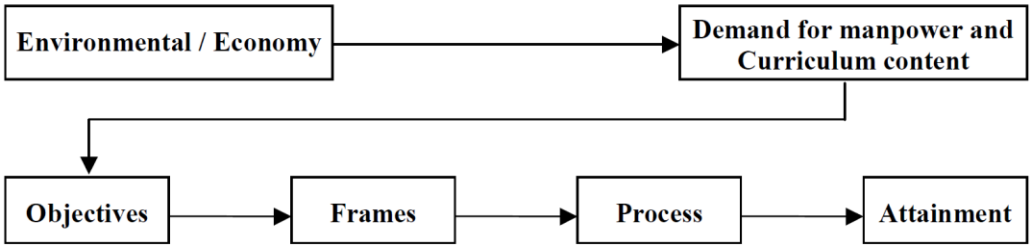

Figure 3 Cross-national model of educational achievement in a national economy (Keeves, 2004).

Another model proposed at the same Lake Mohonk Conference was the Input-Output-Utilisation model (Fig 4). It includes many components related to the monitoring of science education outcomes. These are financial, production conditions, structure, and operations (educational structure, equipment, agents, curriculum, and instructional methods), outputs (knowledge, skills, attitudes, participation, attainment level), and utilisation (employment, community involvement, and family activity). How to measure and at what stage was not considered for this complex model.

The retentivity model of school learning assumed that the underlying distribution of intellectual ability in the complete age cohort is the same in all countries and the differences in mean scores and variances in any cross-country comparisons are due to differences in the selection procedures in different countries. Although this is an oversimplification of a complex subject, the results obtained in many tests validated it.

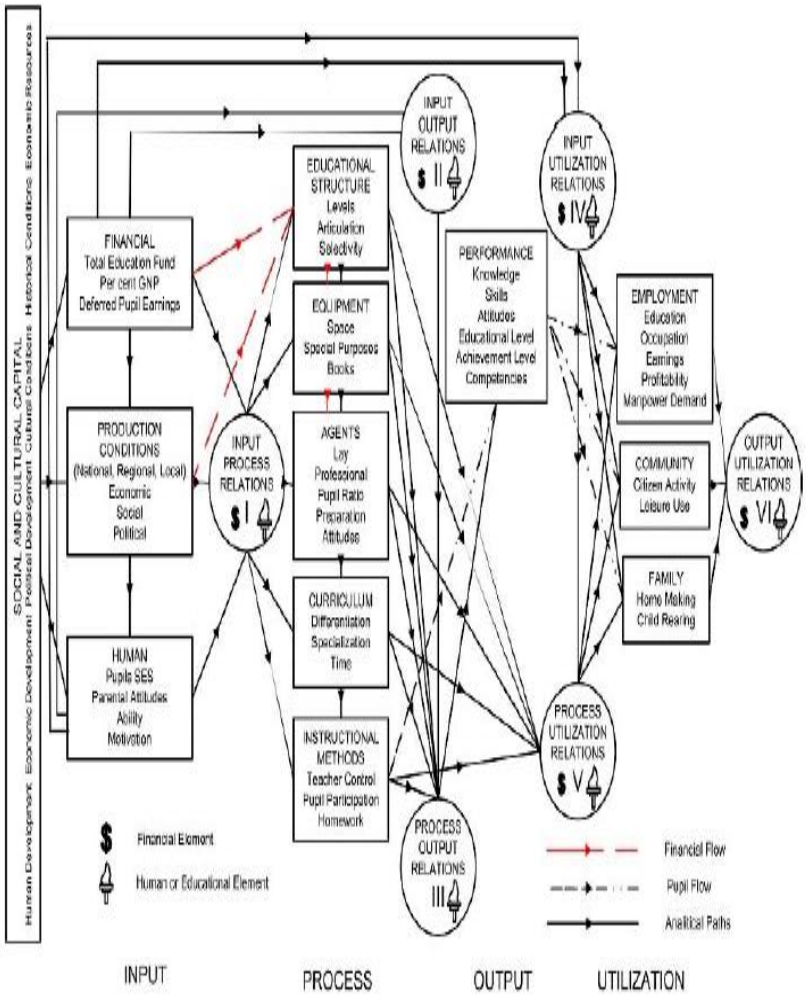

Figure 4 Input-Output-Utilisation model (Keeves, 2004).

Another model from Lake Mohonk Conference was the educational environment model for educational achievement (Fig 5). This model was meant to study the factors of change in the performance of students over time. These factors are related to home, school, and peer groups. The three environments are characterised by the structural, attitudinal and process dimensions of the educational system of the country. The interrelationships among the three environments are also important.

