International Journal of Educational Sciences and Arts (IJESA)

المجلة الدولية للعلوم التربوية والآداب

IJESA

Vol. (5), No. (9) September 2026 الإصدار (5)، العدد (9)

Effectiveness of Argument-Driven Inquiry Combined with Peer Assessment in Developing Epistemic Cognition and Laboratory Report Writing Skills in General Chemistry: A Randomized Cluster Trial

Sarah Waleed Rasheed

Ministry of Higher Education and Scientific Research, Ibn Sina University of Medical and Pharmaceutical Sciences, Iraq

Sarah.waleed@ibnsina.edu.iq

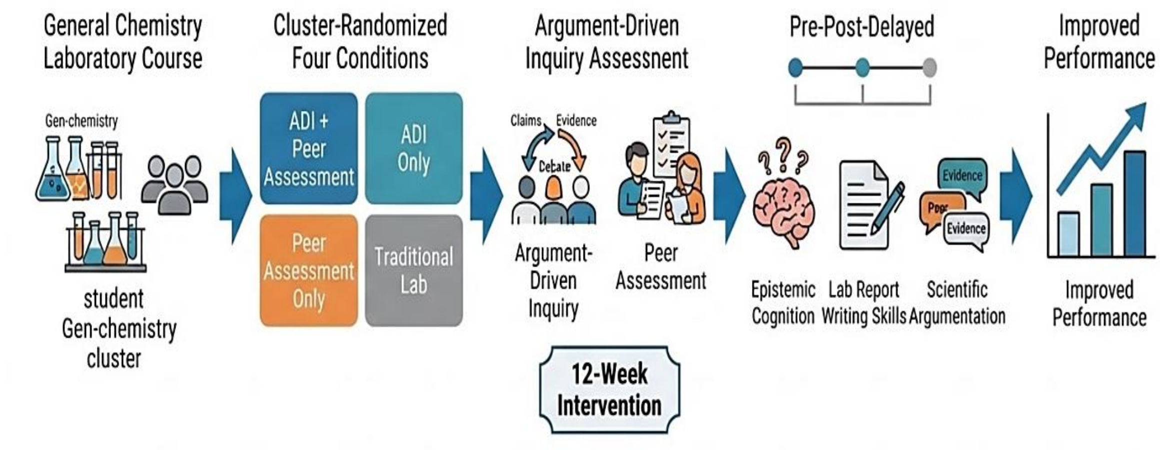
Abstract

We ran a cluster-randomized controlled trial with 228 undergraduate general chemistry students, separating what Argument-Driven Inquiry (ADI) and structured peer assessment each contribute alone, and what they contribute together, to epistemic cognition and laboratory report writing quality. Randomization was at the section level. Twelve laboratory sections went to one of three conditions: ADI only (n = 79), traditional cookbook-style instruction (n = 77), or ADI with structured peer assessment (ADI+PA; n = 72). Writing quality was scored on a six-dimension analytic rubric; for epistemic cognition we adapted a scenario-based instrument from Barzilai and Weinstock (2015). In two-level hierarchical linear models with Kenward–Roger small-sample corrections, ADI+PA students scored significantly higher than both comparison groups on epistemic cognition at posttest ($g = 1.48$, 95% CI [0.89, 2.07], $p < .001$), and five weeks after the intervention ended that advantage, attenuated by then, was still present. Writing quality followed the same ordering (ADI+PA > ADI > Traditional), with especially large effects in claim quality and counter-argument consideration. Causal interpretation is constrained, though, by several methodological caveats: instructor–treatment confounding, rubric familiarity effects, and time-on-task differences we did not control. Across sessions, qualitative discourse analysis of the peer assessment dialogues traced a progressive shift away from surface-level comments toward substantive epistemic engagement. Pairing ADI with peer review may therefore be synergistic, but replication with crossed instructor designs and larger cluster counts is needed before these estimates can be confirmed.

Keywords: Argument-Driven Inquiry, Epistemic Cognition, Peer Assessment, Laboratory Report Writing, Undergraduate Chemistry, Cluster-Randomized Trial.

International Journal of Educational Sciences and Arts (IJESA)
IJESA
المجلة الدولية للعلوم التربوية والآداب
Vol. (5), No. (9) September 2026 الإصدار (5)، العدد (9)

Graphical Abstract



1. Introduction

This is an old story about university chemistry labs. They do not engage children in true scientific thinking. Despite decades of debate about change, the confirmation model has proved remarkably robust, and most laboratory courses still include students doing prescribed procedures, taking notes on the observations they were advised to look for, and confirming what the textbook already says (Hofstein & Lunetta, 2004). A review of over 400 general chemistry classes across the United States by Bretz (2019) revealed that most still employ cookbook type lab instructions, in which students spend more time following processes than forming hypotheses or evaluating evidence.

