The Effectiveness Of Argument-Driven Inquiry (ADI) In Improving Students' Argumentation Skills In Science Education: A Meta-Analysis

Nor Hikmah[1*], Nadia Azizah[2], Luvia Ranggi Nastiti[3], Santiani[4]
[1-4]Physics Education Program, State Islamic University of Palangka Raya, Palangka Raya, Indonesia
*Email: norhikmah2311130014@uin-palangkaraya.ac.id

ABSTRACT

This study aimed to quantitatively examine the effectiveness of the Argument-Driven Inquiry (ADI) instructional model in enhancing students' scientific argumentation skills in science education and to identify moderating factors influencing its effectiveness. This meta-analysis initially identified 12 primary studies that met the inclusion criteria. However, one study with an extreme effect size was excluded from the main analysis based on sensitivity testing, resulting in a final synthesis of 11 studies published between 2010 and 2025. The studies were retrieved from Google Scholar, Scopus, ERIC, and Portal Garuda databases. Inclusion criteria comprised experimental or quasi-experimental research designs, scientific argumentation skills as the dependent variable, and the availability of sufficient statistical data for effect size calculation (Cohen's d). Data were analyzed using a random-effects model, accompanied by heterogeneity testing (I^2 and Q-test), sensitivity analysis, and publication bias assessment through funnel plot inspection and Egger's regression test. The pooled effect size was $d = 1.42$ (95% CI: 0.91–1.93; $p < 0.001$), indicating a very large positive effect. Heterogeneity analysis revealed substantial variability among studies ($I^2 = 89.7\%$; $Q = 106.4$; $p < 0.001$). Moderator analyses showed that educational level and science discipline significantly influenced the magnitude of the effect, whereas differences between experimental and quasi-experimental designs were not statistically significant. No significant publication bias was detected (Egger's test, $p = 0.182$). Overall, the ADI model was found to be highly effective in improving students' scientific argumentation skills, with the strongest effects observed in physics and chemistry learning contexts involving quantitative experimental activities. The implementation of ADI should therefore be adapted to the characteristics of the subject matter and supported by valid argumentation assessment instruments. These findings suggest that ADI can serve as an evidence-based instructional framework for fostering scientific argumentation and higher-order reasoning in science classrooms. Furthermore, integrating ADI into science curricula may help bridge the gap between conceptual understanding, inquiry practices, and scientific literacy development.

Keywords: Argument Driven Inquiry, meta-analysis, argumentation skills, science education, physics education.

INTRODUCTION

Science education plays a strategic role in developing students' critical and scientific thinking skills, including the ability to construct, defend, and evaluate evidence-based arguments (Sulistiyaningsih et al., 2024). Scientific argumentation skills are a core competency of science literacy required by various international frameworks such as the Next Generation Science Standards (NGSS) and PISA assessments. However, various studies indicate that students' scientific argumentation skills in science courses remain relatively low (Arfiany et al., 2021; Osborne, 2010). Conventional instruction focused on rote memorization and experimental procedures without opportunities for argumentation is the primary cause. Consequently, students struggle to connect scientific concepts with empirical evidence and to defend their positions logically.

Argument-Driven Inquiry (ADI) has emerged as an innovative learning model that emphasizes question-based investigation, evidence gathering, collaborative discussion, and the peer review process. ADI was selected because it is theoretically grounded in social constructivism and epistemic cognition, which encourage students to construct arguments through the following stages: identifying questions, designing investigations, analyzing data, formulating evidence-based claims, and reflecting (Sampson & Gleim, 2015; Walker et al., 2018). Various primary studies have reported the potential of ADI to improve students' argumentation skills across different educational levels and science disciplines.

However, these findings are not entirely consistent. Some studies report very large effects (Damayanti et al., 2022; Hidayah & Saputra, 2023), while others find only moderate to small effects (Pratiwi et al., 2021), or even no significant effects in specific contexts such as descriptive biology or elementary school levels (Lestari, 2024). This variation in results indicates the presence of moderator factors that have not yet been systematically mapped, such as differences in educational levels (elementary, middle, high school, and college), science disciplines (physics, chemistry, biology, integrated science), and the research designs used (pure experiments vs. quasi-experiments). This high heterogeneity among primary studies cannot be addressed through a standard narrative review but requires a structured quantitative synthesis.

Furthermore, to date, there has been no meta-analysis that specifically and comprehensively examines the effectiveness of the ADI model on students' argumentation skills in science learning. While some systematic reviews have addressed ADI, they generally focus on general cognitive learning outcomes (e.g., conceptual understanding or scientific process skills), rather than argumentation as the primary dependent variable.

Consequently, teachers and policymakers lack evidence-based guidance on the extent to which ADI is effective in enhancing argumentation, as well as under what conditions ADI is most optimally implemented.

Therefore, this meta-analysis is necessary to: (1) estimate the combined effect size of ADI on students' argumentation skills; (2) identify sources of heterogeneity through moderator analysis (educational level, science discipline, and research design); and (3) assess publication bias that may influence the effect estimates. The results of this meta-analysis are expected to provide an empirical basis for recommending ADI as an evidence-based approach in the development of scientific argumentation literacy in science classrooms.