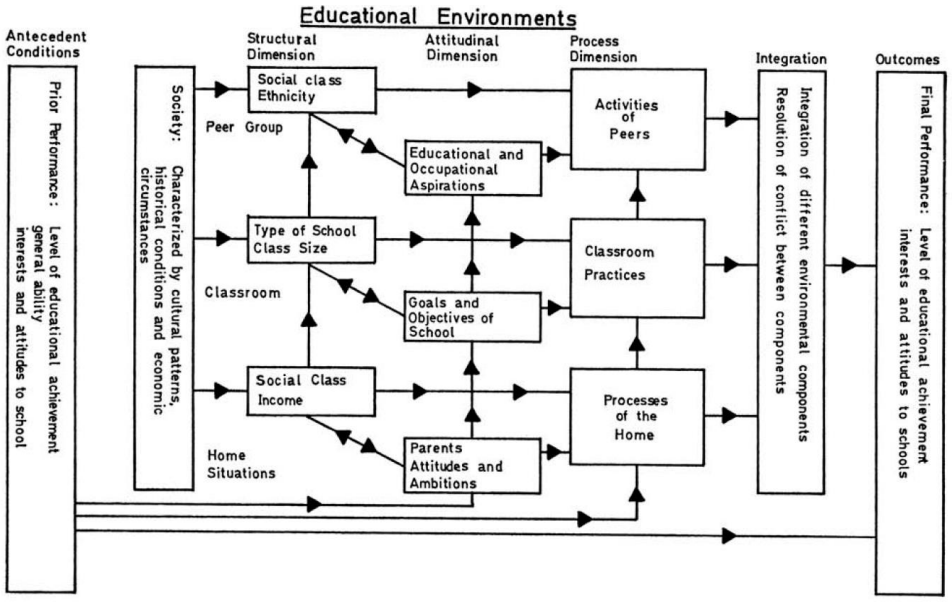

Figure 6. Model for the Study of Educational Environments.

Figure 5 Educational environment model for educational achievement (Keeves, 2004).

All the above models are meant for cross-sectional studies. For longitudinal studies, new models were required. A high level of correlation was observed between achievements in mathematics and science and the labour markets in the Asia-Pacific region, but not in the Western region. These differences were also reflected in the relationship between mathematics and science achievements and labour quality. The need to include information technology in school mathematics and science curricula at appropriate levels is stressed. The choice of a career in science and technology may be easier for achievers of high scores in mathematics and science. More students with a positive attitude towards mathematics and science are essential for the economic growth of the country. Usually, female students perform better in grade examinations, but not in competitive examinations. Female students seem to have a greater capacity to remember what they studied but may not possess the critical thinking and problem-solving capabilities required for competitive examinations. Although gender segregation is an important cultural aspect of Islamic countries, it does not seem to have any effect on female education as per the recent trends. Increased student engagement, parents; investment, the relationship of science with the environment, and the need to create more interest in teaching and learning science are ten issues discussed by the author.

Most of the research on science education outcomes had assessed academic achievement and other significant learning outcomes. Students spend over 15000 hours at school by the end of their high school senior years. Naturally, they are concerned about what lies in future for them. Their reactions, attitudes, and perceptions of their school experiences are important in this regard. (Fraser, 1998) reviewed the progress made over the previous 30 years in conceptualisation, investigation and evaluation of the factors affecting the sociopsychological aspects of the learning environments of classrooms and schools. To study these factors, students are the best sources. Various research methods can be used to extract the feelings of students about their learning environments leading to valuable findings. Learning environment inventory (LEI), Classroom environment scale (CES) were developed and widely used in educational research works over the last 30 years. Lewin's (1936) field theory recognised both the

environment and its interaction with the personal characteristics of the individual as potent determinants of human behaviour. The Lewinian formula, $B = f(P, E)$, stressed the need for new research strategies to consider behaviour as an interaction between the person and the environment. Research using Lewin's theory started in 1938. It is essential to distinguish between class (room) and school environments. The latter is a wider sphere of sociopsychological influence on students, in which class environments and beyond are included. Organisational climate description questionnaire (OCDQ) and College characteristics index (CCI) are adaptations from organisational research and are widely used in research works on school environments. The choice of the unit of analysis is important to define the operational areas of each factor in the enquiry. The author has described and tabulated the characteristics and use contexts of the classroom environment research instruments Learning Environment Inventory (LEI); Classroom Environment Scale (CES); Individualised Classroom Environment Questionnaire (ICEQ); My Class Inventory (MCI); College and University Classroom Environment Inventory (CUCEI); Questionnaire on Teacher Interaction (QTI); Science Laboratory Environment Inventory (SLEI); Constructivist Learning Environment Survey (CLES); and What Is Happening In This Class (WIHIC). For the rarely used school environment research, Organizational Climate Description Questionnaire (OCDQ), Work Environment Scale (WES) and School-Level Environment Questionnaire (SLEQ) are available in their adapted and short forms. Past research using different instruments, different sample sizes and country contexts has firmly established the association between student perceptions of the classroom environment and performance outcomes in science education in any country. Multilevel analyses have been used due to the inherently hierarchical nature of classroom settings. The success of cooperative learning versus competitive or individualistic learning has also been studied well. Classroom and school environment was found to be strong predictor of both achievement and attitudes even when a comprehensive set of other factors was held constant. Students and teachers might have different perceptions of the classroom environment, as evidenced by many studies. Many other variables like class size, grade level, subject matter, the nature of the school-level environment, teacher personality, cultural differences and gender of students and the type of school affect the students' perceptions of the classroom environment. School environment and classroom environment may not be related to each other. Boys preferred competition and individualistic learning and girls preferred cooperative learning. Other gender differences regarding class or school environments also exist. Class assessments, feedback, reflections, discussions, interventions, and re-assessments can help teachers to improve their teaching. Adoption of mixed research methods, greater emphasis on the school environment also linking it with the classroom environment, the help of a psychologist for improvement in perceptions and attitudes, multicountry studies, effects of transition from primary to high school and education and assessment of teachers are some new research trends. Many points in this paper are not current, as the paper was published in 1998. The situation could have changed over time and the status may be entirely different. For instance, the increased use of IT and the effect of covid pandemic on the large-scale adoption of online education could have changed the science education outcomes.