The results of such a model are well known. Students who have experienced a traditional sequence of laboratory work are normally proficient in the methods, able to run a spectrophotometer, prepare standard solutions, and carry out titrations, but they leave the lab without the epistemic dispositions that characterise scientific thought. They cannot discern opposing claims to knowledge, or tell good evidence from bad evidence, or know when a conclusion goes beyond its supporting evidence. Chinn, Buckland, and Samarapungavan (2011) referred to these capacities as epistemic cognition, the beliefs and reasoning processes that people use in their consideration of the nature, justification, and boundaries of knowing.

These shortcomings are what Argument-Driven Inquiry (ADI) was designed to address. Sampson and Gleim (2009) first articulated the model, and a series of later studies refined it (Walker et al., 2011; Walker & Sampson, 2013; Sampson et al., 2011); what ADI does is restructure the laboratory experience around the core practices of scientific argumentation. No script is handed out. Students design investigations of their own, gather data, build evidence-based arguments, and then put those arguments up for peer critique in a gallery-walk format. The rationale: students who spend their time constructing, evaluating, and revising arguments, the epistemic practices of science, should build their scientific reasoning capacities.

Research on ADI has been largely encouraging. Students in ADI sections wrote higher-quality arguments than peers in traditional labs, as Walker and Sampson (2013) showed, and later work carried the findings into conceptual understanding (Kadayifci & Yalcin-Celik, 2016), scientific practices (Walker et al., 2019), and attitudes toward laboratory work (Hasnunidah et al., 2020). Two significant gaps persist all the same. The first: epistemic cognition has not been tested in any published study as an explicit outcome of ADI teaching. That omission is unusual, because ADI's theoretical underpinnings are grounded in epistemological commitments, as the model explicitly invites students to assess the reliability of their evidence, consider alternative explanations, and update statements as appropriate. Are these activities really detectable changes in epistemic beliefs? It is an open question, and an important one.

Peer assessment inside the ADI framework is the second gap. Stage 7 of the conventional ADI model is a double blind peer assessment of written lab reports. Most empirical investigations however ignore the stage altogether or regard it as a procedural detail rather than an independent variable. This is a squandered opportunity because the theoretical rationale for expecting structured peer review to enhance the effects of ADI is strong. Take social constructivism (Vygotsky, 1978): assessing another student's thinking critically exposes a zone of proximal development where evaluative standards become internalised through social interaction. Genre theory (Swales, 1990; Bazerman, 1988) goes in the same direction as reading and analysing several instances of the lab report genre increases disciplinary enculturation. There is also the epistemic cognition reading (Chinn et al., 2014). According to this view, peer review calls for students to use epistemic standards in a tangible, consequential situation. What is evidence that is good? Does this make sense logistically? Are other explanations fully explored?

These theoretical assumptions are frequently confirmed by meta-analytic investigations on peer assessment. Li et al. (2020) synthesised 58 study and identified a moderate favourable effect of peer assessment on academic achievement ($g = 0.31$), with higher effects when structured rubrics were utilised and students were instructed to offer substantive comments. The advantage is cognitive in Kollar and Fischer (2010), and stems from the 'cognitive conflict' that a reviewer experiences when faced with thinking different from their own. The reconciliation of the two helps understanding. But one thing remains a mystery. Does the combination of peer evaluation and an instructional model already infused with argumentation, such as ADI, lead to additive effects on student cognition and writing or synergistic benefits?

A third is methodological in nature. The literature on ADI has a quasi-experimental design with intact groups, therefore it is impossible to tell if a difference found in any specific study is due to the intervention, pre-existing group traits, or instructor effects. Random assignment appears in only a handful of studies. Fewer still handle, through appropriate multilevel analyses, the clustering that classroom-based research carries with it by design. Schochet (2008) has noted what this costs: ignoring clustering in educational studies can inflate Type I error rates by as much as a factor of five, so findings that look significant may not be reliable.

All three gaps are addressed at once. Twelve laboratory sections were randomly assigned, within a cluster-randomized controlled trial, across three conditions: ADI with structured peer assessment, ADI without peer assessment, and traditional instruction. Does ADI develop epistemic cognition? That was one question. We also asked whether structured peer assessment adds measurably to ADI's effects on scientific writing quality, and, should gains appear, whether they would hold over time and carry to novel chemistry tasks. To our knowledge this is the first study that (a) treats epistemic cognition as a primary outcome of ADI instruction, (b) uses an experimental design to isolate what peer assessment contributes on its own within the ADI framework, and (c) tracks retention as well as transfer of epistemic and writing gains in the chemistry laboratory context.