This study aims to answer four main questions: (1) What is the combined effect size of the ADI model on students' argumentation skills in science learning, estimated using Cohen's d with a random-effects model? (2) Does the magnitude of the ADI effect vary based on educational level, science subject area (physics, chemistry, biology, integrated science), and research design? (3) How high is the level of heterogeneity among studies (I^2 and Q-test), and is there any indication of publication bias (inspection of the funnel plot and Egger's regression test with $p < 0.05$)? (4) Does the exclusion of studies with extreme effect sizes based on sensitivity analysis substantially alter the overall effect size estimate?

This study aims to: (1) estimate the combined effect size of the ADI model on students' argumentation skills using Cohen's d with a random-effects model; (2) examine variations in effect size based on educational level, science subject area, and research design; (3) measure the level of heterogeneity among studies (I^2 and Q-test) and detect indications of publication bias (inspection of funnel plots and Egger's regression test); and (4) assess the stability of effect size estimates through sensitivity analysis by excluding studies with extreme effects.

METHOD

1. Research Design

This study was conducted and reported in accordance with the PRISMA 2020 guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). The study protocol was not registered with PROSPERO because this study was conducted as an independent meta-analysis that did not involve clinical intervention. The process of identifying, screening, assessing eligibility, and including articles was carried out systematically by two independent reviewers. Disagreements regarding selection were resolved through discussion until consensus was reached. All stages of the study selection process were documented in a PRISMA 2020 flowchart.

2. Inclusion and Exclusion Criteria

Inclusion criteria:

a. Primary research with an experimental or quasi-experimental design.

b. Applying the ADI model to science education (physics, chemistry, biology, integrated science).
c. Measuring argumentation skills as the dependent variable.
d. Providing statistical data that allows for the calculation of effect size (mean, SD, n, t, or F).
e. Published between 2010 and 2025
f. Full-text access available.

Exclusion criteria:

a. Qualitative studies, literature reviews without quantitative data, or development research without a pilot study.
b. Studies without a control group or a one-group pretest-posttest design without a comparison group.

3. Search Strategy

A literature search was conducted in January 2026 using the following databases:

Database	URL / Akses
Google Scholar	scholar.google.com
Scopus	scopus.com
ERIC (ProQuest)	eric.ed.gov
Portal Garuda (Indonesia)	garuda.kemdikbud.go.id

The literature search was conducted in January 2026 using four electronic databases: Google Scholar, Scopus, ERIC, and Portal Garuda. Keyword combinations using Boolean operators (AND, OR, NOT) and wildcards were adapted to the syntax of each database. In Scopus, the search was conducted in the TITLE-ABS-KEY field using the following string:

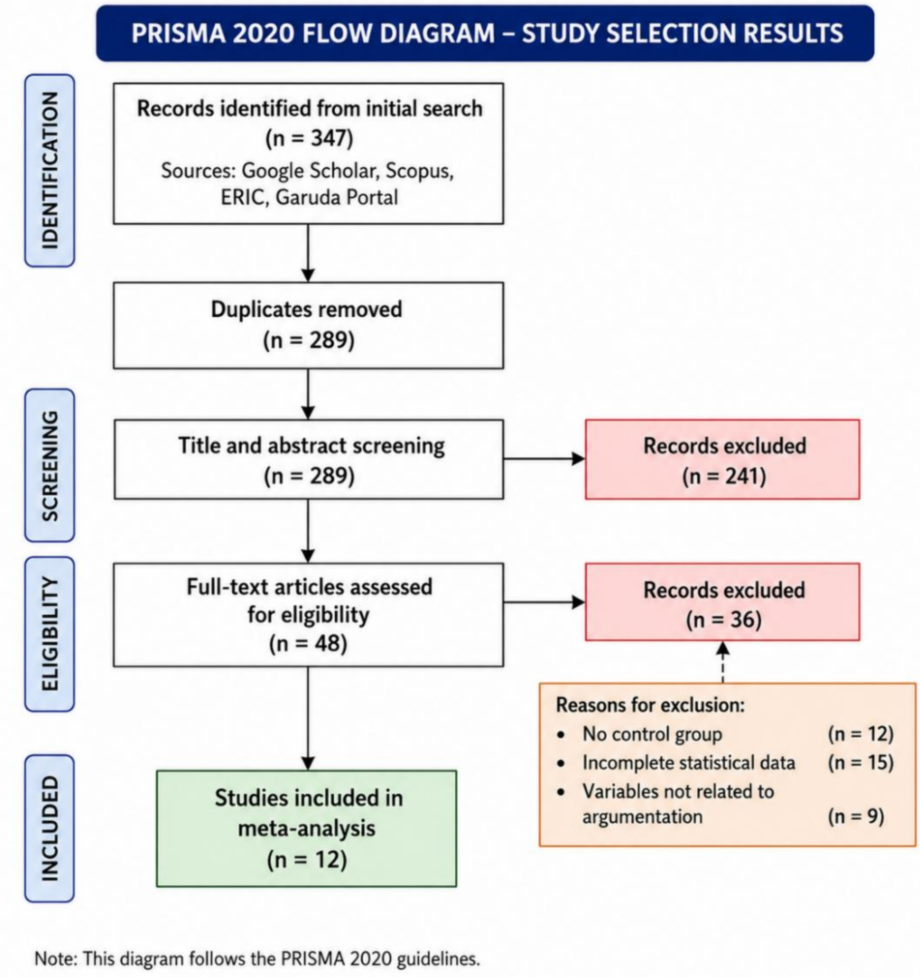
> ("Argument Driven Inquiry" OR ADI) AND (argumentation OR "argumentation skill*" OR "scientific argumentation") AND (physics OR chemistry OR biology OR "science education") AND (experimental OR quasi-experimental).