In a review after 20 years of an older review, Hofstein and Lunetta (2004) suggested that learning science in the laboratory facilitates special attention to the scholarship. The scholarship is associated with models of learning, argumentation and the scientific justification of assertions, students' attitudes, conditions for effective learning, students' perceptions of the learning environment, social interaction, and differences in learning styles and cognitive abilities. These

factors are derived from the goals for learning, discrepancies, and matching goals, students' perceptions of teachers' goals, teachers' expectations and behaviour, the laboratory guide, incorporating inquiry empowering technologies, simulations, and the laboratory, assessing students' skills and understanding of inquiry, and the politics of schooling. Teacher education and professional development are two important aspects to consider. The need for proper assessment techniques for desirable outcomes from science education is obvious. Trained teachers need to practise what was taught to them in this respect. Inadequate resources might stand in the way of the effective implementation of science laboratory reforms. This may lead to poor outcomes. Practice should be led by the goals of science education for desirable performance outcomes.

The paper by Arvanitis, et al. (2009) reported on the human factors and qualitative evaluation of mobile-enabled AR systems for the science education of physically disabled students. The tool was developed for the EU-funded CONNECT project. It assists the users to contextualize and reinforce their learning in schools, science centres and homes. CONNECT AR also encourages learners to visit science centres and perform experiments out of reach in schools. Building on these experiences back at school and home with visual augmentations, they can communicate through web-based streaming technology. Some specific methods of assessment may be required to monitor the science education outcomes of these physically disabled children under the CONNECT project.

The effect of a multilevel multifaceted method for detection of the outcomes of a hands-on science education programme was evaluated by Ruiz-Primo, Shavelson, Hamilton, and Klein (2002). In this approach, the authors used different assessments based on their proximity to the curriculum implemented. Immediate assessments were artefacts (students' products) due to the curriculum. Close assessments were similar to the content and activities of the unit or curriculum. Proximal assessments recorded knowledge and skills related to the curriculum but sometimes on different topics. Distal assessments reflected the state or national standards in a particular knowledge domain. These different types of assessments were tested in a Bay area school district, using the instructional units of variables, mixtures, and solutions. Close assessments were found to be more sensitive to the changes in students' pre- to post-test performance than proximal assessments. Instructions impacted performance. Characteristics of assessments influenced the level of detection of improvements in students' performance. The sensitivity of assessments was influenced by some of their characteristics.

The purpose and method of using rubrics in the assessment of student's performance in science education were discussed by Allen and Tanner (2006). Rubrics help both teachers and students as tools for making learning goals and evaluation criteria explicit for them.

A review of 17 papers led Rutten, Joolingen, and Van Der Veen (2012) to conclude that traditional instruction is enhanced by using computer simulations. It can be used as an add-in like a pre-laboratory exercise or visualization tool. Simulation conditions showed improved learning outcomes, with effect sizes up to 1.54. This improvement in outcome was by improvement in better understanding of the concepts and the ability to predict results with less time. Initiation, participation, perception of the classroom and instructional support improved leading to higher student satisfaction. These results were applied to studies using computer simulations as laboratory exercises also. These results were short-term effects. More studies on long-term effects are required.

One of the scalable ways of involving undergraduates in science research is the introduction of course-based undergraduate research experience (CURE) in colleges. Although the outcomes of CURE are like those who complete research internships, the design and implementation of CURE are quite different. To study the necessary and sufficient aspects of CUREs to achieve desired student outcomes, a systems approach was used by Corwin, Graham, and Dolan (2015). The authors developed pathway models as hypotheses to test and validate or refute in future research. A review of papers suggested increased content knowledge, technical skills, analytical skills, scientific self-efficacy, project ownership, and career clarification as the outcomes of CURE. The likelihood of CURE students pursuing further education in science was also demonstrated. Inadequate evidence existed for increased access to teachers, mentors and understanding of the nature of science. Three mini-models were formed and combined into a single large model (Fig 6) and suggested evaluation at three stages.