2. Theoretical Framework

Three complementary theoretical traditions inform this study; read jointly, they predict the cognitive mechanisms through which ADI and peer assessment ought to operate, and they also identify the specific outcome measures on which their effects should be detectable. Figure 1 depicts the integrated model.

2.1 Epistemic Cognition Framework:

Earlier accounts were stage models (Perry, 1970; King & Kitchener, 1994). Chinn, Buckland and Samarapungavan (2011) went beyond them with a multidimensional paradigm for epistemic cognition that distinguished three connected components. Epistemic purposes are the goals that motivate knowledge-seeking behaviour: a learner may want real comprehension, or just right answers. Epistemic ideals are standards for assessing claims to knowing. Does a learner embrace a conclusion because the argument makes sense logically, because an authority has supported it, or because the evidence is compelling? Reliable epistemic processes are the procedures that produce and justify knowledge. controlled experimentation vs. anecdotal observation, peer-reviewed publishing vs. unvetted web sources.

Chinn et al. (2014) further refined this into the AIR model (Aims, Ideals, Reliable processes) and argued that advanced epistemic cognition involves interaction with all three components simultaneously. Laboratory instruction that merely asks students to follow procedures engages epistemic processes in a limited sense (data collection) but does virtually nothing to develop epistemic aims or ideals. ADI, by contrast, engages all three: students must articulate a clear investigative aim, marshal evidence that meets disciplinary ideals of adequacy, and evaluate the reliability of their own and their peers' epistemic processes during argumentation sessions.

2.2 Social Constructivism:

Vygotsky's (1978) theory of social constructivism holds that higher cognitive functions originate in social interaction before being internalized as individual capacities. Two concepts from this tradition are especially relevant. The zone of proximal development (ZPD) describes the gap between what a learner can accomplish independently and what they can achieve with the support of a more knowledgeable other, or, importantly, a peer whose understanding differs from their own. The process of internalization describes how externally mediated activities (arguing with a peer about the adequacy of evidence, for instance) become self-regulated cognitive processes (evaluating evidence quality when writing alone).

The argumentation sessions in ADI are spaces of social mediation: students formulate claims, defend them, and revise them through discussion. Through this back-and-forth they internalise evaluative standards that had, until then, existed only in the social exchange. That mechanism is extended by peer assessment. The social mediation offered is specifically directed at written scientific communication: a student who reads a peer's lab report and then appraises it according to an explicit rubric is doing externally what they must ultimately accomplish inside, as a self-regulated writer and thinker.

2.3 Genre Theory:

Scientific writing is not a generic ability transferable across settings but a genre as described by Swales (1990) and Bazerman (1988): a socially recognised conventionalised style of communication with its own structural features, rhetorical moves and discourse rules. In chemistry, the genre is the laboratory report. Repeated scaffolded interaction with exemplars supports the process of enculturation. Writing these reports is the main way students learn to articulate scientific thinking. In more traditional training, the report may be reduced to a fill-in-the-blanks exercise with little attention for argumentation structure, hedging, citation procedures for evidence, or examination of counter-arguments.

With ADI, the lab report is a true rhetorical act: students write to convince a critical audience, their peers and instructor, that the argument they make is supported by sufficient data and compelling reasoning. The structured peer assessment adds to the enculturation of genre. Students are shown a number of examples of the genre and have to evaluate each one according to explicit criteria. Genre theory suggests that such exposure could speed up the development of genre awareness and improve writing quality in areas not addressed by traditional instruction: the quality of reasoning, the examination of alternative explanations, and the sophistication of evidence use.

2.4 Integration:

The three frameworks all predict the same thing. Cognitive targets (aims, ideals, dependable processes) are taken from epistemic cognition theory. Social constructivism provides the method by which debate and peer evaluation lead to cognitive change: social mediation leading to internalisation. Genre theory, for its part, predicts that scientific writing will become better as students acquire tacit knowledge of disciplinary conventions by writing and analysing lab reports. What the integrated model critically predicts is that ADI coupled with systematic peer assessment should impact epistemic cognition and writing quality exceed those of ADI alone, because peer assessment activates a further layer of social mediation, one focused specifically on evaluative engagement with scientific text.

