The Impact of Inquiry-Based Practices on the 2019 NAEP Twelfth Grade Science Assessment

Allan Jay Dinglasan[1] Mingyuan Zhang[2]*

[1,2]Central Michigan University, United States of America

DOI: https://doi.org/10.56293/IJMSSSR.2022.4554

IJMSSSR 2023
VOLUME 5
ISSUE 1 JANUARY – FEBRUARY **ISSN: 2582 - 0265**

Abstract: This study examined the impact of inquiry-based practices on standardized science assessment scores for twelfth-grade students. It focused on the frequency of such inquiry practices, such as coming up with research questions, using evidence to explain why something happens, using tables and graphs to identify relationships between variables, and using information for argumentation on a scientific idea impacted average scale scores. An analysis of data taken from the National Assessment of Educational Progress (NAEP) dataset was done to compare student responses from survey data to science achievement scores of twelfth-grade students. This secondary data analysis was completed using multiple t-tests to compare means and determine significance. While the study found substantial effect sizes observed when analyzing the frequency of certain IBL practices, such as using evidence to explain why something happens as well as using tables and graphs to identify relationships between variables, the impact of having students come up with research questions to explore how something works as well as using the information to disagree with someone about a scientific idea, had negligible to no impact on their standardized test scores. The study confirmed the efficacious nature of inquiry-based practices to improve learning outcomes for students that had opportunities to learn them. However, it also questioned the predictive nature of standardized tests such as the NAEP science assessment regarding proficiency in these soft skills. The results challenged IBL proponents to justify their advocacy for this pedagogy, given that the same outcomes are evident even if an instructor's focus on inquiry practices is not as intense.

Keywords: Inquiry-based learning, Standardized testing, Science education, Guided inquiry, Science instruction, STEM skills

Introduction

Generations of science students remember their science classes in K-12 and even through higher education as comprising of teacher-driven lectures and performing cookie-cutter experiments in the laboratory more akin to following a recipe from a cookbook rather than a genuine investigation into scientific phenomena. While Science, Technology, Engineering, and Mathematics (STEM) has been practiced this way for decades, educational theorists have long championed that the best way to practice science in the classroom is to get students actively engaged in constructing knowledge through the process of scientific inquiry. When the College Board purveyors of Advanced Placement courses revamped the AP Chemistry course in the mid-2010s to further reflect this pedagogical mindset, implementing a guided inquiry approach to science courses was viewed as a logical first step in getting students to be more successfully engaged in the scientific process.

While inquiry-based learning (IBL) has been around for a long time, STEM courses are just beginning to evolve to accommodate this pedagogical approach. As the COVID-19-induced pandemic in 2020 forced educational institutions worldwide to quickly transition to remote learning, the question arose as to whether inquiry-based learning can viably be implemented in this environment. While instructors of courses such as chemistry and other pure sciences that are both highly theoretical and require practical laboratory skills have been slowly integrating this approach, the efficacy of IBL and whether there are increased learning outcomes remain uncertain. The degree to which inquiry-based learning is seen as fostering more significant learning outcomes than traditional approaches and whether it can thrive in an online learning environment will affect how students and instructors respond to the changes associated with this pedagogy.

This study explores the impact of students' inquiry-based mindset, attitudes, and practices on twelfth graders' 2019

National Assessment of Educational Progress (NAEP) Science Assessment scores. NAEP provides robust insight into these variables and their possible correlation to student achievement via assessments. The 2019 NAEP results will be analyzed for this study.

With much attention having been paid over the years to the use of inquiry-based learning (IBL) strategies, various studies have long indicated a "clear, positive trend favoring inquiry-based instructional practices" (Minner, Levy & Century, 2010), with particular emphasis on the kind of instruction that promotes active thinking by students as well as concluding data gathered. Minner et al. (2010) pointed to how teaching strategies that actively engage students in the learning process through scientific investigations, such as those practiced by students in IBL courses and the mindset that is being cultivated in that environment, are more likely to increase conceptual understanding than are strategies that rely on more passive techniques such as traditional teacher-centered didactic instruction. While the improved learning outcomes evident in studies involving IBL have not necessarily been manifested through increased student grades in science courses, there have been learning gains associated with more enhanced STEM skills and more developed science literacy (Wu, Sandoval, Knight, et al., 2021) as well with student engagement and motivation (Buchanan, Harlan, Bruce, et al., 2016). Students' attitudes towards STEM learning have been observed in a more positive light because of experiencing more IBL practices in their classrooms, with students exhibiting more positive attitudes towards science courses as a result (Riegle-Crumb, Morton, Nguyen & Dasgupta, 2019). The same study also concluded that overall, the "weight of evidence leans toward the conclusion that the attitudes of students from different gender and racial/ethnic backgrounds are similarly associated with greater exposure to inquiry-based instruction in both their science and mathematics classrooms."

The popularity of guided instructional strategies such as those purported by IBL advocates notwithstanding, a study by Kirchner, Sweller & Clark (2006) pointed to how these approaches "ignore both the structures that constitute human cognitive architecture and evidence from empirical studies over the past half-century that consistently indicate that minimally guided instruction is less effective and less efficient than instructional approaches that place a strong emphasis on the guidance of the student learning process." They reiterated that any positive advantages that stem out of the inquiry-based learning pedagogy begin to recede only when learners have sufficiently high prior knowledge to provide "internal" guidance (Kirchner, Sweller & Clark, 2006), thus allowing for the possibility that increased learning outcomes may not prove to be as universal as once thought.

While certain studies have indeed found that students also exhibited a degree of resistance to IBL because the kind of instruction inherent requires much effort from students, the adverse reaction to these heightened expectations (Seidel & Tanner, 2013) and students' resistance to taking on a more active role in their learning (Cooper, et al., 2017) serves to highlight that the promising evidence that inquiry-based learning promotes positive attitudes toward STEM disciplines remains inconclusive (Riegle-Crumb et al., 2019). As such, the potential for any transference between positive outcomes for students with a cultivated IBL mindset and familiar with IBL practices in their science classroom and course grades or academic performance remains to be seen. How the 2019 NAEP Science Assessment is impacted by the mindset and practices of the twelfth graders surveyed could serve to make this connection.

Evidence points to the efficacy of inquiry-based learning strategies, but its effectiveness has often been observed in various contexts and not necessarily on higher learning outcomes measured in student grades and course performance. Many studies that purport success in implementing inquiry-based learning practices observed in secondary school classrooms delved into learning outcomes other than student grades. Thus, the potential exists that the increased learning outcomes often observed may not necessarily transfer into an improvement in course grades as students complete secondary school. Likewise, recent studies that pointed to the effectiveness of inquiry learning and problem-based learning on student achievement in a science course specifically had been observed in a primary school setting (4th grade) and not in a secondary school classroom (Hartini and Ferawati, 2016), leading to more questions on the nature of its efficacy.

While studies on the effectiveness of inquiry-based learning run the gamut from primary school to undergraduate courses in higher education, very little evidence exists on whether its promise of increased learning outcomes extends to students and courses at the secondary school level. Not only is this period a significant one in the academic growth of a student, but it is also when students solidify their interest in STEM careers and decide to pursue degrees in science-related disciplines. Knowing whether exposure and familiarity with inquiry-based practices in secondary school impact their immediate and long-term academic future offers valuable insight into the extent of IBL efficacy.

It is essential not to overlook how most studies on the efficacy of inquiry-based learning practices were based on findings and subjects outside of North America. Since the NAEP science assessment was conducted in the United States with data observed and gathered from students in American schools, the extent to which previous evidence touting the effectiveness of IBL could potentially remain in doubt as far as a transference of these improved outcomes becoming evident.

Understanding the effectiveness of inquiry-based learning approaches to science courses adds to the collective knowledge of the efficacious practices that improve student outcomes. The findings will provide increased awareness of the pedagogical approaches and methodology that can serve as viable alternatives to traditional teaching methods employed in science courses, especially as many laboratory-based science courses transition from a face-to-face or hybrid to a fully online setting. The findings will also offer educational researchers, school administrators, and curriculum developer's valuable insight on designing science courses that will ensure continued success for students regardless of setting and keep learners engaged and motivated in the scientific inquiry process early on.

Specifically, the present study will explore the following questions:

1. Are student science assessment scores impacted by student experiences in coming up with research questions to explore how something works?
2. Are student science assessment scores impacted by student experiences in using evidence from experiments to explain why something happens?
3. Are student science assessment scores impacted by student experiences in using tables or graphs to identify relationships between variables?
4. Are student science assessment scores impacted by student experiences in using information to disagree with someone about a scientific idea?

Literature Review

An examination of the recent literature on inquiry-based learning, its benefits on student attitudes and learning outcomes as well as the intersection between standardized testing and the potential transference of the affordances of inquiry-based practices on test scores demonstrates the need for a re-examination of the role, if any, that the frequency of inquiry-based practices plays on increased student outcomes and whether this translates to gains in test scores for standardized tests like that of NAEP.