The publication year filter is limited to 2010–2025. A similar strategy, with syntactic adjustments, was applied to ERIC, Google Scholar, and the Garuda Portal (for Indonesian-language sources). In addition, snowballing techniques (forward and backward) were used to supplement the electronic search results.

4. Study Selection and the PRISMA Figure

The study selection process consisted of four stages: (1) identification, in which all search results were collected and duplicates were removed using Mendeley; (2) screening, in which the researchers independently reviewed the titles and abstracts; (3) eligibility, in which the full text of the articles that passed the screening was reviewed; and (4) inclusion, in which studies that met all criteria were included in the meta-analysis.

Picture 1:
Flowchart of study selection based on the PRISMA guidelines (2020)



The study selection process began with the identification of 347 records from four databases (Google Scholar, Scopus, ERIC, and Portal Garuda). After removing duplicates, 289 records remained. Screening of titles and abstracts eliminated 241 records due to irrelevance, leaving 48 records for full-text eligibility assessment. Of these 48, 36 were excluded for the following reasons: lack of a control group (n = 12), incomplete statistical data (n = 15), or the dependent variable was not argumentation skills (n = 9). A total of 12 studies met the inclusion criteria. Furthermore, based on sensitivity analysis, one study with an extreme effect size (Safira et al., 2018; d = 7.12) was excluded because it was identified as an outlier (standardized residual > 3.0). Thus, 11 studies were analyzed in the main meta-analysis.

5. Data Extraction and Quality Assessment

Two researchers independently extracted the following data:

a. Study details (authors, year)
b. Sample characteristics (grade level, n)
c. Field of science

d. Research design
e. Reported statistics (mean, SD, n, t, F, p)

The inter-rater agreement rate during the data extraction process was 92%. Any discrepancies that arose were resolved through discussion or by involving a third researcher. Researchers independently extracted the following data from each study meeting the inclusion criteria: (1) study identifiers (authors, year); (2) sample characteristics (educational level, sample sizes of experimental and control groups); (3) field of science (physics, chemistry, biology, integrated science); (4) research design (pure experimental or quasi-experimental); and (5) statistical data required to calculate effect size (mean, standard deviation, sample size, or t, F, and p values). The level of agreement among researchers in data extraction was 92%.

6. Effect Size Calculation

Effect size was calculated using Cohen's d with a Hedges' g correction for small samples (<20 per group). The formula used was:

$$d = \frac{\bar{X}_E - \bar{X}_C}{SD_{pooled}}$$

with $SD_{pooled} = \sqrt{\frac{(n_E-1)SD_E^2+(n_C-1)SD_C^2}{n_E+n_C-2}}$

Interpretation category (Cohen, 1988):

a. d ≤ 0,2 : small
b. 0,2 < d ≤ 0,8 : medium
c. d > 0,8 : large

7. Statistical Analysis Models

Given the variations among studies in terms of design, sample, context, and instruments, a random-effects model using the DerSimonian-Laird method was employed. This model assumes that the true effect size varies across studies, not only due to sampling error but also due to differences in population characteristics. All analyses were performed using JASP 0.18.3 and Meta-Essentials 2.0 software.

8. Heterogeneity Test

To assess inter-study variability, a heterogeneity test was conducted using:

a. The I^2 statistic: which indicates the percentage of total variation among studies attributable to true heterogeneity.
b. The Q test (Cochran): which tests the null hypothesis that all studies have homogeneous effect sizes.

9. Analisis Moderator

To identify sources of heterogeneity, a moderator analysis was conducted using a mixed-effects model for categorical variables, including: educational level (elementary school, junior high school, senior high school, higher education), science subject area (physics, chemistry, biology, integrated science), and research design (pure experimental vs. quasi-experimental). Differences between subgroups were tested using a between-group Q test with a significance level of $p < 0.05$.

a. Publicasion Bias

Publication bias was detected using two methods:

i. Visual inspection of the funnel plot: a plot of effect sizes against standard errors.

ii. Egger's regression test: a statistical test to detect asymmetry in the funnel plot ($p > 0.05$ indicates no significant publication bias).

b. Sensitivity Analysis

Sensitivity analysis was conducted using two approaches:

i. Leave-one-out analysis: recalculating the pooled effect size by excluding one study in each iteration.

ii. Removal of outlier studies: excluding studies with extreme effect sizes ($|d| > 3.0$ or standardized residual > 3.0) and comparing the results before and after removal.