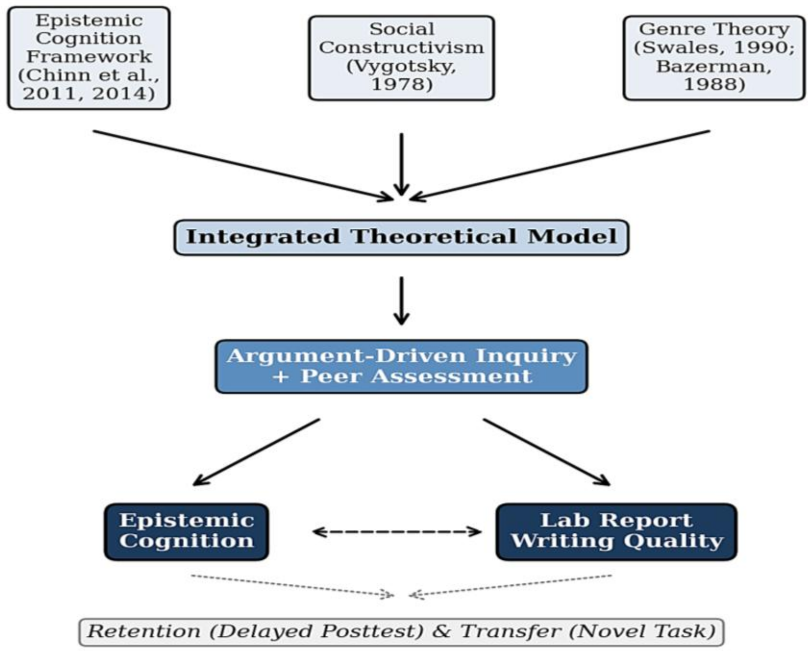


Figure (1): Integrated theoretical framework depicted as a conceptual diagram showing the convergence of three theoretical traditions: epistemic cognition theory (identifying cognitive targets: aims, ideals, reliable processes), social constructivism (specifying the mechanism: social mediation leading to internalization), and genre theory (predicting writing quality improvements through disciplinary enculturation), in generating the central prediction that ADI combined with structured peer assessment will produce synergistic effects on epistemic cognition and laboratory report writing quality that exceed those of either component alone. Arrows indicate hypothesized causal pathways; dashed lines represent the additional mediation layer introduced by peer assessment

3. Literature Review

3.1 Argument-Driven Inquiry in Science Laboratories:

Sampson and Gleim (2009) adapted ADI as an instructional model in undergraduate science laboratories and utilised a seven-stage framework that substituted student-designed investigations for cookbook methods, evidence-based argumentation for rote reporting, and iterative report writing for linear report writing. The first empirical study was conducted by Walker et al. (2011) in which the model was introduced in undergraduate chemistry laboratories. Students in ADI sections showed improved argument quality on written assessments compared to students receiving normal instruction. The results were validated in a bigger study. Walker and Sampson (2013) followed a full semester of general chemistry laboratory and found significant increases in the structural complexity of student arguments and in their evidentiary quality, with medium-to-large effect sizes ($d = 0.42–0.68$).

Further studies have broadened the evidence base across disciplinary contexts and outcome metrics. Sampson, Grooms, and Walker (2011) found that in general chemistry ADI also increased students'

https://doi.org/10.59992/IJESA.2026.v5n9p5 E-ISSN 2976-7237

argument construction abilities and subject understanding, which addresses a common concern that inquiry-based instruction compromises topic coverage for process skills. Grooms, Enderle, and Sampson (2015) found that students taught in ADI produced more epistemically sophisticated arguments than students in comparison groups, arguments that drew more on empirical evidence and less on appeals to authority. Kadayifci and Yalcin-Celik (2016) reported that conceptual knowledge was improved in biology. Attitudes toward laboratory work were also enhanced. In a thorough assessment of 23 ADI research, Hasnunidah et al. (2020b) found that the model was effective in a wide variety of educational contexts, although the methodological quality of the underlying work varied greatly.

One line of work is directly relevant to scientific writing: Sampson et al. (2013) found that students in ADI sections produced lab reports that were much better structured and more rhetorically sophisticated than reports produced by students in regular laboratories. Evidence selection, reasoning, and claim quality showed the biggest improvement. These are the qualities most directly connected to reasoning, while changes in lower-order dimensions like as formatting and language were modest, often non-significant. The pattern is consistent with a mediation. The effects of ADI on writing are mediated by its effects on scientific reasoning, and are not due to a general boost in writing proficiency.

That's true, the ADI literature has several important shortcomings. Most research uses quasi-experimental methods with intact comparison groups that pose problems of selection bias and confounding. Few studies have studied long-term retention or transfer of improvements beyond the initial training context. And as noted in the introduction, epistemic cognition has never been measured as a primary outcome variable in any published study, leaving a crucial theoretical relationship empirically untested.

3.2 Epistemic Cognition in Science Education:

Over the last two decades, research on epistemic cognition in science education has experienced rapid growth, most of which has been inspired by the theoretical frameworks of Hofer and Pintrich (1997), Greene et al. (2010), and Chinn et al. (2011, 2014). Take-home message: Epistemic cognition is adaptable. It is a capacity, not a fixed characteristic, and it is changed by educational treatments, but the majority of treatments have had a modest impact. Muis et al. (2015) conducted a meta-analysis of 26 studies and found that explicit epistemic interventions had an average effect of $d = 0.34$ on measures of epistemic beliefs. Interventions that involved argumentation activities, collaborative reasoning, or the presentation of conflicting evidence had larger effects.