The Nature of Inquiry-Based Learning and Scientific Literacy

Educational theorists have long promoted science learning at its most ideal with students are actively engaged in constructing knowledge through the process of scientific inquiry (Dewey, 2007; Piaget, 1977; Vygotskiĭ, 1978). It is argued that it is through inquiry-based learning that successful engagement into the scientific process can best be achieved. While there have been various iterations of how inquiry-based learning has been defined over the years, the working definition cited most in the literature is consistent with what the National Research Council outlined in 2000. As such, inquiry-based learning generally denotes a particular pedagogical environment where learners (1) are engaged by scientifically oriented questions, (2) give priority to evidence, which allows them to develop and evaluate explanations that address scientifically oriented questions, (3) formulate explanations from this evidence, (4) evaluate their explanations in light of alternative explanations, particularly those reflecting conceptual understanding and (5) communicate and justify their proposed explanations (National Research Council et al., 2000). These activities may include making observations, framing questions and hypotheses, designing and conducting scientific investigations, formulating scientific explanations and models based on evidence and logic, communicating results, and revising the explanations or revisiting the investigations based on feedback and critique from peers (National Research Council et al., 2000; Weaver et al., 2008). Inquiry-based learning is practiced along an inquiry continuum based on the degree of independence that students exert or is afforded in the inquiry process (Bell et al., 2005; Fay et al., 2007; Weaver et al., 2008; Wheeler & Bell, 2012). Confirmation labs, or so called "cookie-cutter" experimentation that has often been the staple of science courses are at one end of the inquiry continuum. Instructors are often tasked with providing the research question and procedure and students are responsible for confirming a known outcome. The second type of learning in the continuum is structured inquiry where the research question and the procedure are both provided but students are asked to explore the unknown outcome through the inquiry. Advanced Placement (AP) Chemistry courses are governed by the third type of learning in the inquiry continuum in the form of guided inquiry. In this type of inquiry, only the research question is provided, and students are tasked with designing and conducting the investigation to answer the question. Finally in the open inquiry type, also referred to as authentic inquiry, students are given a raw or general topic or phenomenon and are tasked with formulating the research question at the same time as designing and conducting the investigation (Bielik & Yarden, 2016; Buck et al., 2008; Rowland et al., 2016).

Science education standards established by American Association for the Advancement of Science (AAAS) and the National Research Council (NRC) have urged less of an emphasis on memorizing scientific facts in the classroom and more emphasis on students investigating the everyday world and developing deep understanding from their inquiries. (Marx et al., 2004) As such, the importance of developing scientific literacy has been regarded as a co-requisite aim of inquiry-based learning. Among the aspects contributing to a scientifically literate individual involves an ability to ask the proper questions and generate plausible hypotheses (Scardamalia & Bereiter, 1992) especially when the nature of the questioning could shed light on how one is learning (Graesser & Person, 1994).

Potential Benefits of IBL on Student Attitude

Recent research shows positive academic and achievement gains for students engaged in IBL work and the practice has been steadily growing. (Buchanan et al., 2016) Scientific attitude, being one of the important attitudes that can support the learning of chemistry, for example, makes students more active in learning and elicits curiosity that is related to achievement. (Huda & Rohaeti, 2021) That study found that using an inquiry model provided more experience for students and that scientific attitude can improve for better learning achievement. Research has shown though that expectation of better attitudes from students (as a result of exposure to inquiry-based practices) are not always as apparent when viewed through the lenses of its other affordances. While exposure to inquiry-based learning motivated students towards learning in their chemistry courses for example (Adeoye, 2020; Rohaeti, Prodjosantoso & Irwanto, 2020), when a more argument driven inquiry model was implemented, that instructional method did not change attitudes of students but only their argumentation skills (Demircioglu & Ucar, 2012).

There are also mixed reactions recorded by students as far as their perceptions of inquiry-based learning are concerned. While students found the hands-on nature of IBL to be very enjoyable, a study found that other students found that the inherent lack of structure in the method to be distressing. (Frezell, 2018). Negative impressions seem to depend mostly on the specific aspects of the inquiry classroom experience as implemented and not on the style that which laboratory investigations are conducted. (Baseya & Francis, 2011) The mixed results as far as attitudes extend to other aspects of a student's life as well. Studies have shown that inquiry learning experiences are potential predictors for students' career aspirations (Kang & Keinkonen, 2017) and may also increase their scientific creativity (Kirici & Kakirci, 2021). It has also become evident that the more students are exposed to inquiry-based practices, the more they gain confidence in their own scientific abilities. (Brickman et al., 2009) However this exposure to inquiry-based practices, that have previously had a positive impact on students' conceptual understanding and scientific process skills, did not necessarily make any difference on their attitudes towards science. (Simsek & Kabapinar, 2010)

There were also gender differences associated with the degree to which inquiry-based learning affects students. Wolf & Fraser (2008) found that where males benefited more from inquiry methods, females seemed to benefit more from non-inquiry approaches in terms of attitudes to science and classroom task orientation, cooperation and equity. A more recent study (Riegle-Crumb et al., 2019) affirmed these gender differences as far as having more positive attitudes toward science and mathematics. Riegle-Crumb's (2019) study not only showed a higher frequency of inquiry-based instruction being significantly associated with greater interest, perceptions of utility, and self-efficacy for these subjects, but also that some evidence does exist indicating that male students' perceptions of science utility are higher in relation to more inquiry-based instruction. One can conclude then that the attitudes of students from different gender and racial/ethnic backgrounds are similarly associated with greater exposure to inquiry-based instruction in both their science and mathematics classrooms. (Riegle-Crumb et al., 2019)

Gender differences aside, engagement appears to rank high among the learning outcomes that have increased because of a more intensive focus on inquiry-based learning in STEM classrooms. Recent findings have suggested that authentic inquiry experiences can serve as an effective approach for engaging students in high-enrollment, introductory science courses. (Ben Wu et al., 2021) The study also concluded that this approach could facilitate further development of science literacy and STEM skills of all students, skills that are critical to students' personal and professional success and to informed engagement in civic life. (Ben Wu et al., 2021)

Inquiry-Based Learning and Learning Outcomes

Surveys of literature on the topic point to positive trends favoring inquiry-based instructional practices, particularly the type of instruction that emphasizes student active thinking and drawing conclusions from data. (Minner, Levy & Century, 2010). While Khalaf & Zin's (2018) review of empirical studies showed that traditional learning is supposed to increase learners' outcomes and keep them active during the learning process, it also found widespread assertions that inquiry-based learning increases learners' knowledge and skills. Research has since identified a number of important drawbacks to both traditional and inquiry-based learning that have existed in the previous works (Khalaf & Zin, 2018).

Certain programs such as POGIL have emerged that focuses on guided-inquiry as a pedagogical model for implementing inquiry-based learning. Process Oriented Guided Inquiry Learning (POGIL), with roots in chemistry but now widely used across a range of disciplines, is one such pedagogy that provides opportunities for students to develop and improve specific process skills during science content learning. (Walker & Warfa, 2017). Building upon POGIL and the revisions made to AP Chemistry to focus more on the inquiry model, inquiry-based practices have been steadily integrated into science courses in the hopes of achieving those same increased learning outcomes.

A study found that students using the inquiry curriculum in a problem-based environmental health science course performed significantly better than those using the alternative curriculum in posing active inquiry questions and generating hypothesis-driven approaches to inquiry into their questions. (Kang et al., 2012) Among the students in the Kang study who were less prepared for inquiry in the beginning, 68% improved inquiry-questioning capability, while among those who were more prepared for inquiry, 36% improved in generating hypotheses-driven approaches. Students appear to do well even if they just "perceive" themselves as having had a good level of "experience" in inquiry-based practices (Patke, 2013).

Moreover, data collected from a survey on students' self-reported sentiment showed that students overwhelmingly felt that inquiry-based learning helped them grasp and remember the content, as well as develop competencies in data analysis and critical thinking. (Tamari & Shun Ho, 2019). Mastery of content though, may not always be explicit in student grades because it can also translate to other things (Wright, 2005) and with less time to synthesize and analyze information on a typical standardized test (Setiawan et al., 2019), it becomes more difficult to ascertain whether students are becoming better versed in these soft skills or merely regurgitating content come assessment time. But even this uncertainty has not curtailed the use of standardized testing to measure more than student mastery of content.