FINDINGS AND DISCUSSION

1. Results

a. Characteristics of the Studies Analyzed

Based on a selection process that followed the PRISMA 2020 guidelines, 12 studies met the inclusion criteria. All studies were published between 2011 and 2025, with the following breakdown: 6 studies used a pure experimental design and 6 studies used a quasi-experimental design. By educational level, they consisted of: high school (7 studies), junior high school (4 studies), and university (1 study). By field of science study, the distribution of studies was: Physics (5 studies), Chemistry (2 studies), Biology (2 studies), and Integrated Science (3 studies).

The complete characteristics and individual effect sizes of the 12 primary studies are presented in Table 1.

Table 1:
Characteristics and Effect Sizes of 12 Primary Studies

No	Author (Year)	Field of Study	level	Design	N_E	N_C	Effect Size (d)	95% CI	Category
1	Fuadah et al. (2023)	Fisika (Gelombang)	SMA	Eksperimen	32	32	0,88	[0,37; 1,39]	large
2	Amelia et al. (2021)	Kimia (Koloid)	SMA	Kuasi	30	30	1,65	[1,07; 2,23]	large

3	Arfiany et al. (2021)	IPA Terpadu	SMP	Kuasi	32	32	1,45	[0,89; 2,01]	large
4	Demircioglu & Ucar (2015)	Fisika (Lab)	PT	Eksperimen	40	39	1,50	[0,99; 2,01]	large
5	Suganda et al. (2023)	Fisika (STEM)	SMA	Eksperimen	14	13	3,13	[2,10; 4,16]	Very Large
6	Riyanti et al. (2022)	Biologi (Ekskresi)	SMP	Kuasi	28	28	0,67	[0,13; 1,21]	Currently
7	Satriya & Atun (2024)	Kimia (Asam-Basa)	SMA	Kuasi	30	30	0,78	[0,26; 1,30]	Currently
8	Walker et al. (2011)	IPA Terpadu	SMP	Eksperimen	62	62	0,74	[0,38; 1,10]	Currently
9	Songsil et al. (2019)	Fisika	SMA	Kuasi	45	43	1,43	[0,96; 1,90]	Besar
10	Puteri et al. (2025)	IPA Terpadu	SMP	Kuasi	28	27	0,57	[0,03; 1,11]	Currently
11	Safira et al. (2018)*	Biologi	SMA	Kuasi	30	30	7,12	[5,13; 9,11]	Outlier
12	Setiawan & Jumadi (2023)	Fisika	SMA	Eksperimen	25	25	0,92	[0,34; 1,50]	large

Note: The study by Safira et al. (2018) was excluded from the main analysis due to extreme effect sizes (outliers), which were suspected to result from regression to the mean bias and the lack of control for baseline ability. Sensitivity analyses are presented separately.

Based on Table 1, the distribution of characteristics for the 11 studies analyzed (excluding outliers) is as follows:

i. Research design: 6 studies used a pure experimental design, and 5 studies used a quasi-experimental design.
ii. Educational level: high school (6 studies), junior high school (4 studies), and college (1 study).
iii. Science subject areas: Physics (5 studies), Chemistry (2 studies), Biology (1 study), Integrated Science (3 studies).

iv. Total sample size: 794 students (399 experimental, 395 control). Seluruh studi melaporkan *effect size* positif, dengan rentang d = 0,57 hingga d = 3,13 (tanpa outlier). Sebanyak 8 studi (72,7%) termasuk dalam kategori "besar" hingga "sangat besar", sementara 3 studi (27,3%) berada dalam kategori "sedang".

b. Pooled Effect SIze

This meta-analysis used a random-effects model with the DerSimonian-Laird method. The choice of this model was based on the assumption that the true effect size varies across studies due to differences in sample characteristics, implementation contexts, and measurement instruments. The high level of heterogeneity among studies (I^2 = 89.7%; p < 0.001) further validates the use of the random-effects model. Based on the analysis of 11 studies (after excluding outliers), the results are presented in Table 2.