In chemistry, in particular, the data is scant. Deng et al. (2011) investigated how epistemic beliefs relate to conceptual understanding in general chemistry. They found that students who held more sophisticated epistemic beliefs, especially in relation to the tentative and evidence-based nature of chemical knowledge, were better at transfer tasks, even after controlling for prior knowledge and general cognitive ability. Soysal (2024) examined the epistemic anatomy of reform-based chemistry curricula and found that these curricula, while expressing sophisticated epistemological positions, often revert to a transmissionist epistemology in the implemented curriculum, where knowledge is presented as settled and students are positioned as passive recipients rather than active evaluators.

An significant methodological contribution was provided by Barzilai and Weinstock (2015) who developed a scenario-based method for assessing epistemic reasoning which overcomes the well-documented disadvantages of self-report Likert scales. Their instrument presents students with realistic scenarios in which experts disagree, and asks students to evaluate the knowledge claims, the evidence, and the reasoning process. This approach captures the situated, practice-based nature of epistemic cognition in a way that decontextualized belief questionnaires cannot, and it is this instrument that we adapted for use in the present study.

3.3 Peer Assessment in STEM Laboratories:

Peer assessment in higher education is well studied. Evidence points consistently toward positive effects on learning outcomes, and in the most wide-ranging meta-analysis to date Li et al. (2020) synthesized 58 studies to find an overall effect of $g = 0.31$ on academic performance, with substantial heterogeneity across studies. Effect sizes ran larger when peer assessment was structured (i.e., guided by rubrics or checklists), when students had been trained, and where multiple rounds of feedback and revision were built in. Used in a single-shot, unstructured format, peer assessment yielded effects that were smaller or non-significant.

Within laboratory instruction, peer review of written reports has been studied mostly in engineering (Lee & Lim, 2019) and biology (Mynlieff et al., 2014). There aren't very many of them in chemistry. What is the evidence that it is the act of evaluating rather than simply the receipt of feedback that is primarily responsible for learning gains? Lundstrom and Baker (2009) discovered that students who read and evaluated peers' scientific papers improved their own writing more than students who just got reviewer feedback. This finding supports the social constructivist hypothesis that the evaluative act itself generates cognitive progress.

It has not been explored if peer assessment has distinct effects, in magnitude or in kind, whether it is located within an argumentation-rich framework like ADI vs a more typical instructional environment. There is a theoretical potential for synergy. ADI already has the conceptual basis for evaluating scientific arguments in gallery tours and classroom discussions, so structured peer review may build on that basis and apply it to the unique genre of the written lab report. This investigation tests that prediction directly.

3.4 Scientific Writing Quality in Chemistry:

In science education scientific writing is becoming recognised as a basic activity of learning rather than only a form of assessment. Writing may be an epistemic tool, a technique of creating understanding, not simply communicating it, as the Writing-to-Learn movement has proved (Klein, 2004; Bangert-Drowns et al., 2004). In chemistry, however, the tradition has been otherwise. Lab report writing is usually viewed as a reporting rather than a thinking task: students write down observations, compare outcomes with expected values, and summarise procedures, without much emphasis on formulating claims, evaluating evidence or argumentation.

Several research have tried to improve the quality of scientific writing in chemistry courses via specific measures. Sampson et al. (2013) demonstrated that ADI greatly enhanced the quality of lab reports in undergraduate chemistry, with the most pronounced improvements in the argumentation-related aspects. Quitadamo and Kurtz (2007) found that a writing-intensive biology lab course resulted in increased critical thinking skills as measured by standardised instruments. Most recently, Moon et al. (2021) reported that specific teaching in genre rules and structured feedback increased the quality of chemistry laboratory reports among first-year students. Taken together, these findings imply that writing quality increases when training makes the rhetorical and epistemic demands of scientific writing obvious and provides structured opportunities for practice and review.

3.5 Synthesis of Gaps:

The above evaluation points to three interrelated gaps that inspire the present study. First, ADI has been found to improve argument quality and, in some situations, conceptual understanding, but no study has tested its impact on epistemic cognition, despite ADI's theoretical underpinnings making significant epistemic claims. Second, the role of peer assessment within the ADI framework has not been isolated as an independent variable; we do not know whether including or omitting the peer review stage changes the magnitude or nature of ADI's effects. Third, the absence of delayed posttests and transfer tasks in the ADI literature leaves open the question of whether observed gains are durable and generalizable. This study addresses all three gaps within a single experimental design.