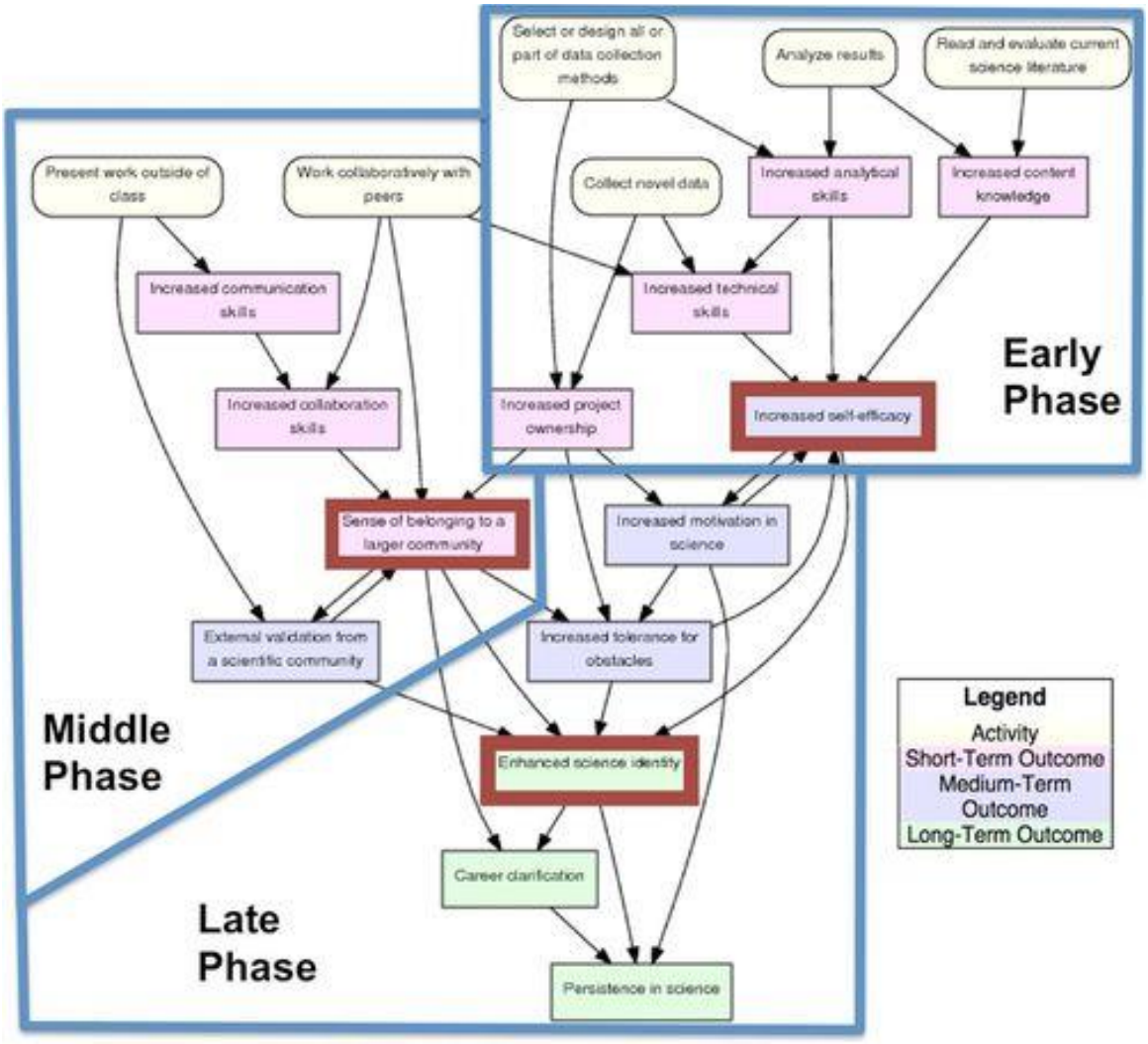

Figure 6 A large CURE model showing different phases and hubs. Arrows for positive directional relationships, blue boxes for the evaluation phases, and red borders for outcomes (Corwin, Graham, & Dolan, 2015).

To develop scientific attitudes among communities, large-scale citizen science programmes were implemented in the USA and southern Canada. The Monarch Larva Monitoring Project aimed to study the distribution and abundance of monarch butterflies in the project regions. An evaluation of youth involvement in the monitoring programme Kountoupes and Oberhauser (2008) showed the success of the programme in terms of successful activities by the youth enjoying them all along. The success and educational value of the programme for children was enhanced by innovations by adults. This was achieved without compromising the data integrity. Many adults conducted independent research and worked on extension activities to enhance monitoring observations on the butterfly.

An instrument was proposed to evaluate systemic reforms in science education and was validated by Scantlebury, Boone, Kahle, and Frase (2001) using a survey of 8000 science and mathematics students taught by 1000 teachers in 200 schools in the USA. The instrument consisted of four factors and 20 items. The survey showed that the class, home, and peer environments influenced the achievement of students. Class environment (standards-based teaching practices) was the best independent predictor of achievement and attitude of students.

The results of a survey of 1063 undergraduates by Huang, Bernacki, Kim, and Hong (2022) showed that their final examination performances were predicted strongly by their self-efficacy perceptions. This influence decreased with decreasing online monitoring behaviour. The previous GPA predicted one of the three combinations of high and low self-efficacy and metacognitive monitoring. High levels of both led to the highest performance outcomes followed by high self-efficacy and low monitoring activities and then by low levels of both.

In the education system of the USA, the education of students is monitored using standardized tests regularly. The monitoring tools are used to compare the average performance of students living in different states or belonging to different subgroups like gender, race, ethnicity, or parental income and to monitor their progress over time. The country's only large-scale monitoring system is the National Assessment of Educational Progress (NAEP). The NAEP data has been utilised for the construction of the Stanford Education Data Archive (SEDA). This publicly available database can be used to analyse and identify patterns of achievement for any school district in the country (Fahle, Shear, & Shores, 2019).

In a German mixed methods study of third-year medical students, the traditional and blended problem-solving learning did not differ in student performance. However, the blended intervention increased subjective learning and satisfaction as the students assessed working with the web-based learning environment as very good (Woltering, Herrler, Spitzer, & Spreckelsen, 2009).

Student attendance is a dominant factor in academic performance. This relationship, reported by several works was confirmed by Newman-Ford, Fitzgibbon, Lloyd, and Thomas (2008) The authors used UniNanny®, an electronic attendance monitoring system developed at the University of Glamorgan, to evaluate 22 first-year modules within four separate award programmes using attendance data gathered and stored electronically. While confirming the earlier reports on the relationship between attendance and performance, it was found that the more a student attends classes, the higher the chances of failure in academic achievements. This enhances the chance to score high grades. However, attendance decreased substantially over time. The early morning lectures were not affected by the decline in attendance.

An android-based monitoring application for students' performance was developed and validated by Sulistyowati, Setyaningrum, Kumala, and Hudha (2018). The results of a survey of experts on content, media, language, teachers and 40 parents showed the usefulness of the application for both teachers and parents was demonstrated by the results. Especially, parents found it useful to monitor how their children are performing in the classes. Some technical difficulties were experienced in the process of data transfer to the application.