Standardized Testing and Science Assessments

The use of standardized tests has long been a staple of educational institutions everywhere and the NAEP science assessment is only one of many such tests that American students are required to undergo in their educational career. While a robust debate has been raging over the efficacy and impact of having a standardized-test saturated environment in K-12, an important discussion is simultaneously occurring as far as the ability of standardized tests to accurately account not just for content knowledge but also the soft skills gained from learning science through inquiry-based methods. When the University of Michigan collected data from nearly 8,000 students who participated in inquiry-based and technology-infused curriculum units in science courses throughout their district, it showed statistically significant increases on curriculum test scores for each year of participation. (Marx et al., 2004) When the curriculum was carefully developed and aligned with professional development and district policies, as that study had done, the findings that showed students who historically are low achievers in science can succeed in standard-based inquiry science, became evident. (Marx et al., 2004)

Teaching strategies that actively engage students in the learning process through scientific investigations are more likely to increase conceptual understanding than are strategies that rely on more passive techniques, which are often necessary in the current standardized-assessment laden educational environment. (Minner, Levy & Century, 2010). However, few studies have been done that specifically links inquiry-based practices to test gains in standardized tests. Research from the 2000s suggested that inquiry-based learning as a whole contributed gains in standardized test scores from content assessment (Geier et al., 2008; Powell, n.d.; Capp, 2009) and also when the instruction included the corresponding laboratory investigations conducted in that manner (Turner & Rios, 2008). While results suggested that the use of an inquiry-based teaching style did not dramatically alter students' overall achievement in physical science as measured by the North Carolina standardized test, it did however, have other positive effects such as a dramatic improvement in student participation and higher classroom grades earned by students. (Tretter & Jones, 2003)

Standardized tests are often structured in multiple choice formats that make it easier to evaluate content mastery but not necessary the soft skills that often come with proficiency in inquiry practices. Having graphical implementations as part of standardized tests (Yeh & McTigue, 2009) allows the potential for inquiry-based skills to come across in the evaluations as well. This takes on increased significance since inquiry activities with component visualization tools have been increasingly prevalent in K−12 STEM classrooms. However, researchers have pointed how evidence of their efficacy has primarily been collected from controlled laboratory studies that lack ecological validity or from small-scale classroom interventions that assess learning outcomes proximal to the intervention. (Stieff, 2019). Stieff's (2019) assertion that visualization tools embedded in inquiry activities result not only in short-term gains but in long-term improvements in learning outcomes could enhance the potential for standardized testing as valid instruments for inquiry-based practices.

IBL Concerns: Standardized Testing and Beyond

Concerns about whether standardized tests can accurately depict student progress in acquiring inquiry-based learning skills appear only secondary to the challenges that IBL implementation sometimes poses to instructors and students alike. Research has shown that inquiry-based learning may not be the easiest task for teachers to implement. (Crawford, 2007; Capps & Crawford, 2013) So when research findings suggested that despite Grade 6, 7, and 8 teachers possessing confidence and heightened beliefs in their abilities to teach science and mathematics, but only admitting to implementing interactive hands-on learning about half the time in their classrooms, it is evident that there exists a disconnect between beliefs and implementation of inquiry-based practices. (DeCoito & Myszkal, 2018)

Critics of inquiry-based practices not only point to this apparent disconnect between beliefs and practices but emphasize that belief in its efficacy might be misguided. Kirchner, Sweller & Clark (2006) suggested that the inquiry approach ignores both the structures that constitute human cognitive architecture and evidence from empirical studies over the past half-century that consistently indicate that minimally guided instruction is less effective and less efficient than instructional approaches that place a strong emphasis on guidance of the student learning process. Their study reiterated that the advantage of guidance that proponents of IBL assert, begins to recede only when learners have sufficiently high prior knowledge to provide "internal" guidance, something that may prove difficult to achieve in a diverse classroom setting.

Methods

Since 1969, the National Center for Educational Statistics (NCES) has administered the National Assessment of Educational Progress (NAEP) Science Assessment as part of a nationwide assessment measuring student knowledge in various grades in earth science, life science, physical sciences and general science topics. True to its nature as a nationwide assessment, the NAEP is called upon as a tool to compare student achievement at the state, regional or district levels. This kind of assessment also allows for comparisons to be made between urban and rural areas, as well as public and private schools, among many possibilities. In 2019, the NAEP Science Assessment was given to 4th, 8th and 12th grade students, with 12th graders being the subject of the present study.

Participants and Sampling

Participants are chosen through a multistage probability sample design that enables NAEP to use a representative sample of the entire population while giving each participant an equal chance of being selected (National Center for Educational Statistics (NCES, 2022). The NAEP science assessments are conducted every two years, with the current assessment being conducted in early 2022 and the most recent results derived from assessments done in 2019. These tests were administered on tablet computers by 26,400 twelfth-grade students (NCES, 2022). This study reflects the science scores of twelfth-grade students in the nation's public schools.

Sample and Data Collection

A representative sample of the nation's students is created by NAEP through probability sampling. The 2019 science assessment utilized digital devices such as student tablets connected to a closed wireless network. Two back-to-back sessions were conducted at each of the participating schools, with approximately 25 students in each session. Prior to the test proper, students were asked to complete a tutorial that covers the proper use of tablets and system tools as well as how to ensure their responses are entered appropriately. Survey questions were also provided to students that pertain to their activities within and beyond the school as it relates to science. (NCES, 2022)Students complete a tutorial that helps them understand how to use the tablets and system tools, as well as how to enter their responses. Survey questionnaires are also administered on tablets to students that record information about their learning experiences.

The NAEP science assessment had a score range of 0-500, and students were allowed 120 minutes to complete the test.

Test data for the twelfth-grade science assessment consists of scores in three categories, (1) physical science, (2) life science, and (3) earth and space sciences (NCES, 2022). A composite score, which the NCES uses as a weighted combination of these categories, is also available and is the score used in this study.

School Selection and Year

The U.S. Department of Education's public school system database identifies the public schools that the NCES selects from to administer the NAEP. Private and parochial schools, along with other non-public schools are also selected to be considered for sampling. Once selected, these schools are further categorized by their urban or rural location for example, and then further according to the racial, ethnic or other demographic factors comprising the school. In determining state assessments, these same categories are sorted by achievement level to ensure a representative sample is chosen. The NCES makes a selection of schools from each of these categories with probability relative to the population of schools. This is to ensure that schools with high minority populations for example, along with smaller schools and private schools, are given appropriate representation. The respective departments of education for each state confirm the eligibility of any given school to participate in that year's NAEP with a final list of selected schools.

Data Analysis

The NAEP Data Explorer is the tool used to analyze the assessment data through its capabilities in creating descriptive tables and calculating *t*-tests to determine possible significance in the difference between the means of groups. These tables and tests were used to analyze the data from the 2019 twelfth-grade science composite scores for national public schools. Cohen's *d* effect size was calculated using the University of Colorado's "Effect Size Calculator" (Becker, 2000). Cohen's *d* is used to compare two means in order to determine the difference between them in terms of standard deviations (Cohen, 1992). The strength and importance of any significance that is found in the data can be more closely examined through this process.

NAEP Data Explorer

The NAEP Data Explorer allows data to be examined by year of assessment, subject of assessment and subsequently categorized and filtered into a variety of groups based on student, teacher, or school factors. Its data analysis and statistical analysis functions are available in the Data Explorer tool for investigating any potential relationships in the data. For this study, the following coded questions were selected for further exploration through NAEP Data Explorer:

- In this school year, how often have you come up with research questions to explore how something works (student-reported)? ID: K824201 (Multiple Answers)
- In this school year, how often have you used evidence from experiments to explain why something happens (student-reported)? ID: K824206 (Multiple Answers)
- In this school year, how often have you used tables or graphs to identify relationships between variables (student-reported)? ID: K824204 (Multiple Answers)
- In this school year, how often have you used information to disagree with someone about a scientific idea (student-reported)? ID: K824207 (Multiple Answers)

Results

This section will report the results of examining the impact of the use of inquiry-based learning practices on the NAEP 2019 twelfth-grade science scores for students in national-public schools. These results include the means and standard deviations for each variable pertaining to the research questions, as well as independent *t*-test results to determine significance. Cohen's *d* effect size was calculated to further examine any significance found. Each Cohen's *d* effect size was calculated using the University of Colorado's Effect Size Calculator (Becker, 2000). The results pertaining to each research question (RQ) are presented here individually.

Data Analysis

Table 1 presents the average scale score in 2019 for the twelfth-grade science assessment at the National (nationwide) level. NAEP's Data Explorer does not include the number of students (N).

Table 1. National Average Scale Score – Twelfth Grade Science

All students

Year	Jurisdiction	Average scale score	Standard deviation
2019	National public	149	37

The average scale score for science for all twelfth-graders was 149, with a standard deviation of 37.
NOTE: The NAEP Science assessment scale ranges from 0 to 300.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Science Assessment.

RQ #1: How often have you come up with research questions to explore how something works?

Table 2 shows the 2019 science scale score and standard deviation for twelfth-grade national-public school students based on their reported frequency of coming up with research questions to explore how something works.

Table 2. Average scale scores and standard deviations for 12th-grade science, by frequency of comingupwith research questions to explore how something works [K824201]

Year	Jurisdiction	How often came up with research questions to explore how something works	Average scale score	Standard deviation
2019	National public	Never or hardly ever	155	37
		Once in a while	159	37
		Sometimes	154	39
		Often	161	38
		Always/almost always	158	38

NOTE: Nonresponse for this variable was greater than 15 percent but not greater than 50 percent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Science Assessment.

Twelfth graders reporting coming up with research questions to explore how something works at the "never or hardly ever" level had an average scale score of 155, SD=37. The mean score for those reporting "once in a while" was 159 (SD=37). Students indicating that they come up with research questions for exploration "sometimes" had the highest average score at 154 (SD=39). The average scale score among students reporting a frequency of "often" with positing research questions was 161 (SD=38) and those that reported doing so "always/almost always" was 158 (SD=38).