4. Research Questions and Hypotheses

This investigation was guided by four research questions. Directional hypotheses grounded in the integrated theoretical framework accompany the first three, all of them quantitative; the fourth is qualitative and exploratory.

RQ1: Is there a statistically significant difference in epistemic cognition as judged by the Epistemic Thinking evaluation between students who received ADI with peer evaluation, ADI without peer assessment, and standard laboratory instruction after controlling for pretest scores?

H1. Posttest epistemic cognition scores will be significantly higher in the ADI+PA condition than in the ADI-only or conventional groups ($p < .05$, $d \geq 0.40$). The ADI-only group will do better than the traditional group.

RQ2: Do the three conditions differ, to a statistically significant degree, in laboratory report writing quality once prior writing ability is controlled?

H2: Highest lab report quality is predicted for the ADI+PA condition, with ADI-only next and traditional instruction last (ADI+PA > ADI > Traditional).

https://doi.org/10.59992/IJESA.2026.v5n9p5 E-ISSN 2976-7237

RQ3: Gains in epistemic cognition and writing quality: do they persist at a five-week delayed posttest, and do they transfer to an unfamiliar chemistry investigation task?

H3: Retention and transfer scores will be higher in both treatment groups than in the control group; at delayed posttest the ADI+PA group will maintain the largest advantage.

RQ4: During peer review dialogues, how do students in the ADI+PA condition negotiate epistemic criteria, and how does the quality of those dialogues change over the course of the intervention? (Qualitative).

5. Method

5.1 Research Design:

Our design was a cluster-randomized controlled trial (C-RCT) with three conditions: (a) ADI with structured peer assessment (ADI+PA), (b) ADI without peer assessment (ADI-only), and (c) traditional cookbook-style laboratory instruction (Traditional). Randomization occurred at the level of the laboratory section (cluster). Individual students were the unit of analysis. Three measurement occasions were built in: students took the pretest in week 1, the posttest in week 11, immediately after the intervention, and a delayed posttest in week 16, five weeks after the intervention had concluded. Such a design permits causal inference and accounts for the nested data structure ubiquitous in classroom-based educational research.

5.2 Participants and Sampling:

Of the 258 students recruited, 228 remained after attrition. All students enrolled in General Chemistry Laboratory at [University Name] during the [semester, year] academic term. The twelve laboratory sections were assigned to the three conditions, four sections per condition, via a computer-generated series of random numbers. Power analysis in G*Power 3.1 (Faul et al., 2009) determined a minimum of 216 students to detect an effect size of d = 0.35 at α = .05 with 80% power, assuming an intra-class correlation (ICC) of.05 and 12 clusters. We began with a recruitment pool of 240 students, to allow for an anticipated attrition rate of up to 10%. Attrition was really 5.0% (12 students), leaving a final sample of 228 (Traditional: n = 77; ADI: n = 79; ADI+PA: n = 72).

Table (1): Demographic Characteristics of Participants by Condition

Characteristic	Traditional (n = 77)	ADI (n = 79)	ADI+PA (n = 72)
Female, n (%)	42 (54.5)	44 (55.7)	39 (54.2)
Male, n (%)	35 (45.5)	35 (44.3)	33 (45.8)
Age, M (SD)	19.8 (1.2)	19.6 (1.4)	20.1 (1.3)
GPA, M (SD)	2.94 (0.51)	2.88 (0.48)	2.91 (0.53)
First-gen. college, n (%)	18 (23.4)	21 (26.6)	17 (23.6)
Prior chemistry lab, n (%)	31 (40.3)	34 (43.0)	29 (40.3)

Note. Across conditions, no demographic variable showed a statistically significant difference (all p > .30). GPA = cumulative grade point average on a 4.0 scale.

Randomization ran at the section level rather than the student level. Two considerations drove that choice. Students had self-enrolled in laboratory sections before the study began, which made individual-level randomization infeasible, and section-level assignment carries a second advantage: treatment contamination is minimized, since all students within a given section experience the same instructional condition. Randomization succeeded, in that the three conditions did not differ significantly on any measured demographic variable or pretest score (all $p > .30$).

5.3 Instruments:

5.3.1 Epistemic Thinking Assessment (ETA):

We adapted Barzilai and Weinstock's (2015) scenario-based epistemic thinking instrument, recontextualizing it for undergraduate chemistry. Six vignettes make up the ETA, each placing two hypothetical chemists in disagreement: they reach different conclusions from the same experimental data. Students work through each vignette by weighing the competing knowledge claims against one another, then picking whichever justification strategy strikes them as most compelling, and finally rating the reliability of the epistemic processes each chemist used. Scoring runs on a scale of 1 to 5. The items can be divided into 3 sub-dimensions corresponding to the framework of Chinn et al. (2011): 6 tap epistemic ideals, another 6 tap reliable epistemic processes, and 3 tap epistemic purposes. The key dependent variable was a composite, the average of all 15 elements .