According to Georghiades (2004), metacognition improved the learning skills of primary school students, especially in science education. The need for more research blending metacognition with science education was stressed.

Studies (Davis, 2000) showed that self-monitoring prompts encouraged planning for and reflection on activities. Thus, it helped the students to display an integrated understanding of the relevant science. On the other hand, activity prompts guided the inquiry process but were less successful in prompting knowledge integration.

In e-learning of medical education, the evaluation needs to include peer reviews and assessment of outcomes concerning learner satisfaction, content usability, and demonstration of learning. Educators will perform the role of facilitators and assessors in this system rather than acting as content distributors to the students (Ruiz, Mintzer, & Leipzig, 2006).

According to Caldwell (2007) the clickers used in audience response systems have either a gentle or positive effect on student performance on exams. The specific effect depends on the method and extent of their use. The positive effect is higher and there is a more active atmosphere in large classrooms. These systems can be used for introducing and monitoring peer learning methods in large classrooms. Clickers improve attendance and retention when they are linked to grades and are used daily. Decreased content coverage due to the use of clickers is more than compensated by the benefits. Some practical tips have been offered on the correct use of clickers to maximise benefits.

The factors affecting visitor learning from a science museum were studied by Falk and Storksdieck (2005) using mixed methods. Prior knowledge, interest, motivation, choice and control, social interactions within and between groups, orientation, advance organisers, architecture, and exhibition design affected visitor learning. All these factors could individually explain visitor learning, but not with adequate explanatory power. The framework of Falk and Dierking's Contextual Model of Learning was useful to understand the complex interactions between factors.

In a bibliometric analysis, Arici, Yildirim, Caliklar, and Yilmaz (2019) observed that the use of augmented reality (AR) and mobile learning in science education enhanced motivation and attitude of students leading to better performance achievements. Mobile applications and marker-based materials on paper were used most for AR as they were easy to use and could be developed easily and practically. In a related study on the impact of using AR for learning, Sahin and Yilmaz (2020) observed higher levels of achievement and more positive attitudes towards the course by the students compared to the control group (without AR). The students were happy to use AR and wanted to continue using it in the future. There was no anxiety among them when using AR applications. There was a significant positive correlation between academic achievement and attitude with the use of AR technology for learning.

Technology-enabled mathematics and science education at grade levels of 5 to 13 leads to positive impacts on attitude and student learning. The overall effect is moderated by the provision for teachers' training. Effect sizes on the positive impact on learning were larger if digital tools were used along with other tools, and not when used as a substitute for other methods. These results were obtained by Hillmayr, Ziernwald, Reinhold, Hofer, and Reiss (2020) from a meta-analysis of 92 papers.

The issues related to lower educational outcomes in Uzbekistan and Indonesia were compared by Shaturaev and Bekimbetova (2021). In the case of Uzbekistan, despite spending about 24% of the budget on primary education, the student achievement levels are low due to a shortage of teachers in rural areas, traditional teaching methods, frequently updating textbooks and an excessive number of students per classroom. The increasing population adds to the problem. In the case of Indonesia, the low achievement levels, despite nearly 100% enrolment in primary classes, are due to the high rates of dropouts, the high cost of education, and difficulties to access schools. Both countries are spending almost a quarter of their budgets on primary education. After investigating these causes, the authors recommended solutions for them.

Adaptive learning, smart campus, teacher evaluation, intelligent tutoring robots, and virtual classrooms are only a few of the applications of educational AI. An evaluation of AI in teaching and learning science led Alam (2022) to conclude that AI has a beneficial effect on both the quality of instruction provided by teachers and on the learning outcomes of students. However, some challenges related to complying with the rules of law, preventing the digital divide, algorithmic divide, lack of basic technologies, skills, infrastructure, safety, ethical issues, accountability, and decrease of social communication skills with increasing use of AI.

Discussions

The trends of international assessments of science education (TIMSS, PISA) impact the policies of some countries to elevate the performance levels of their children. It is widely recognised that regular monitoring of science education outcomes is the only way to improve the achievement levels of science students. This review showed that much research has been done in modelling, and analysing the data on science performance outcomes, interventions, methods, and tools used for enhancing science performance outcomes. However, the problem remains due to various challenges related to infrastructure, resources, and implementation issues. These observations were derived from the 25 reviewed papers in this article. Some common trends may be interesting to understand the effects of the nature of the enquiry, methods used, country contexts and limitations affecting the findings. These trends are discussed below with the help of data.

Years of publication

The number of reviewed papers published in different years is given in Fig 7.

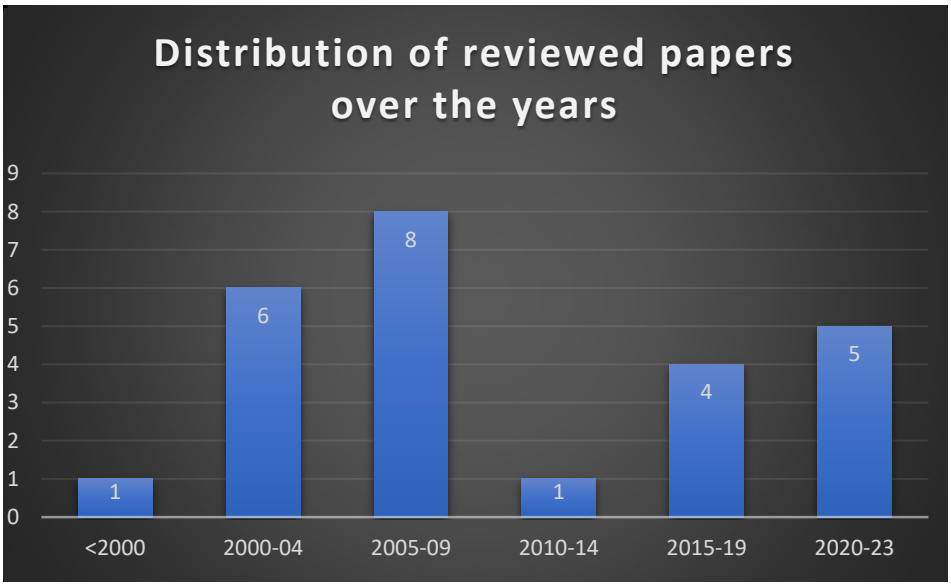

Figure 7 Frequency distribution of reviewed papers over the years.