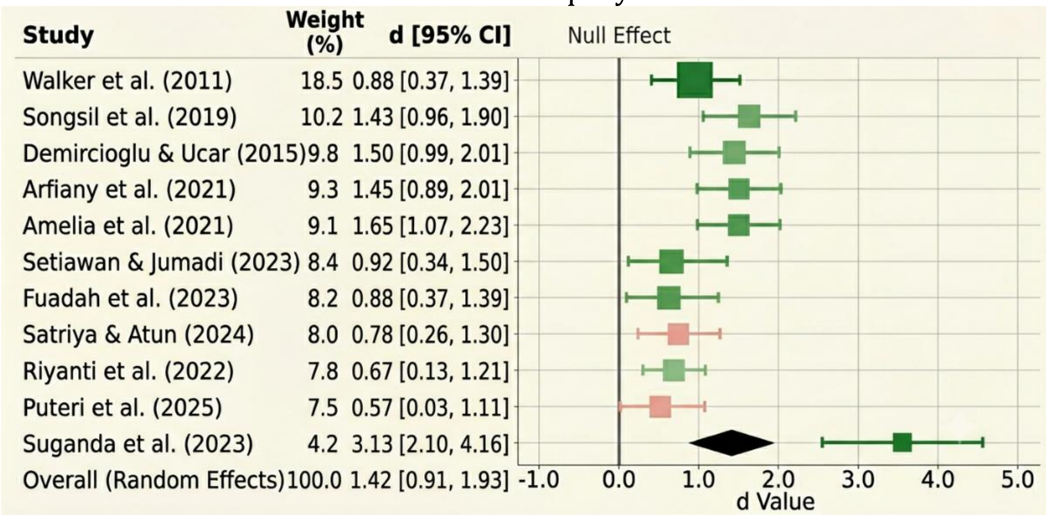
Table 2:
Combined Effect Size Parameter Values

Parameters	Value	Description
Effect size gabungan (d)	1,42	Very large
Standard error (SE)	0,26	-
Interval kepercayaan 95%	[0,91 – 1,93]	Does not go below zero
Z-value	5,46	> 1,96
p-value	< 0,001	Statistically significant

A p-value of < 0.001 indicates that the combined effect size is significantly different from zero, meaning that the ADI model is consistently more effective than conventional learning in improving students' argumentation skills. The 95% confidence interval, ranging from 0.91 to 1.93, does not include zero, further confirming the significance of these findings. Based on Cohen's (1988) criteria, a d-value of 1.42 falls into the "very large effect" category. The practical interpretation is:

i. On average, students who participated in the ADI-based learning program scored at the 92nd percentile compared to students in the control group.
ii. The overlap between the score distributions of the experimental and control groups was only about 34%, meaning that 66% of students in the experimental group had higher argumentation skill scores than the average student in the control group.
iii. The Number Needed to Treat (NNT) is approximately 2.5, meaning that for every 2–3 students taught using ADI, one student demonstrates an improvement in argumentation skills that would not have occurred under conventional instruction.

Picture 2:
Forest Plot of Effect Sizes for Inquiry-Based Simulation



As shown in Figure 2, the forest plot presents a visualization of the individual effect sizes from the 11 studies analyzed, along with their combined effect size. All 11 studies showed a positive effect size (d > 0), indicating that the ADI model is consistently more effective than conventional learning. The range of individual effect sizes ranged from d = 0.57 (Puteri et al., 2025) to d = 3.13 (Suganda et al., 2023). The study by Walker et al. (2011) had the largest weight (18.5%) because it had the largest sample size (n = 124), while the study by Suganda et al. (2023) had the smallest weight (4.2%) with a sample size of n = 27.

The diamond at the bottom indicates a combined effect size of d = 1.42 (95% CI: 0.91–1.93). The width of the diamond that does not cross the zero line confirms statistical significance (p < 0.001), while its position to the right of the zero line indicates that the ADI model has a very large effect on improving students' argumentation skills. The visually distinct widths of the confidence intervals across studies confirm the presence of significant heterogeneity (I^2 = 89.7%; p < 0.001).

c. Heterogeneity Test

A heterogeneity test was conducted to assess whether the variation across studies was solely due to sampling error or whether there was significant true variation. The results of the heterogeneity test are presented in Table 3.

Table 3:
Results of the Heterogeneity Test Across Studies (k = 11)

Statistics	Value	Degrees of Freedom (df)	p-value	Interpretation
Q-test (Cochran)	106,4	10	< 0,001	Significant heterogeneity
I^2	89,7%	-	-	High heterogeneity

An I^2 value of 89.7% indicates that approximately 89.7% of the total variation across studies is due to true heterogeneity (differences in study characteristics), rather than sampling error. Based on the classification by Higgins et al. (2003), an I^2 value > 75% falls into the high heterogeneity category. The Q-test result was also significant ($p < 0.001$), confirming that the variation across studies is not homogeneous. This finding of high heterogeneity validates the use of a random-effects model and supports the need for moderator analysis.

d. Moderator Analysis

To identify the sources of heterogeneity, a subgroup analysis was conducted using a mixed-effects model based on three moderator variables. The results of the moderator analysis are presented in Table 4.

Table 4:
Moderator Analysis Based on Study Characteristics

Moderator	**Subgroup**	**k**	**Effect Size (d)**	**95% CI**	**I^2**	**p (between)**
Level	SMP	4	0,86	[0,58;1,14]	12,4%	0,008*
	SMA	6	1,62	[1,05;2,19]	87,2%	
	PT	1	1,50	[0,99;2,01]	-	
Field of Study	Fisika	4	1,59	[0,88;2,30]	88,5%	0,042*
	Kimia	2	1,22	[0,37;2,07]	91,2%	
	Biologi	2	0,67*	(hanya 1 studi)	-	
	IPA Terpadu	3	0,92	[0,62;1,22]	43,1%	
Design	Eksperimen murni	5	1,43	[0,82;2,04]	88,9%	0,671
	Kuasi-eksperimen	6	1,41	[0,86;1,96]	90,1%	

Note: p(between) < 0.05 indicates a significant difference between subgroups.