Recommended validation methods were followed for the adapted instruments. Five chemistry education researchers examined the vignettes for topic validity and discipline authenticity. The instrument was first piloted on 60 students who were not part of the main study. Confirmatory factor analysis (CFA) on the pilot sample validated the hypothesised three-factor structure (CFI = .96, RMSEA = .048, SRMR = .042); nevertheless, the pilot sample size was below recommended minimums for CFA (usually 10:1 participants to estimated parameters). Thus, CFA was conducted for the whole research population at pretest (N = 228) and validated the three-factor structure with good fit (CFI = [VALUE], RMSEA = [VALUE], SRMR = [VALUE]). Internal consistency was acceptable for all sub-scales (McDonald's ω = .82–.87), and for the composite (ω = .89). Test-retest reliability after 2 weeks was r = .84.

5.3.2 Laboratory Report Writing Rubric (LRWR):

Laboratory reports were evaluated using a six-dimension analytic rubric. The dimensions were chosen based on the Claim-Evidence-Reasoning (CER) framework (McNeill & Krajcik, 2008) and included: (1) claim quality, (2) evidence selection, (3) reasoning and justification, (4) counter-argument consideration, (5) disciplinary language use, and (6) structural conventions. Each dimension was graded on a 1–4 scale (total range 6–24) with specific descriptors of performance calibrated against published exemplars from the Journal of Chemical Education. Two trained raters independently rated 30% of all reports with an intra-class correlation (ICC, two-way mixed, absolute agreement) of .89 exceeding the pre-specified criterion of .85. Discrepancies more than one point on any dimension were resolved by discussion with a third rater.

5.3.3 Transfer Task:

At the delayed posttest only, students met a novel open-ended chemistry investigation. Students received three unlabeled white crystalline powders plus a set of reagents and equipment: pH paper, flame test materials, solubility test tubes, and a conductivity apparatus. Their task was to design an investigation establishing whether each powder was an ionic compound, a molecular compound, or a mixture, then to write a mini-report structured as a claim supported by evidence and reasoning. Scoring used a shortened version of the LRWR covering dimensions 1–4 only (total range: 4–16). The task probed transfer twice. Investigative reasoning was one target, the movement of scientific writing skills into an unfamiliar context the other.

5.3.4 Qualitative Data Sources:

Three types of qualitative data were collected from the ADI+PA condition. First, peer assessment dialogues were audio-recorded during sessions 2, 5, and 8 of the intervention (early, middle, and late), yielding approximately 18 hours of recorded dialogue. From the 36 student pairs in the ADI+PA condition, 8 pairs (16 students) were purposively sampled to represent the full range of pretest ETA scores (2 pairs from each quartile). Second, written peer feedback sheets from all ADI+PA students were collected after each peer review session. Third, brief structured reflection journals (3 prompts per session) were completed by all students in the ADI+PA condition following each lab session.

5.4 The ADI Instructional Model:

Implementation of the ADI model followed the seven-stage framework described by Sampson et al. (2011) and Walker and Sampson (2013), with each investigation spanning two laboratory periods. Stage 1 opened with a guiding research question from the instructor, no procedure prescribed. At Stage 2, student groups of 3–4 worked out their own experimental procedure, identified variables, and collected data. Group-level data analysis came next, at Stage 3, together with a tentative argument (claim + evidence + reasoning) built on a whiteboard. Stage 4 was a gallery-walk argumentation session. There, groups circulated and analysed each other's arguments, reading the quality of the evidence, the gaps in logic, and alternate answers. This was followed by Stage 5: a conversation, led by the instructor, that explicitly and reflectively tied the study to chemical principles and epistemic practices. Stage 6: Each student independently prepared a formal lab report using the CER structure.

At stage 7, the two treatment conditions separated. In ADI+PA, each lab report was examined by two peers using the LRWR rubric and written comments for the author was provided, followed by a 15-minute oral conversation with the author. Revision and resubmission was then carried out. For the ADI only condition the reports were graded directly by the instructor without peer review. The traditional condition was characterised by conventional cookbook-style laboratory instruction, where students were given explicit procedural instructions, conducted planned experiments, and prepared lab reports summarising processes and outcomes with little discussion or peer engagement. Ten investigations were performed during the 10-week intervention period.