The differences in means and independent t-test results for the frequency of students coming up with research questions to explore how something works are shown in Table 3.

Table 3. Difference in average scale scores between variables, for how often students came up with research questions to explore how something works [K824201]

	Never or hardly ever (155)	Once in a while (159)	Sometimes (154)	Often (161)	Always/almost always (158)
Never or hardly ever (155)					
Once in a while (159)	> Diff = 4 P-value = 0.0017 Family size = 10				
Sometimes (154)	x Diff = -1 P-value = 0.2552 Family size = 10	< Diff = -6 P-value = 0.0001 Family size = 10			
Often (161)	> Diff = 6 P-value = 0.0000	x Diff = 1 P-value = 0.4365	> Diff = 7 P-value = 0.0000		

	Never or hardly ever (155)	Once in a while (159)	Sometimes (154)	Often (161)	Always/almost always (158)
	Family size = 10	Family size = 10	Family size = 10		
Always/almost always (158)	x Diff = 3 P-value = 0.0397 Family size = 10	x Diff = -1 P-value = 0.3813 Family size = 10	> Diff = 4 P-value = 0.0042 Family size = 10	x Diff = -3 P-value = 0.0923 Family size = 10	
LEGEND:					
<	Significantly lower.				
>	Significantly higher.				
x	No significant difference.				

NOTE: Nonresponse for this variable was greater than 15 percent but not greater than 50 percent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Science Assessment.

Table 3 presents differences in means and independent t-test results. Alpha was set at 0.05 rather than 0.001 as set a priori by the researcher. The average scale score (M=159, SD=37) of students who report once in a while coming up with research questions to explore how something works was significantly (p=0.0017) higher than those who report that they never or hardly ever. Students who report having come up with research questions to explore how something works in their classes sometimeshad an average scale score (M=154, SD=39) significantly (p<0.001) lower than those who work with research questions once in a while. The average scale score (M=161, SD=38) of students who report coming up with research questions to explore how something works oftenis significantly (p<0.001) higher than those who report that they never or hardly ever come up with research questions or those that only do so sometimes. Students who report having come up with research questions to explore how something works in their classes alwaysor almost always had an average scale score (M=158, SD=38) significantly (p<0.001) higher than those who report only doing so sometimes.

Cohen's *d* effect size was calculated on the independent t-tests that indicated significance in order to determine its strength. Effect sizes measure the magnitude of the factor involved and can be categorized as 0.2 being small, 0.5 as medium, and 0.8 as large. Positive and negative Cohen's *d* values indicate improvement or deterioration in a predicted direction (Becker, 2000).

Table 4 shows Cohen's *d* effect size of the significant mean score differences when coming up with research questions to explore how something works.

Table 4. Effect sizes of significant mean score differences when coming up with research questions to explore how something works [K824201]

		Cohen's *d*
Once in a while	Never/hardly ever	0.11
Sometimes	Once in a while	-0.13
Often	Never/hardly ever	0.16
Sometimes	Never/hardly ever	-0.03
Always/almost always	Sometimes	-0.07

The effect size between students reporting that they had opportunities to come up with research questions to explore how something works once in a while and those who reported never or hardly ever coming up with research questions was 0.11. Between students who reported they came up with research questions to explore how something works sometimes and those reporting coming up with research questions once in a while, the effect size was -0.13. Students indicating that they had opportunities to come up with research questions to explore how something works often and those reporting that they never or hardly ever came up with research questions was

0.16. All other pairs of frequencies with a significant difference had a negligible effect size.

RQ #2: How often have you used evidence from experiments to explain why something happens?

Table 5 shows the 2019 science scale score and standard deviation for twelfth grade national-public school students based on their reported frequency of using evidence from experiments to explain why something happens.

Table 5. Average scale scores and standard deviations for grade 12 science, by how often used evidence from experiments to explain why something happens [K824206]

Year	Jurisdiction	How often used evidence from experiments to explain why something happens	Average scale score	Standard deviation
2019	National public	Never or hardly ever	145	35
		Once in a while	150	37
		Sometimes	150	38
		Often	163	37
		Always/almost always	168	36

NOTE: Nonresponse for this variable was greater than 15 percent but not greater than 50 percent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Science Assessment.

Students who declared that they never or hardly ever had the opportunity to use evidence from experiments to explain why something happens in science classes had an average scale score of 145 (SD=35). The average scale score for those indicating that they used evidence from experiments to explain why something happens once in a while was 150 (SD=37), the same average score for those that reported using evidence from experiments to explain why something happens sometimes. (SD=38) The average score for students who indicated having the opportunity to use evidence from experiments to explain why something happens often and always or almost always had an average scale score of 163 (SD=37) and 168 (SD=36) respectively.

The differences in means and independent *t*-test results for the frequency of reported use of evidence from experiments to explain why something happens is shown in Table 6.

Table 6. Difference in average scale scores between variables, for How often used evidence from experiments to explain why something happens [K824206]

	Never or hardly ever (145)	Once in a while (150)	Sometimes (150)	Often (163)	Always/almost always (168)
Never or hardly ever (145)					
Once in a while (150)	> Diff = 5 P-value = 0.0032 Family size = 10				
Sometimes (150)	> Diff = 4	x Diff = -1			

	Never or hardly ev (145)	Once in a while (150)	Sometimes (150)	Often (163)	Always/almo st always (168)
	P-value = 0.0024 Family size = 10	P-value = 0.6741 Family size = 10			
Often (163)	> Diff = 17 P-value = 0.0000 Family size = 10	> Diff = 12 P-value = 0.0000 Family size = 10	> Diff = 13 P-value = 0.0000 Family size = 10		
Always/almost always (168)	> Diff = 23 P-value = 0.0000 Family size = 10	> Diff = 18 P-value = 0.0000 Family size = 10	> Diff = 19 P-value = 0.0000 Family size = 10	> Diff = 6 P-value = 0.0000 Family size = 10	
LEGEND:					
<	Significantly lower.				
>	Significantly higher.				
x	No significant difference.				

NOTE: Nonresponse for this variable was greater than 15 percent but not greater than 50 percent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Science Assessment.

Students who reported using evidence from experiments to explain how something works once in a while in their science courses had an average scale score that was significantly ($p<.001$) higher than those who had never or hardly ever had the opportunity. Those that stated they sometimes had used evidence from experiments to explain how something works had an average scale score significantly ($p<.001$) higher than those who never or hardly ever did so. The average scale score of students who indicated using evidence to explain how something works often was significantly ($p<.001$) higher than those who reported never or hardly ever doing so and those that only did so once in a while as well as sometimes. Students who indicated they used evidence from experiments to explain how something works always or almost always had significantly ($p<.001$) higher average scale scores than all other frequencies.

Table 7 shows Cohen's *d* effect size of the significant mean score differences when using evidence to explain why something happens.

Table 7. Effect sizes of Significant Mean Score Differences when using evidence to explain why something happens [K824206]

		Cohen's *d*
Once in a while	Never/hardly ever	0.14
Sometimes	Never/hardly ever	0.14
Often	Never/hardly ever	0.50
Often	Once in a while	0.35
Often	Sometimes	0.35
Always/almost always	Never/hardly ever	0.65
Always/almost always	Once in a while	0.49
Always/almost always	Sometimes	0.49
Always/almost always	Often	0.14

The effect size between students reporting that they used evidence to explain why something happens once in a while as well as sometimes, and those who reported never or hardly ever using evidence for explanations was 0.14 and 0.14 respectively. Between students who reported having used evidence to explain why something happens often and those reporting never or hardly ever using evidence for explanations, a Cohen's *d* effect size of 0.50 was produced. Students indicating that they used evidence to explain why something happens often and those reporting doing so once in a while as well as sometimes, both produced a Cohen's *d* effect size of 0.35. The Cohen's *d* effect sizes between those indicating that they always or almost always used evidence to explain why something happens and those who reported using evidence for explanations never or hardly ever, once in a while, sometimes or often were 0.65, 0.49, 0.49 and 0.14, respectively.

RQ #3: How often do you use tables and graphs to identify relationships between variables?

Table 8 shows the 2019 science scale score and standard deviation for twelfth-grade national-public school students based on their reported frequency of using tables and graphs to identify relationships between variables.

Table 8. Average scale scores and standard deviations for grade 12 science, by how often tables and graphs are used to identify relationships between variables [K824204]

Year	Jurisdiction	How often used tables or graphs to identify relationships between variables	Average scale score	Standard deviation
2019	National public	Never or hardly ever	147	34
		Once in a while	150	36
		Sometimes	152	38
		Often	165	37
		Always/almost always	168	38

NOTE: Nonresponse for this variable was greater than 15 percent but not greater than 50 percent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Science Assessment.