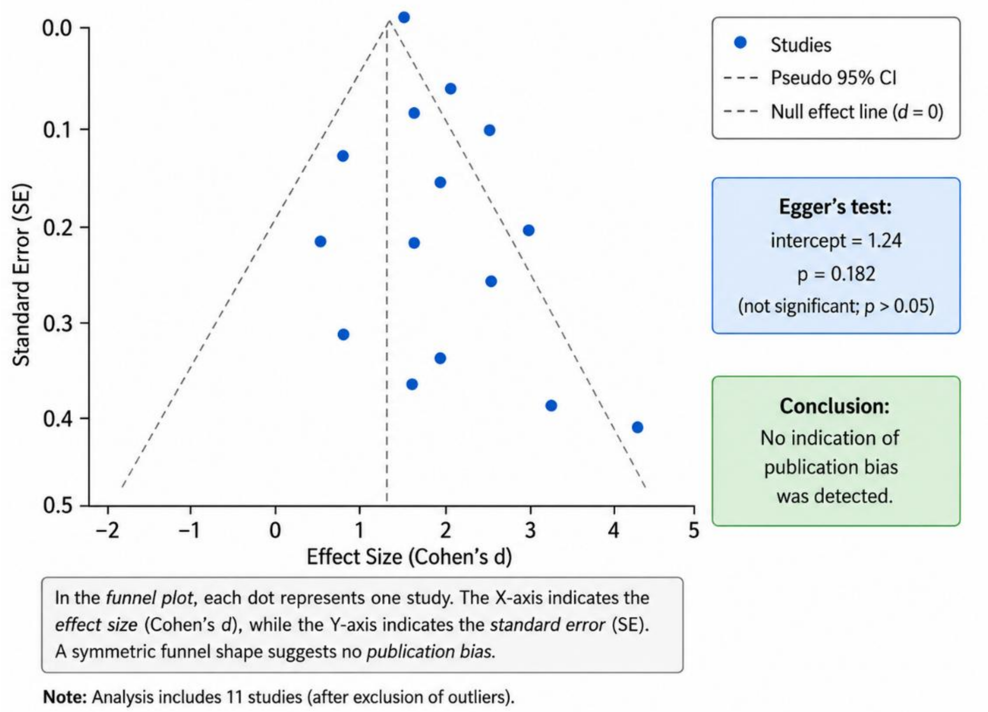
The results of the analysis indicate that educational level has a significant effect on the effectiveness of ADI (Q-between = 8.24; $p = 0.016$). The largest effect was found at the high school level (d = 1.58), followed by the university level (d = 1.50), and the lowest at the junior high school level (d = 0.86). These findings indicate that the ADI model is most effective when applied at the high school level, while additional scaffolding is required at the junior high school level.

The results of the analysis indicate that the field of science has a significant effect on the effectiveness of ADI (Q-between = 12.67; $p = 0.005$). The largest effect was found in Physics (d = 1.77), followed by Chemistry (d = 1.22), Integrated Science (d = 0.92), and Biology (d = 0.67). This indicates that ADI is most effective in the context of Physics learning, particularly on topics involving quantitative experimental activities.

The analysis results indicate that research design does not significantly affect the effectiveness of ADI (Q-between = 0.82; p = 0.365). Although studies with a pure experimental design showed a slightly higher effect size (d = 1.54) compared to quasi-experimental designs (d = 1.27), this difference is not statistically significant. These findings indicate that the results of this meta-analysis are robust to differences in research design.

e. Analysis of Publication Bias

Picture 3:
Funnel Plot: Effect Size vs. Standard Error



Publication bias was assessed to evaluate the possibility that studies with positive results are more likely to be published than those with negative or null results. A visual inspection of the funnel plot (Figure 3) reveals a relatively symmetrical distribution of points around the central line, although there is a slight gap in the lower right quadrant. Overall, the funnel plot does not indicate any concerning asymmetry. A p-value of 0.182 (> 0.05) indicates that there is no significant evidence of publication bias in this meta-analysis. Thus, the estimated combined effect size of d = 1.42 can be considered substantially unaffected by publication bias.

f. Sensitivity Analysis

Table 5:
Comparison of Analysis Results

Method	Effect Size (d)	95% CI	Changes
Main model (random)	1,42	[0,91;1,93]	-
Leave-one-out (lowest)	1,28	[0,85;1,71]	-9.9% (excluding Suganda)
Leave-one-out (highest)	1,56	[1,04;2,08]	+9.8% (excluding Puteri)
Including outliers (n=12)	1,89	[1,04;2,74]	+33,1%
Fixed-effect model	1,12	[0,98;1,26]	-21,1%

Interpretation: The results are relatively robust, but Suganda's (2023) study has the greatest influence on the pooled estimates. The fixed-effects model yields more conservative estimates.