Out of the total 25 papers reviewed, eight were published during 2005-2009 and six were published during 2000-2004. Thus, the total number of papers published during the 10 years from 2000 was 14 (56%). There were five recent publications belonging to 2020-2023 also. More papers would have been possible with a wide search. But this was avoided due to the fear of increasing the length of the paper beyond the limit.

Aims of papers

The frequency distribution of reviewed papers according to their aims is given in Fig 8.

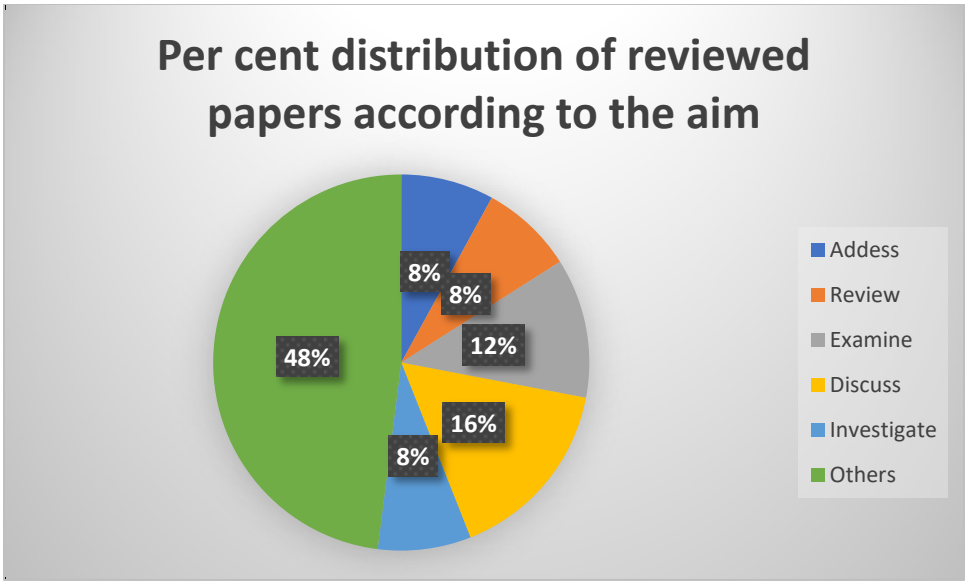

Figure 8 Per cent distribution of reviewed papers according to their aims.

The aim to discuss some aspects related to the review topic dominated with 16%. About 12% of the papers examined various aspects. Some aims had only one paper each and these were

grouped as Others, which accounted for 48% of the papers. About 8% of the papers each investigated, addressed or reviewed the topic of interest.

Methods

The reviewed papers categorised according to the methods of study are given in Fig 9.

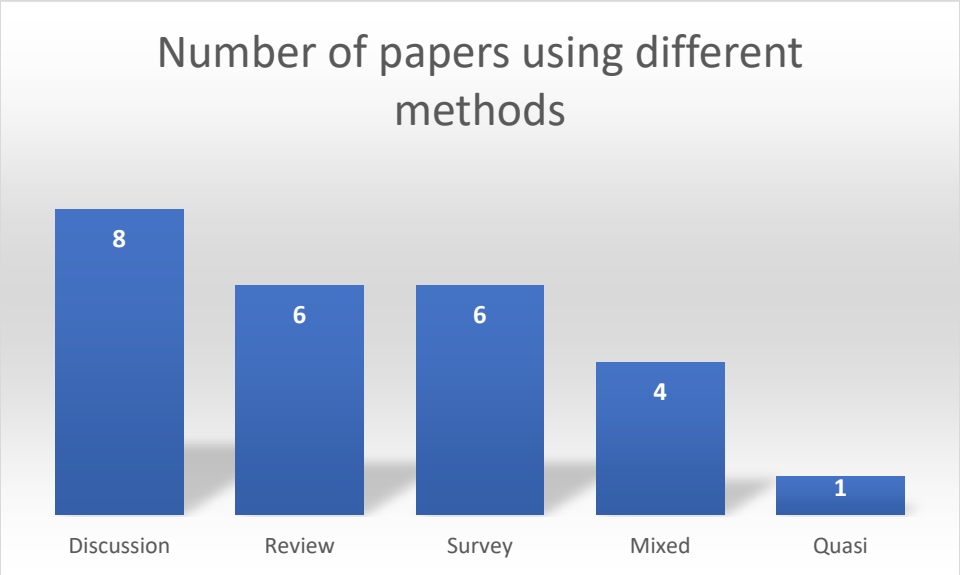

Figure 9 Number of papers using different methods.

Out of the total 25 papers, eight were discussion papers, six each were reviews and case studies and four used mixed methods. One paper used a quasi-experimental design. The trend to use existing information to generate new concepts is evident from the fact that 14 out of 25 papers were discussion or review papers. Discussions are also based on past research.