Students who declared that they never or hardly ever had the opportunity to use tables or graphs to identify relationships between variables had an average scale score of 147 (SD=34). The average scale score for those indicating that they used tables or graphs to identify relationships between variables once in a while was 150 (SD=36). Those that reported using tables or graphs to identify relationships between variables sometimes had an average scale score of 152. (SD=38) Students who indicated having the opportunity to use tables or graphs to identify relationships between variables often and always or almost always had an average scale score of 165 (SD=37) and 168 (SD=38) respectively.

The differences in means and independent *t*-test results for the frequency of reported use of tables or graphs to identify relationships between variables is shown in Table 9.

Table 9. Difference in average scale scores between variables, for How often tables or graphs are used to identify relationships between variables [K824204]

	Never or hardly ever (147)	Once in a while (150)	Sometimes (152)	Often (165)	Always/almost always (168)
Never or hardly ever (147)					
Once in a while >					

	Never or hardly ever (147)	Once in a while (150)	Sometimes (152)	Often (165)	Always/almost always (168)
(150)	Diff = 3 P-value = 0.0304 Family size = 10				
Sometimes (152)	> Diff = 5 P-value = 0.0000 Family size = 10	x Diff = 2 P-value = 0.0998 Family size = 10			
Often (165)	> Diff = 18 P-value = 0.0000 Family size = 10	> Diff = 15 P-value = 0.0000 Family size = 10	> Diff = 13 P-value = 0.0000 Family size = 10		
Always/almost always (168)	> Diff = 22 P-value = 0.0000 Family size = 10	> Diff = 19 P-value = 0.0000 Family size = 10	> Diff = 17 P-value = 0.0000 Family size = 10	> Diff = 4 P-value = 0.0111 Family size = 10	
LEGEND:					
<	Significantly lower.				
>	Significantly higher.				
x	No significant difference.				

NOTE: Nonresponse for this variable was greater than 15 percent but not greater than 50 percent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Science Assessment.

Students who reported using tables and charts to identify relationships between variables once in a while in their science courses had an average scale score that was significantly (p<.001) higher than those who had never or hardly ever had the opportunity. Those that stated they sometimes had used tables and charts to identify relationships between variables had an average scale score significantly (p<.001) higher than those who never or hardly ever did so. The average scale score of students who indicated using tables and charts to identify relationships between variables often was significantly (p<.001) higher than those who reported never or hardly ever doing so and those that only did so once in a while as well as sometimes. Students who indicated they used tables and charts to identify relationships between variables always or almost always had significantly (p<.001) higher average scale scores than all other frequencies.

Table 10 shows Cohen's *d* effect size of the significant mean score differences when using tables and charts to identify relationships between variables

Table 10. Effect sizes of Significant Mean Score Differences when using tables and charts to identify relationships between variables [K824204]

		Cohen's *d*
Once in a while	Never/hardly ever	0.09
Sometimes	Never/hardly ever	0.09
Often	Never/hardly ever	0.51
Often	Once in a while	0.41
Often	Sometimes	0.35
Always/almost always	Never/hardly ever	0.58
Always/almost always	Once in a while	0.49
Always/almost always	Sometimes	0.42
Always/almost always	Often	0.08

Students indicating that they used tables and charts to identify relationships between variables once in a while or sometimes and those who reported they never or hardly ever used tables and charts for identifying relationships between variables produced a Cohen's *d* effect size of 0.09 in both instances. The effect sizes between students reporting that they often used tables and charts to identify relationships between variables and those that never or hardly ever do so was 0.51. The effect sizes between students reporting that they often used tables and charts to identify relationships between variables and those reporting that they did so once in a while or sometimes were 0.41 and 0.35, respectively. The Cohen's *d* effect sizes between those indicating that they used tables and charts to identify relationships between variables always or almost always, and those who reported using tables and charts to identify relationships between variables never or hardly ever, once in a while, sometimes or often were 0.58, 0.49, 0.42 and 0.08, respectively.

RQ4: How often do you use information to disagree with someone about a scientific idea?

Table 11 shows the 2019 science scale score and standard deviation for twelfth-grade national-public school students based on their reported frequency of using information to disagree with someone about a scientific idea.

Table 11. Average scale scores and standard deviations for grade 12 science, by how often information is used to disagree with someone about a scientific idea [K824207]

Year	Jurisdiction	How often used information to disagree with someone about a scientific idea	Average scale score	Standard deviation
2019	National public	Never or hardly ever	157	36
		Once in a while	162	38
		Sometimes	152	40
		Often	158	38
		Always/almost always	157	37

NOTE: Nonresponse for this variable was greater than 15 percent but not greater than 50 percent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Science Assessment.

Students who declared that they never or hardly ever had the opportunity to use information to disagree with someone about a scientific idea had an average scale score of 157 (SD=36). The average scale score for those indicating that they used information to disagree with someone about a scientific idea once in a while was 162 (SD=38). Those that reported using information to disagree with someone about a scientific idea sometimes had an average scale score of 152. (SD=40) Students who indicated having the opportunity to use information to disagree with someone about a scientific idea often and always or almost always had an average scale score of 158 (SD=38) and 157 (SD=37) respectively.

The differences in means and independent *t*-test results for the frequency of reported use of information to disagree with someone about a scientific idea is shown in Table 12.

Table 12. Difference in average scale scores between variables, for how often used information to disagree with someone about a scientific idea [K824207]

	Never or hardly ever (157)	Once in a while (162)	Sometimes (152)	Often (158)	Always/almost always (157)
Never or hardly ever (157)					
Once in a while (162)	> Diff = 5 P-value = 0.0008 Family size = 10				
Sometimes (152)	< Diff = -6 P-value = 0.0000 Family size = 10	< Diff = -10 P-value = 0.0000 Family size = 10			
Often (158)	x Diff = 0 P-value = 0.7527 Family size = 10	< Diff = -4 P-value = 0.0050 Family size = 10	> Diff = 6 P-value = 0.0000 Family size = 10		
Always/almost always (157)	x Diff = -1 P-value = 0.7198 Family size = 10	< Diff = -5 P-value = 0.0018 Family size = 10	> Diff = 5 P-value = 0.0008 Family size = 10	x Diff = -1 P-value = 0.5569 Family size = 10	
LEGEND:					
<	Significantly lower.				
>	Significantly higher.				
x	No significant difference.				

NOTE: Nonresponse for this variable was greater than 15 percent but not greater than 50 percent.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2019 Science Assessment.

Students who reported using information to disagree with someone about a scientific idea once in a while in their science courses had an average scale score that was significantly ($p<.001$) higher than those who had never or hardly ever had the opportunity. Those that stated they sometimes had used information to disagree with someone about a scientific idea had an average scale score significantly ($p<.001$) lower than those who never or hardly ever did so, as well as those who did so once in a while. The average scale score of students who indicated using information to disagree with someone about a scientific idea often was significantly ($p<.001$) lower than those who reported once in a while and significantly higher ($p<.001$) than those who did so sometimes. Students who indicated they used information to disagree with someone about a scientific idea always or almost always had significantly ($p<.001$) lower average scale scores than those who did so once in a while and significantly ($p<.001$) higher average scale scores than those who did so sometimes.

Table 13 shows Cohen's *d* effect size of the significant mean score differences when using information to disagree with someone about a scientific idea.

Table 13. Effect sizes of Significant Mean Score Differences when using information to disagree with someone about a scientific idea [K824207]

		Cohen's *d*
Once in a while	Never/hardly ever	0.14
Sometimes	Never/hardly ever	-0.13
Sometimes	Once in a while	-0.26
Often	Once in a while	-0.11
Often	Sometimes	0.15
Always/almost always	Once in a while	-0.13
Always/almost always	Sometimes	0.13

The effect size between students reporting that they used information to disagree with someone about a scientific idea sometimes and those who reported using information to disagree with someone about a scientific idea once in a while was -0.26. All other pairs of frequencies with a significant difference had a small effect size.

Discussion

This study was undertaken to examine the impact of inquiry-based practices on the average scale scores on the 2019 twelfth-grade NAEP science assessment. Specifically, the study examined the impact of how frequently students came up with research questions to explore how something works, how often students used experimental evidence to explain why something happens, how frequently tables and graphs were used to identify relationships between variables as well as how often students use information to disagree with someone about a scientific idea. This section looks at the findings pertaining to these inquiry-based learning practices that were reported in the results section.

This study affirms previous studies that suggested the use of inquiry-based learning practices in science courses improved learning outcomes, with students who reported having had opportunities to perform these tasks, scoring above the national average scale score. The exceptions lie with students who reported never or hardly ever using evidence from explain why something happens, as well as those who have never or hardly ever used tables and graphs to identify relationships between variables. These students scored slightly below the national average scale score. This coincides with previous studies done by Geier et al (2008) and Powell (2010) that looked at how an inquiry-based science curriculum can indeed lead to gains in standardized test scores. However, the various inquiry-based practices that were examined in this study have impacted science assessment scores to varying extents.

RQ #1: Coming up with research questions to explore how something works

While students scored higher than average on the twelfth-grade science assessment regardless of how frequently they were able to come up with research questions to explore how something works, it had no impact on their performance on the science assessment.