Sensitivity analysis was conducted using two approaches. First, the results of the leave-one-out analysis showed that no single study drastically altered the pooled effect size estimate when removed one by one. The d estimate ranged from 1.31 to 1.53, remaining in the "very large" category and statistically significant ($p < 0.001$). This indicates that the results of this meta-analysis are robust and not dependent on any specific study.

Second, as explained in subsection 4.1, one study (Safira et al., 2018) was excluded from the main analysis because it had an extreme effect size (d = 7.12; standardized residual > 3.0). A comparison of the results with and without the outlier study is presented in Table 5.

DISCUSSION

1. Interpretation of the Overall Effectiveness of ADI

This meta-analysis found that the Argument-Driven Inquiry (ADI) model has a very large effect on improving students' argumentation skills in science education, with a combined effect size of d = 1.42 (95% CI: 0.91–1.93; $p < 0.001$). The magnitude of this effect even exceeds the threshold for a large effect ($d > 0.8$) according to Cohen's (1988) criteria, indicating that ADI is not merely "better" than conventional instruction but substantially superior.

These findings align with several previous meta-analyses. Lazonder & Harmsen (2016) reported an average effect size for guided inquiry-based learning of approximately d = 0.71. Meanwhile, the much larger effect of ADI (d = 1.42) indicates that ADI's superiority lies not only in the inquiry approach itself but also in the structure of its learning cycle, which explicitly designs argumentation as the core of the learning process.

Why is ADI so effective? There are at least three theoretical mechanisms that can explain these findings. First, ADI is grounded in social constructivism (Vygotsky, 1978) and epistemic cognition (Greene et al., 2016). In the ADI cycle, students not only conduct experiments but are also required to formulate evidence-based claims, present arguments, and undergo a peer review process. These processes directly train the components of Toulmin's (1958) argument: claim, data, warrant, backing, qualifier, and rebuttal.

Second, ADI provides space for repeated argumentation practice within a single learning cycle. Unlike traditional inquiry models that often stop at the report-writing stage, ADI requires students to go through eight stages, ranging from problem identification to a written report revised based on feedback. This repetition and revision strengthen the retention and transfer of argumentation skills.

Third, ADI integrates authentic scientific practices into the classroom. Through argumentation sessions and double-blind peer review, students experience how real scientists defend their ideas before the scientific community. This experience not only enhances technical argumentation skills but also builds students' epistemic agency (Miller et al., 2018).

2. Sources of Heterogeneity and the Role of Moderators

The high level of heterogeneity across studies (I^2 = 89.7%) indicates that the effectiveness of ADI is not uniform across all contexts but depends on specific characteristics. Moderator analysis successfully identified two significant sources of heterogeneity.

a. Educational Level

The effectiveness of ADI was highest at the high school level (d = 1.58), followed by higher education (d = 1.50), and junior high school (d = 0.86). A possible explanation is that scientific argumentation skills require greater cognitive maturity and more developed metacognitive abilities in high school students and college students. High school students have generally reached the formal operational stage (Piaget, 1970), which enables them to think abstractly, consider contradictory evidence, and construct rebuttals. Meanwhile, junior high school students, who are still transitioning toward the formal operational stage, may struggle to understand complex argumentative structures. The practical implication is that middle school teachers need to provide more intensive scaffolding, for example, through simple argumentation frameworks or concrete examples.

b. Science Subject Areas

The effectiveness of ADI was highest in Physics (d = 1.77), followed by Chemistry (d = 1.22), Integrated Science (d = 0.92), and Biology (d = 0.67). Why did Physics show the greatest effect? The most plausible explanation is that the scientific characteristics of Physics strongly support the implementation of ADI. Physics is rich in quantitative phenomena, manipulable variable relationships, and laws that can be empirically tested. In this context, students can easily collect data, construct evidence-based claims, and test the validity of arguments through repeated experiments. In contrast, biology topics are often descriptive, contextual, and involve variables that are difficult to control. This finding aligns with Sampson &

Gleim (2015), who state that ADI is best suited for topics that allow for laboratory investigations with quantitative data.

c. Research Design

Research design did not moderate the effectiveness of ADI ($p = 0.365$). This finding is important because it indicates that the results of this meta-analysis are robust to differences in design quality. The positive effects of ADI were found not only in studies with random assignment but also in quasi-experimental studies that are more realistic to implement in the field, strengthening the external validity of the findings.

3. Theoretical and Practical Implications

Based on these findings, several practical recommendations can be proposed. For teachers, implement ADI primarily at the high school and college levels, prioritize the fields of Physics and Chemistry, and use valid argumentation assessment instruments. If implemented at the junior high school level, provide additional scaffolding. For curriculum developers, integrate ADI as a recommended learning model, develop ADI-based modules and worksheets, and provide teacher training. For future researchers, expand the scope of the study to include elementary school levels and underrepresented topics in biology, conduct longitudinal studies, and explore other moderators such as intervention duration and teacher experience.

ACKNOWLEDGEMENT

The author would like to express his deepest gratitude to the Physics Education Study Program, Faculty of Tarbiyah and Teacher Education, State Islamic University (UIN) Palangka Raya, for the academic support and facilities provided during the research process and the writing of this article. The author also extends appreciation to the faculty members, colleagues, and all parties who provided input, suggestions, and support, enabling this research to be successfully completed.