Limitations of papers

Limitations were reported or could be identified only from eight out of 25 papers. Some reviews or discussion papers did not mention anything about the limitations of their works. Some abstracts included in the papers also did not mention anything about limitations. A few of the limitations were derived from the full-text papers.

Countries

The topic of review had a global dimension. Hence, the countries from where the data were collected were important. Sometimes, the countries from where the data were collected were different from countries from the countries to which the authors belonged. The data considered in reviews and discussions were from different countries and hence global. The frequencies of countries from where the data were collected are provided in Fig 10.

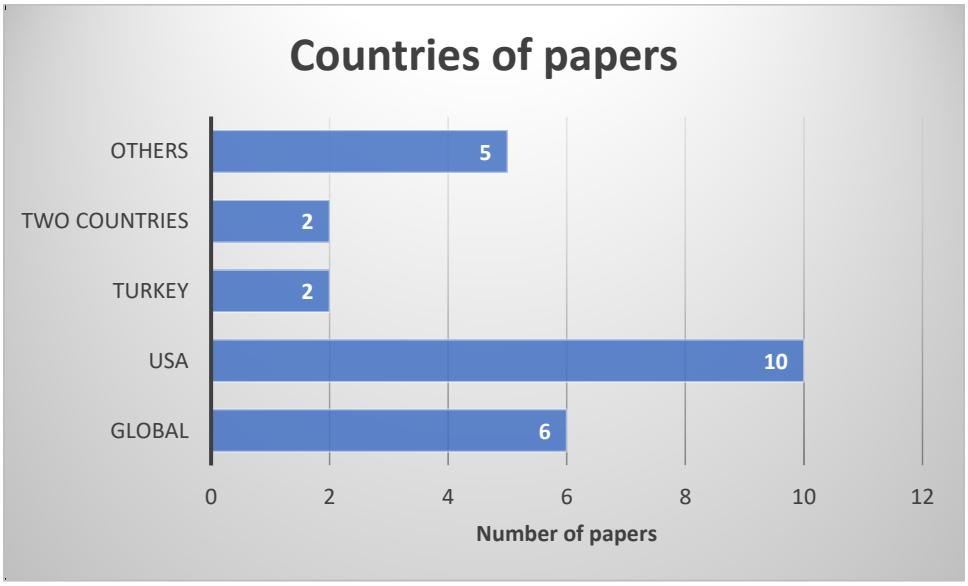

Figure 10 Countries of papers.

The maximum number of papers were published in the USA. There were six papers of global nature. In the case of two papers, data from two countries were compared. These were comparisons between USA and Canada and between Uzbekistan and Indonesia. There were single papers from Germany, Cyprus, the Netherlands, the UK, and Indonesia.

Overall, the reviewed papers covered a wide range of country contexts, aims and methods to identify different models, techniques, and tools to measure science education outcomes. Thus, it can be said that the review is a fair evaluation of the status of science education outcomes monitoring globally.

Conclusions

International competitions like TIMSS, STEM and PISA use very detailed methods to assess science education outcomes. Countries can monitor science education outcomes through the data published over the years. However, these international competitions are conducted at intervals of 3-5 years. Therefore, countries need their science education monitoring systems for short-term regular monitoring.

Models of science education monitoring have been proposed by different authors. However, the relative merits or demerits of these models have not been compared yet. The countries can use what is best for them by adapting some of these models to their contexts.

Many authors described definite projects to apply and evaluate science education monitoring systems. It is very difficult to choose the best one from them. In most cases, the outcomes are positive. But in the absence of comparing these improvements with the TIMSS, PISA or STEM scores, the adequacy of these improvements cannot be evaluated.

The reviewed papers were categorised based on their year of publication, aim, method and country. The wide range of areas of the topics covered is evident from these analyses.

Limitations

Many papers selected were published earlier than 2010. Special efforts had to be made to select more recent papers. Abstracts did not include methodological details in a few papers.

References

Alam, A. (2022). Employing Adaptive Learning and Intelligent Tutoring Robots for Virtual Classrooms and Smart Campuses: Reforming Education in the Age of Artificial Intelligence. In R. N. Shaw, S. Das, V. Piuri, & M. Bianchini (Ed.), *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022. Lecture Notes in Electrical Engineering, vol 914*, pp. 395-406. Singapore: Springer Nature Singapore. doi:10.1007/978-981-19-2980-9_32

Allen, D., & Tanner, K. (2006). Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners. *CBE—Life Sciences Education, 5*(3), 197-203. doi:10.1187/cbe.06 – 06 – 0168

Arici, F., Yildirim, P., Caliklar, Ş., & Yilmaz, R. M. (2019). Research trends in the use of augmented reality in science education: Content and bibliometric mapping analysis. *Computers & Education, 142*(December), 103647. doi:10.1016/j.compedu.2019.103647

Arvanitis, T. N., Petrou, A., Knight, J. F., Savas, S., Sotiriou, S., Gargalakos, M., & Gialouri, E. (2009). Human factors and qualitative pedagogical evaluation of a mobile augmented reality system for science education used by learners with physical disabilities. *Personal and ubiquitous computing, 13*(November), 243-250. doi:10.1007/s00779-007-0187-7

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sciences Education, 6*(1), 9-20. doi:10.1187/cbe.06 –12– 0205

Corwin, L. A., Graham, M. J., & Dolan, E. L. (2015). Modelling course-based undergraduate research experiences: An agenda for future research and evaluation. *CBE—Life Sciences Education, 14*(1), es 1. doi:10.1187/cbe.14-10-0167