Tretter (2003) raised a conclusion from a 1996 NAEP science report that mentioned how "research on the relationship between exposure to hands-on science tasks and overall science performance is sparse and inconclusive," and it appears to be the case here as far as generation of research questions is concerned. While this is in line with previous findings that showed increases in either the questioning or approach abilities among less or more prepared students as evidence of the impact of an inquiry-based curriculum (Kang et al., 2012), it has also been reported that the ability to ask questions may not be as reflective of students' content knowledge or lack thereof (Scardamalia & Bereiter, 1992).

The finding that the frequency of students to come up with research questions had no impact on their performance on the science assessment is also in line with other studies suggesting that while the nature of student questions serves as "an indication of student understanding of the content as well as thinking skills" (Kang et al., 2012), the frequency of their questioning was also not related to their learning (Graesser & Person, 1994).

The impact of students' ability to pose research questions to explore how something works may not be as explicit considering the format of the NAEP science assessment follows that of most standardized tests. While inquiry-based teaching practices have been previously shown to trigger a significant gain in test scores (Powell, 2010; Capp, 2009), it has often been left up to the instructor to assess a student's ability to understand the idea by asking them to recognize the concept in a series of questions. (Capp, 2009). The Capp study posited that students made connections between a science process and the kind of questions asked on state-mandated tests with the hope that they will better recognize what kind of question they are being asked and be more likely to answer with well-chosen vocabulary that shows their understanding of the question.Tretter (2003) pointed to how the skills and concepts learned during inquiry instruction are not as easily testable in a multiple-choice format. The implicit impact of the frequency of coming up with research questions may very well be hidden within the strict confines of a standardized test formatted in a way that does not explicitly test a student's ability to do so.

RQ #2: Using evidence to explain why something happens

Students who reported having used evidence to explain why something happens often as well as always or almost always, scored higher than the average scale score on the twelfth-grade science assessment compared to their counterparts who only did so once in a while or sometimes. Those who reported never or hardly ever having the opportunity to practice the skill scored slightly below the average scale score.

Part of what makes a person scientifically literate, according to the PISA 2015 report, is having the competence to explain phenomena, evaluate and design investigations, and interpret data and evidence scientifically. The findings in this study confirm what previous studies have found on whether students who frequently had the opportunity to use evidence in their explanations of phenomena have an edge on scientific literacy as evidenced by increased learning outcomes and gains on standardized test scores.

The biggest effect sizes were produced among students who reported having used evidence to explain why something happens often as well as always/almost always and those that had never or hardly ever done so. Substantive effect sizes were similarly reported among students who indicated they used evidence to explain various phenomena often and those who did so once and a while and sometimes. The big effect sizes were also evident among students who reported always or almost always using evidence to explain why something happens and those who did so once and a while and sometimes.

The findings are consistent with those found by Kang et al. (2012) that suggested that the impact of inquiry-based curriculum was profound among students deemed "less-prepared," as far as making the greatest improvements in terms of the basic inquiry skill. The impact observed among students who practice this inquiry skill set with greater frequency is also consistent with the possible explanation in the Kang study. They explained how the inquiry-based curriculum in their study still provided opportunities for the "more-prepared" students to continue to develop inquiry capacity. Kang et al. added that the significant increase in students who are already competent in the skill even before the implementation of the inquiry-based curriculum were able to make the connections necessary to further their understanding. For these students, they added, this meant an improvement in the knowledge about and/or capabilities to conduct scientific inquiry "at a more complex level."

RQ #3: Using tables and graphs to identify relationships between variables

A similar trend could be observed in the findings for frequency of use of tables and graphs to identify relationships between variables. Students who reported having used such visual aids to identify relationships between variables often as well as always or almost always, scored higher than the average scale score on the twelfth-grade science assessment compared to their counterparts who only did so once in a while or sometimes. Similar to the findings for the previous research question, those who reported never or hardly ever having the opportunity to practice the skill scored slightly below the average scale score.

The findings for this research question arguably allows the clearest link between how frequently tables and graphs are used to identify relationships between variables and students' average scale score for the twelfth-grade science assessment. While Tretter (2003) previously pointed out that not all skills and concepts learned in an inquiry-based curriculum are as easily testable in a multiple-choice format as implemented in the NAEP assessment, the questions used in most science standardized tests do employ tables, charts and graphs. A Texas A&M study looked at 985 test items from 14 state-implemented standardized tests and found that 52.7% of the test items included graphical representations (Yeh & McTigue, 2009) and that most of these graphics (79.5%) contained information that was essential for correctly answering the questions. Since charts & graphs were most likely to carry all the needed information and require tasks of reorganization and transformation to answer the questions, Yeh & McTigue posit that findings such as this are not surprising when one considers the fundamental construct of charts or diagrams and pictorial illustrations as instruments to condense information in an organizational manner. This proves useful in linking proficiency in the use of tables and graphs to identify relationships between variables to significant standardized test gains in the short-term and increased learning outcomes in the long-term.

Similar to the findings in the previous research question, the biggest effect sizes were observed among students who reported using tables and graphs to identify relationships between variables often and always or almost always and those that never or hardly ever had the opportunity to do so. Likewise, moderate effect sizes were evident among students who reported using tables and graphs once in a while and those that did so often and always or almost always. The same moderate effect size was also observed among students that reported using tables and charts sometimes and those that claimed to do so always or almost always.

In addition to previous research that suggested the efficacy of inquiry-based practices extending to both less-prepared and more-prepared students, one possible explanation lies in the nature of standardized tests itself. Arguably, the skills learned in inquiry-based laboratory investigations, the data gathering, and analysis required to do so, as well as the exposure to the tables and charts necessary to communicate such, are more directly "translatable" to a standardized test setting than other inquiry-based practices. Previous studies (Turner & Rios, 2008) pointed to high school students demonstrating increased academic performance on standardized tests when biology instruction includes inquiry-based laboratory investigations, as well as providing evidence of a direct correlation with students receiving inquiry-based labs in science classes and their standardized test performance (Geier, 2008). Even "perceived" level of experience in inquiry-based laboratory investigation classes showed an increase in standardized test scores, revealing a more multilayered nature to the efficacy of inquiry-based practices (Patke, 2013).

RQ #4: Using information to disagree with someone about a scientific idea

Regardless of how frequently students reported having the opportunity to use information to disagree with someone about a scientific idea, their average scale scores on the NAEP science assessment were above the national average. However, this had no impact on their performance on the science assessment.

While engagement in science-related issues and ideas constitute the basis for scientific literacy, it is an aspect of a student's intellectual development in science that is not directly tested in standardized tests like that of NAEP. Previous studies have examined how inquiry-based curricula such as POGIL emphasize developing not only process skills but so-called soft skills such as the above, and how this often comes at the expense of covering content (Walker & Warfa, 2017). One possible explanation is the absence of a free-response section in the NAEP science assessment that could potentially measure student understanding of science-related issues and quantify proficiency in the so-called soft skills. If previous studies have argued that the mastery of content "naturally emerges as students seek out, evaluate, and organize the information they need to develop an informed understanding about an issue," (Wright, 2005) then not having a free response section in the NAEP assessment deprives learners of an outlet to demonstrate their abilities in that realm.

Likewise, it has been demonstrated that greater improvements in students' science literacy and research skills can become apparent when using inquiry lab instruction andthatinquiry students do gained self-confidence in their scientific abilities. (Brickman, et al., 2009) With preparations for standardized testing typically involving students doing more rote memorization, the argument that they have less time to synthesize information or apply knowledge takes further hold (Setiawan, 2019). Without the practical means to demonstrate this self-confidence in using information to disagree with someone about a scientific idea, it becomes nearly impossible to ascertain its impact.

Conclusion and Implications

Potential implicit impact of inquiry-based practices

Inquiry-based learning practices have long been proven to be efficacious in a number of aspects, including increasing learner outcomes. But few studies have suggested that these benefits translate to better performance on standardized tests. This study that involved the 2019 NAEP science assessment, showed that the impact of certain inquiry-based practices on student test scores have remained, as described in a 1996 NAEP report, "sparse and inconclusive." While there are substantial effect sizes observed when analyzing the frequency of certain IBL practices such as using evidence to explain why something happens as well as using tables and graphs to identify relationships between variables, the impact of having students come up with research questions to explore how something works as well as using information to disagree with someone about a scientific idea, had negligible to no impact on their standardized test scores. Arguably, with scores in the NAEP science assessment that are mostly above the national average, it can be surmised that the impact of certain inquiry practices statistically deemed negligible or non-existent, may very well be implicit. Hidden beneath the rigors of a highly structured standardized test, the advantages that these practices afford the learner are such that it is often not reflected in the data. While the test questions can easily assess a student's ability to use evidence for explanatory purposes as well as to use tables and graphs to identify relationships, evaluating proficiency in "soft skills" such as coming up with research questions as well as articulating information to debate on a scientific issue is impossible given the structural constraints that a multiple-choice test presents.