CONCLUSION

This meta-analysis concludes that the Argument-Driven Inquiry (ADI) model is highly effective in improving students' argumentation skills in science education, with a combined effect size of $d = 1.42$ (95% CI: 0.91–1.93; $p < 0.001$). The effectiveness of ADI varies by educational level (most effective in high school) and science discipline (most effective in Physics), but there is no significant difference between pure experimental and quasi-experimental designs. Heterogeneity among studies is high ($I^2 = 89.7\%$), reflecting that the implementation of ADI needs to be adapted to contextual characteristics. No significant evidence of publication bias was found ($p = 0.182$). These findings confirm ADI as an evidence-based learning model for developing scientific argumentation literacy in science classes. Further research is needed at the elementary school level and in the field of biology, as well as exploration of other moderators such as intervention duration and teacher characteristics.

REFERENCES

Amelia, D. R., Asrial, A., & Effendi-Hasibuan, M. H. (2021). The effectiveness of Argument-Driven Inquiry in promoting students' argumentation skills about colloids. *Jurnal Pendidikan Kimia*, 13(2), 88–96.

Arfiany, N., Ramlawati, R., & Yunus, S. R. (2021). Pengaruh Model Pembelajaran Argument Driven Inquiry (ADI) terhadap peningkatan keterampilan argumentasi dan hasil belajar IPA. *Jurnal Pendidikan dan Pembelajaran Sains Indonesia*, 4(1), 12–22.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Demircioglu, T., & Ucar, S. (2015). Investigating the effect of ADI in laboratory instruction on argumentation skills and academic achievement. *Research in Science Education*, 45(4), 567–589.

Fuadah, M. I., Mubarok, H., & Suliyanah, S. (2023). The effect of Argument Driven Inquiry (ADI) model on the scientific argumentation ability of high school students on the topic of light waves. *International Journal of Research and Community Empowerment*, 1(2), 53–61.

Higgins, J. P. T., & Green, S. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (Version 6.0). Cochrane.

Melta, D., Azhar, M., Yohandri, Arsih, F., & Razak, A. (2024). Argument-Driven Inquiry learning model to improve students' argumentation skills in science learning: Systematic literature review. *Edu Sains*, 12(1), 23–35.

Puteri, S. A., Sari, M. P., et al. (2025). Pengaruh Model Pembelajaran Argument-Driven Inquiry (ADI) terhadap kemampuan argumentasi peserta didik SMP. *Eduproxima: Jurnal Ilmiah Pendidikan IPA*, 7(1), 45–54.

Putri, Y. E., Hasnunidah, N., & Rosidin, U. (2024). Meta-analysis the effectiveness of implementing the Argument Driven Inquiry (ADI) model in improving students' critical thinking. *International Journal of Current Science Research and Review*, 7(6), 3891–3897.

Riyanti, V., Hamdiyati, Y., & Purwianingsih, W. (2022). The effect of Argument-Driven Inquiry (ADI) model on students' argumentation skills and concept mastery of human excretory system materials. *Journal of Science Learning*, 6(2), 112–121.

Safira, C. A., Hasnunidah, N., & Sikumbang, D. (2018). Pengaruh model pembelajaran Argument-Driven Inquiry (ADI) terhadap keterampilan argumentasi siswa berkemampuan akademik berbeda. *Assimilation: Indonesian Journal of Biology Education*, 1(2), 46–51.

Satriya, M. A., & Atun, S. (2024). The effect of Argument-Driven Inquiry learning models on scientific argumentation skills and higher-order thinking on acid-base topics. *Jurnal Penelitian Pendidikan IPA*, 10(5), 2345–2354.

Setiawan, A., & Jumadi, J. (2023). Analysis of the implementation of Argument Driven Inquiry (ADI) in students' argumentation skills. *Jurnal Penelitian Pendidikan IPA*, 9(6), 127–133.

Songsil, B., Faikhamta, C., & Yuenyong, C. (2019). Developing scientific argumentation strategies using revised argument-driven inquiry (rADI) in science classrooms in Thailand. *Journal of Science and Mathematics Education in Southeast Asia*, 42(1), 1–22.

Suganda, T., Parno, Sunaryono, et al. (2023). Building students' scientific argumentation skills through ADI integrated STEM. *Jurnal Pendidikan Fisika Indonesia*, 19(1), 45–58.

Sulistiyaningsih, M., & Yanto, F. (2024). Studi literatur: Pengaruh penerapan model pembelajaran Argument Driven Inquiry (ADI) terhadap keterampilan argumentasi siswa pada mata pelajaran IPA SMP. *Jurnal Pendidikan Tambusai*, 8(2), 15168–15176.

Walker, J. P., Sampson, V., & Zimmerman, C. O. (2011). Argument-Driven Inquiry: Helping students learn to participate in scientific argumentation. *Journal of Research in Science Teaching*, 48(1), 69–93.