Davis, E. A. (2000). Scaffolding students' knowledge integration: Prompts for reflection in KIE. *International Journal of Science Education, 22*(8), 819-837. doi:10.1080/095006900412293

Fahle, E. M., Shear, B. R., & Shores, K. A. (2019). Assessment for monitoring of education systems: The US example. *The ANNALS of the American Academy of Political and Social Science, 683*(1), 58-74. doi:10.1177/0002716219841014

Falk, J., & Storksdieck, M. (2005). Using the contextual model of learning to understand visitor learning from a science center exhibition. *Science education, 89*(5), 744-778. doi:10.1002/sce.20078

Fraser, B. J. (1998). 5.1 science learning environments: Assessment, effects and determinants. In *International handbook of science education* (pp. 527-564). Kluwer Academic Publishers. Retrieved May 26, 2023, from https://surveylearning.moodle.com/cles/papers/Handbook98.htm

Georghiades, P. (2004). From the general to the situated: Three decades of metacognition. *International journal of science education, 26*(3), 365-383. doi:10.1080/0950069032000119401

Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., & Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education, 153*(August), 103897. doi:10.1016/j.compedu.2020.103897

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science education, 88*(1), 28-54. doi:10.1002/sce.10106

Huang, X., Bernacki, M. L., Kim, D., & Hong, W. (2022). Examining the role of self-efficacy and online metacognitive monitoring behaviors in undergraduate life science education. *Learning and Instruction, 80*(August), 101577. doi:10.1016/j.learninstruc.2021.101577

Keeves, J. P. (2004). Monitoring the Learning and Teaching of Science in a changing world. *International Education Journal, 5*(3), 275-293. Retrieved May 26, 2023, from https://files.eric.ed.gov/fulltext/EJ903855.pdf

Kountoupes, D. L., & Oberhauser, K. S. (2008). Citizen science and youth audiences: educational outcomes of the Monarch Larva Monitoring Project. *Journal of Community Engagement and Scholarship, 1*(1), 10-20. Retrieved May 27, 2023, from https://pdfs.semanticscholar.org/d09c/6892bdde7524967c259d8da7b303a5000bbc.pdf

Mullis, I. V., Martin, M. O., Foy, P., L, K. D., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science.* TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). Retrieved May 26, 2023, from https://timss2019.org/reports/wp-content/themes/timssandpirls/download-center/TIMSS-2019-International-Results-in-Mathematics-and-Science.pdf

National Academies of Sciences, E. a. (2018). *Indicators for Monitoring Undergraduate STEM Education.* (M. B. Rosenberg, M. L. Hilton, & K. A. Dibner, Eds.) Washington, DC: The National Academies Press. doi:10.17226/24943

Newman-Ford, L., Fitzgibbon, K., Lloyd, S., & Thomas, S. (2008). A large-scale investigation into the relationship between attendance and attainment: a study using an innovative, electronic attendance monitoring system. *Studies in Higher Education, 33*(6), 699-717. doi:10.1080/03075070802457066

OECD. (2018). *PISA 2018 Results: Are students ready to thrive in an interconnected world?* OECD Publishing, Paris. doi:10.1787/d5f68679-en

Ruiz, J. G., Mintzer, M. J., & Leipzig, R. M. (2006). The impact of e-learning in medical education. *Academic medicine, 81*(3), 207-212. doi:10.1097/00001888-200603000-00002

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 39*(5), 369-393. doi:10.1002/tea.10027

Rutten, N., Joolingen, W. R., & Van Der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & education, 58*(1), 136-153. doi:10.1016/j.compedu.2011.07.017

Sahin, D., & Yilmaz, R. M. (2020). The effect of Augmented Reality Technology on middle school students' achievements and attitudes towards science education. *Computers & Education, 144*(January), 103710. doi:10.1016/j.compedu.2019.103710

Scantlebury, K., Boone, W., Kahle, J. B., & Frase, B. J. (2001). Design, validation, and use of an evaluation instrument for monitoring systemic reform. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 38*(6), 646-662. doi:10.1002/tea.1024

Shaturaev, J., & Bekimbetova, G. (2021). Indigent condition in education and low academic outcomes in public education system of Indonesia and Uzbekistan. *Scientific Research Archive, 1*(1), 1-11. Retrieved May 28, 2023, from https://www.researchgate.net/profile/Jakhongir-Shaturaev/publication/357271097_INDIGENT_CONDITION_IN_EDUCATION_AND_LOW_ACADEMIC_OUTCOMES_IN_PUBLIC_EDUCATION_SYSTEM_OF_INDONESIA_AND_UZBEKISTAN/links/61c466afabcb1b520adb0427/INDIGENT-CONDITION-IN-EDUCATION-

Sulistyowati, P., Setyaningrum, L., Kumala, F. N., & Hudha, M. N. (2018). Android-based monitoring applications of students' learning outcomes. *IOP Conference Series: Materials Science and Engineering: 3rd Annual Applied Science and Engineering Conference (AASEC 2018). IOP Conf. Series: Materials Science and Engineering 434, No 1*, p. 012036. IOP Publishing. doi:10.1088/1757-899X/434/1/012036

Woltering, V., Herrler, A., Spitzer, K., & Spreckelsen, C. (2009). Blended learning positively affects students' satisfaction and the role of the tutor in the problem-based learning process: results of a mixed-method evaluation. *Advances in Health Sciences Education, 14*(January), 725-738. doi:10.1007/s10459-009-9154-6