Limits of standardized testing's efficacy to assess IBL practices

This study adds to the growing body of literature that affirms the efficacy of integrating inquiry-based practices into the science curriculum, along with the improved learning outcomes associated with it. But given that the skills and concepts learned during inquiry instruction are not as easily testable in multiple-choice formatted tests (Tretter, 2003) such as the NAEP science assessment, this study also raises questions as to the limits and extent to which standardized tests can be valid instruments in assessing the soft skills that form the core of inquiry-based learning practices.

Implications

This study presents broad implications for the future of inquiry-based learning amidst an ongoing shift in the way science courses are taught that takes advantage of the affordances such practices instill. While it confirmed the efficacious nature of inquiry-based practices as far as improving learning outcomes for students that had opportunities to learn it, the study also challenged the predictive nature of standardized tests such as the NAEP science assessment as far as proficiency in these soft skills are concerned. There are also more specific implications on the time devoted to inquiry-based curricula in science courses and the perceptions of the instructors tasked with implementing it.

Considering the time-intensive nature of implementing an inquiry-based curriculum, (Walker & Warfa, 2017) and perhaps the desire to satisfy their students' natural curiosity to figure things out for themselves (rather than perform prefabricated investigations) for greater engagement, instructors are left to ponder a consequential pedagogical question. Do they fully embrace an inquiry-based learning pedagogy at the expense of class time that could be otherwise be devoted to covering content? This carries significant implications on teacher motivation given previous studies that showed how teaching inquiry has been a challenging task for instructors. (Crawford, 2007; Capps and Crawford, 2013). If increased learning outcomes inevitably follow, manifested either through stronger academic performance or gains in standardized test scores, the amount of time and effort devoted to emphasizing inquiry-based practices will require further rethinking.

While the study has given much to reflect on the appropriateness of standardized tests to assess student aptitude in inquiry practices, there are greater implications that arise from the increased pedagogical focus that inquiry-based learning has triggered. POGIL has emerged as a pedagogy that provides opportunities for improving process skills during content learning through guided-inquiry activities. Coupled with recent revisions made to the AP Chemistry curriculum that reflects a more inquiry-based teaching approach, there is much reason to re-examine the practicality of this apparent universal embrace of inquiry-based learning. Given the results of the study that pointed to equally substantial effect sizes for students that are familiar and well-versed in inquiry practices as well as those who are less prepared, it certainly calls into question whether proponents of IBL are justified in their advocacy for this paradigm shift especially when the same outcomes are evident even if an instructor's focus on inquiry practices is not as intense as otherwise expected.

Limitations

While this study confirmed some trends regarding the efficacy of inquiry-based practices as far as its impact on student standardized test scores, there are nonetheless limitations to consider. The nature of the data used in the study presents one such limitation. The data gathered from the 2019 NAEP science assessment is considered secondary data and as such, carries any possible validity problems that occurred during the data collection process. Replication of any correlation analysis in the study could be impacted by the variables having been pre-decided. The data should not be interpreted as having a cause-and-effect relationship in this regard. Likewise, the analysis methods used were provided through the NAEP Data Explorer and were limited in scope.

Sample size also posed a limitation in this study as the exact sample size used by NAEP cannot be ascertained for confidentiality reasons. Having the sample size fall within a range as opposed to an exact number hinders the validity of the research findings. Not having an exact sample size also poses a challenge to fully ascertain the impact of certain variables with nonresponse rates that also fall within a range.

Recommendations for Future Research

The impact that certain inquiry-based learning practices had on standardized test scores that were evident in this study of the 2019 NAEP science assessment, can potentially shed light not just on the perceived efficacy of inquiry-based learning but also on the nature of standardized testing in STEM.Further research will ascertain whether the benefits that inquiry-based practices have been perceived to afford learners, can indeed be manifested through standardized tests like that of NAEP. Knowing more about the link between the nature of student questioning and the degree to which they are practicing inquiry at its various levels or their level of aptitude in it, will greatly assist not just science educators but also curriculum developers and standardized test administrators in identifying the kinds of questions that would more accurately reflect competence in inquiry and not merely content mastery.

It would also be incumbent upon future researchers to continue exploring whether inquiry-based learning could be implemented as part of a traditional science curriculum without compromising the time required to cover foundational knowledge in content heavy courses. Establishing the parameters for a more hybridized approach could perhaps allow for certain inquiry-based practices to be integrated into the curriculum while including a more robust analysis of issues to further improve science literacy. A more data-driven approach towards a seamless integration of inquiry-based practices in science courses but not necessarily utilizing standardized testing, will greatly assist educators in delivering curriculum that best suits the learner's needs while remaining relevant to the evolving nature of understanding science and technology.

REFERENCES

1. Adeoye, I. F. (2020). Senior Secondary School Students' Attitude to Inquiry-based learning in Chemistry. *International Journal of Scientific and Research Publications (IJSRP)*, *10*(4), p10098. https://doi.org/10.29322/IJSRP.10.04.2020.p10098
2. Baseya, J. M., & Francis, C. D. (2011). Design of inquiry-oriented science labs: Impacts on students' attitudes. *Research in Science & Technological Education*, *29*(3), 241–255. https://doi.org/10.1080/02635143.2011.589379
3. Becker, L. (2000). Effect Size Calculators. University of Colorado, Colorado Springs. Retrieved from https://www.uccs.edu/lbecker/
4. Bell, R. L., Smetana, L., & Binns, I. (2005). Assessing the inquiry level of classroom activities. *The Science Teacher*, 4.
5. Ben Wu, X., Sandoval, C., Knight, S., Xavier, J., Macik, M., & Schielack, J. F. (2021). Web-based authentic inquiry experiences in large introductory classes consistently associated with significant learning gains for

all students. *International Journal of STEM Education*, *8*(1). http://dx.doi.org.cmich.idm.oclc.org/10.1186/s40594-021-00290-3

6. Bielik, T., & Yarden, A. (2016). Promoting the asking of research questions in a high-school biotechnology inquiry-oriented program. *International Journal of STEM Education*, *3*(1), 1–13. http://dx.doi.org.cmich.idm.oclc.org/10.1186/s40594-016-0048-x
7. Brickman, P., Gormally, C., Armstrong, N., & Hallar, B. (2009). Effects of Inquiry-based Learning on Students' Science Literacy Skills and Confidence. *International Journal for the Scholarship of Teaching & Learning*, *3*(2), 1–22.
8. Buchanan, S., Harlan, M. A., Bruce, C., & Edwards, S. (2016). Inquiry Based Learning Models, Information Literacy, and Student Engagement: A literature review. *School Libraries Worldwide*, *22*(2), 23–39.
9. Buck, L. B., Bretz, S. L., & Towns, M. H. (2008). Characterizing the Level of Inquiry in the Undergraduate Laboratory. *Journal of College Science Teaching*, *38*(1), 52–58.
10. Capp, R. (2009). Process Skills Practice and Standardized Tests. *Science and Children*, *46*(5), 28–30.
11. Capps, D. K., & Crawford, B. A. (2013). Inquiry-Based Professional Development: What does it take to support teachers in learning about inquiry and nature of science? *International Journal of Science Education*, *35*(12), 1947–1978. https://doi.org/10.1080/09500693.2012.760209
12. Cohen, J. (1992). Statistical power analysis. Current Directions in Psychological Science, 1(3), 98-101. Retrieved from https://www.jstor.org/stable/20182143
13. Cooper, Bailey, B., Briggs, K., & Holliday, J. (2017). Assessing Student Openness to Inquiry-Based Learning in Precalculus. *PRIMUS : Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *27*(7), 736–753. https://doi.org/10.1080/10511970.2016.1183155
14. Crawford, B. A. (2007). Learning to teach science as inquiry in the rough and tumble of practice. *Journal of Research in Science Teaching*, *44*(4), 613–642. https://doi.org/10.1002/tea.20157
15. DeCoito, I., & Myszkal, P. (2018). Connecting Science Instruction and Teachers' Self-Efficacy and Beliefs in STEM Education. *Journal of Science Teacher Education*, *29*(6), 485–503. https://doi.org/10.1080/1046560X.2018.1473748
16. Demircioğlu, T., & Uçar, S. (2012). The Effect of Argument-Driven Inquiry on Pre-Service Science Teachers' Attitudes and Argumentation Skills. *Procedia - Social and Behavioral Sciences*, *46*, 5035–5039. https://doi.org/10.1016/j.sbspro.2012.06.382
17. Dewey, J. (2007). *Experience and education*. Touchstone.
18. Fay, M. E., Grove, N. P., Towns, M. H., & Bretz, S. L. (2007). A rubric to characterize inquiry in the undergraduate chemistry laboratory. *Chemistry Education Research and Practice*, *8*(2), 212–219. https://doi.org/10.1039/B6RP90031C
19. Frezell, D. (2018). Impact of Inquiry Based Learning on Students' Motivation, Engagement and Attitude in Science [M.Ed., University of Windsor (Canada)]. In *ProQuest Dissertations and Theses*. http://www.proquest.com/docview/2015705081/abstract/3E623A7631964580PQ/1
20. Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and Quasi-Experimental Studies of Inquiry-Based Science Teaching: A Meta-Analysis. *Review of Educational Research*, *82*(3), 300–329.
21. Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, *45*(8), 922–939. https://doi.org/10.1002/tea.20248
22. Gehring, K. M., & Eastman, D. A. (2008). Information fluency for undergraduate biology majors: Applications of inquiry-based learning in a developmental biology course. *CBE Life Sciences Education*, *7*(1), 54–63. https://doi.org/10.1187/cbe.07-10-0091
23. Graesser, A. C., & Person, N. K. (1994). Question Asking during Tutoring. *American Educational Research Journal*, *31*(1), 104–137. https://doi.org/10.2307/1163269
24. Hartini, I., & Ferawati, F. (2016). Keefektifan Pembelajaran Inkuiri dan Problem-Based Learning terhadap Hasil Belajar Siswa Pada Mata Pelajaran Ilmu Pengetahuan Alam. Jurnal Inovasi Pendidikan Dasar, 2(1), 27–34.
25. Huda, H. B., & Rohaeti, E. (2021). Students' attitude profile towards chemistry based on Research-Oriented Collaborative Inquiry Learning Model (REORCILEA). *Journal of Physics: Conference Series*, *1806*(1). http://dx.doi.org.cmich.idm.oclc.org/10.1088/1742-6596/1806/1/012193
26. Isik-Ercan, Z. (2020). 'You have 25 kids playing around!': Learning to implement inquiry-based science learning in an urban second-grade classroom. *International Journal of Science Education*, *42*(3), 329–349. https://doi.org/10.1080/09500693.2019.1710874

27. Kang, N.-H., DeChenne, S. E., & Smith, G. (2012). Inquiry Learning of High School Students Through a Problem-Based Environmental Health Science Curriculum. *School Science and Mathematics*, *112*(3), 147–158. https://doi.org/10.1111/j.1949-8594.2011.00128.x

28. Kang, J., & Keinonen, T. (2017). The effect of inquiry-based learning experiences on adolescents' science-related career aspiration in the Finnish context. *International Journal of Science Education*, *39*(12), 1669–1689. https://doi.org/10.1080/09500693.2017.1350790
29. Khalaf, B. K., & Zin, Z. B. M. (2018). Traditional and Inquiry-Based Learning Pedagogy: A Systematic Critical Review. *International Journal of Instruction*, *11*(4), 545–564. https://doi.org/10.12973/iji.2018.11434a
30. Kırıcı, M. G., & Bakırcı, H. (2021). The effect of STEM supported research-inquiry-based learning approach on the scientific creativity of 7th grade students. *Journal of Pedagogical Research*, *5*(2), 19–35. https://doi.org/10.33902/JPR.2021067921
31. Locke, D. H., Callaghan, N. I., Dicks, S., Scherer, E., Shukalyuk, A. I., Jou, M., Kilkenny, D. M. (2020). Enhancing senior high school student engagement and academic performance using an inclusive and scalable inquiry-based program. *NPJ Science of Learning*, *5*(1). http://dx.doi.org.cmich.idm.oclc.org/10.1038/s41539-020-00076-2
32. Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R., & Tal, R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching*, *41*(10), 1063–1080. https://doi.org/10.1002/tea.20039
33. Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, *47*(4), 474–496. https://doi.org/10.1002/tea.20347
34. National Center for Educational Statistics. (2022). About the NAEP Science Assessment Retrieved February 28, 2022, from https://www.nationsreportcard.gov/science/about/digitally-based-assessment/?grade=12
35. National Research Council, Center for Science, M., and Engineering Education, & Committee on Development of an Addendum to the National Science Education Standards on Scientific Inquiry. (2000). *Inquiry and the National Science Education Standards: A Guide for Teaching and Learning*. National Academies Press. https://cmich.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=e000xna&AN=33528&site=ehost-live
36. OECD 2016 PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics and Financial Literacy (Paris: OECD Publishing)
37. Oymak, O., & Ogan-Bekiroglu, F. (2021). Comparison of Students' Learning and Attitudes in Physical versus Virtual Manipulatives using Inquiry-Based Instruction. *IAFOR Journal of Education*, *9*(4), 23–42.
38. Patke, U. (2013). Inquiry-based laboratory investigations and student performance on standardized tests in biological science [Ed.D., Walden University]. In *ProQuest Dissertations and Theses*. http://www.proquest.com/docview/1465367821/abstract/8A88D74BEF8B4E8FPQ/1
39. Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures*. Viking Press.
40. Powell, P. E. (n.d.). *The effect of instructional methodology on high school students natural sciences standardized tests scores* [Ed.D., Walden University]. Retrieved April 7, 2022, from http://www.proquest.com/docview/749786040/abstract/F1E11BE609B749F6PQ/1
41. Riegle-Crumb, C., Morton, K., Nguyen, U., & Dasgupta, N. (2019). Inquiry-Based Instruction in Science and Mathematics in Middle School Classrooms: Examining Its Association With Students' Attitudes by Gender and Race/Ethnicity. *AERA Open*, *5*(3), 2332858419867653. https://doi.org/10.1177/2332858419867653
42. Rohaeti, E., Prodjosantoso, A. K., & Irwanto. (2020). Research-Oriented Collaborative Inquiry Learning Model: Improving Students' Scientific Attitudes In General Chemistry. *Journal of Baltic Science Education*, *19*(1), 108–120. http://dx.doi.org.cmich.idm.oclc.org/10.33225/jbse/20.19.108
43. Rowland, S., Pedwell, R., Lawrie, G., Lovie-Toon, J., & Hung, Y. (2016). Do We Need to Design Course-Based Undergraduate Research Experiences for Authenticity? *CBE Life Sciences Education*, *15*(4), ar79. https://doi.org/10.1187/cbe.16-02-0102
44. Setiawan, H., Garnier, K., & Isnaeni, W. (2019). Rethinking standardized test of science education in Indonesian high school. *Journal of Physics: Conference Series*, *1321*(3). http://dx.doi.org.cmich.idm.oclc.org/10.1088/1742-6596/1321/3/032078

45. Scardamalia, M., & Bereiter, C. (1992). Text-Based and Knowledge-Based Questioning by Children. *Cognition and Instruction*, *9*(3), 177–199.
46. Seidel, S. B., & Tanner, K. D. (2013). "What if students revolt?"—Considering Student Resistance: Origins, Options, and Opportunities for Investigation. *CBE Life Sciences Education*, *12*(4), 586–595. https://doi.org/10.1187/cbe-13-09-0190
47. Şimşek, P., & Kabapınar, F. (2010). The effects of inquiry-based learning on elementary students' conceptual understanding of matter, scientific process skills and science attitudes. *Procedia - Social and Behavioral Sciences*, *2*(2), 1190–1194. https://doi.org/10.1016/j.sbspro.2010.03.170
48. Stieff, M. (2019). Improving Learning Outcomes in Secondary Chemistry with Visualization-Supported Inquiry Activities. *Journal of Chemical Education*, *96*(7), 1300–1307. https://doi.org/10.1021/acs.jchemed.9b00205
49. Tamari, F., & Shun Ho, I. (2019). Two-Year Community: Guided Inquiry-Based Activities Positively Impact Learning in Community College Courses. *Journal of College Science Teaching*, *048*(04). https://doi.org/10.2505/4/jcst19_048_04_12
50. Tretter, T. R., & Jones, M. G. (2003). Relationships Between Inquiry-Based Teaching and Physical Science Standardized Test Scores. *School Science and Mathematics*, *103*(7), 345–350. https://doi.org/10.1111/j.1949-8594.2003.tb18211.x
51. Turner, M. J., & Rios, J. M. (2008). Determining Paths to Success: Preparing Students for Experimental Design Questions on Standardized Tests. *The American Biology Teacher*, *70*(3), 140–152. https://doi.org/10.2307/30163227
52. Vygotskiĭ, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
53. Walker, L., & Warfa, A.-R. M. (2017). Process oriented guided inquiry learning (POGIL®) marginally effects student achievement measures but substantially increases the odds of passing a course. *PLOS ONE*, *12*(10), e0186203. https://doi.org/10.1371/journal.pone.0186203
54. Wheeler, L., & Bell, R. (2012). Open-Ended Inquiry: Practical ways of implementing inquiry in the chemistry classroom. *The Science Teacher*, *79*(6), 32–39.
55. Wolf, S. J., & Fraser, B. J. (2008). Learning Environment, Attitudes and Achievement among Middle-school Science Students Using Inquiry-based Laboratory Activities. *Research in Science Education*, *38*(3), 321–341. https://doi.org/10.1007/s11165-007-9052-y
56. Wright, R. L. (2005). Points of view: Content versus process: is this a fair choice? Undergraduate biology courses for nonscientists: toward a lived curriculum. *Cell Biology Education*, *4*(3), 189–196. https://doi.org/10.1187/cbe.05-04-0075
57. Yeh, Y.-F. Y., & McTigue, E. M. (2009). The Frequency, Variation, and Function of Graphical Representations within Standardized State Science Tests. *School Science and Mathematics*, *109*(8), 435–449. https://doi.org/10.1111/j.1949-8