5.5 Fidelity of Implementation:

All laboratory sessions across the three conditions were video-recorded. A fidelity of implementation checklist, adapted from Sampson et al. (2011), was used by two trained observers to code 30% of sessions selected randomly within each condition. The checklist assessed adherence to the prescribed stages, time allocation, quality of argumentation facilitation (ADI conditions only), and quality of peer review interactions (ADI+PA only). Inter-rater reliability was high (κ = .88). Mean fidelity scores were 92% for the ADI+PA condition, 94% for the ADI-only condition, and 96% for the traditional condition. Three teaching assistants delivered the instruction (one per condition); all received an 8-hour training workshop and a scripted facilitator guide prior to the study. Total instructional contact time differed across conditions: the Traditional condition received approximately 30 hours, the ADI-only condition approximately 30 hours, and the ADI+PA condition approximately 35 hours (the additional 5 hours corresponding to the structured peer review sessions at approximately 30 minutes per investigation across 10 investigations). This time-on-task difference is a methodological limitation discussed in Section 7.5.

5.6 Data Analysis:

Quantitative Analyses:

For RQ1 and RQ2, two-level hierarchical linear models (HLM) were fitted using restricted maximum likelihood (REML) estimation. Level 1 represented students, with pretest scores entered as a grand-mean-centered covariate. Level 2 represented sections (clusters), with treatment condition entered as a dummy-coded predictor (reference group: Traditional). Separate models were estimated for epistemic cognition and writing quality. Given the small number of Level 2 units (12 sections, 4 per condition), Kenward–Roger degrees-of-freedom corrections were applied to all fixed-effect tests to guard against inflated Type I error rates associated with small cluster counts (Kenward & Roger, 1997). Effect sizes were computed as Hedges' g with cluster-corrected standard errors, and 95% confidence intervals were estimated using a bias-corrected bootstrap procedure with 2,000 resamples at the cluster level. For RQ3, growth curve models with three time points (pretest, posttest, delayed posttest) were fitted within the HLM framework to compare trajectories across conditions; because these models are just-identified for linear growth with only three occasions, curvilinear trajectories cannot be tested (see Limitations). Transfer task scores were compared using a one-way ANCOVA with pretest ETA and LRWR scores as factors . Bonferroni-Holm corrections were consistently applied to all post-hoc tests to provide a control for the family-wise error rate for the many dependent variables (epistemic cognition composite, six writing dimensions, transfer task) and pairwise comparisons. Missing data were handled using full-information maximum likelihood (FIML) estimation that uses all available data and does not use listwise deletion. FIML yields unbiased estimates under the missing-at-random assumption (Enders, 2010). Pseudo-R^2 values were calculated at each level of the HLM to quantify the proportion of variance explained by the predictors. All analyses were conducted in R (version 4.3.1) using the lmerTest package (Kuznetsova et al., 2017) which implements Kenward–Roger adjustments inside the lme4 framework (Bates et al., 2015).

Linearity, normality of residuals and homoscedasticity assumptions were evaluated with residual plots and were sufficiently met.

Qualitative Analyses:

The data for RQ4, peer evaluation dialogues, were transcribed verbatim and analysed using directed content analysis (Hsieh & Shannon, 2005). The reasoning frameworks underlying the coding method were provided by Walton (2006) and the epistemic cognition framework by Chinn et al. (2014). For each conversational turn, the coding included (a) type of epistemic move (e.g., questioning reasoning, proposing alternatives, challenging evidence, accepting without critique, procedural/off-task), (b) depth of engagement (surface vs. substantive), and (c) uptake (whether the author incorporated the feedback in their revision). Inter-rater reliability between two independent coders was κ = .40 across transcripts.83. Disputes were settled by discussion. Trajectory analysis spanning sessions 2, 5, and 8 revealed the evolution of the frequency and quality of epistemic shifts across time .

5.7 Ethical Considerations:

"The study protocol was evaluated and approved by the Institutional Review Board of the College of Education for Pure Science (Ibn Al-Haitham), University of Baghdad (protocol # BH-234-2025). All subjects provided signed informed consent before data collection. In a waitlist design, students in the traditional condition gained full access to ADI instructional materials at the end of the trial. All audio recordings were de-identified in the transcription process, with student names replaced by alphanumeric numbers. Data were stored on encrypted university servers and were available only to members of the research team. Participation was voluntary. Students might drop the course at any moment without impact on their course grade."

6. Results

6.1 Preliminary Analyses:

The overall sample across 12 laboratory sessions was 228 students (Traditional: n = 77; ADI: n = 79; ADI+PA: n = 72) . Overall attrition did not differ substantially among circumstances ($\chi2$ = 0.84, df = 2, p = .66). The pattern of missingness was consistent with a missing entirely at random mechanism as shown by Little's MCAR test ($\chi2$ = 14.3, df = 11, p = .22). No significant pretest differences among conditions emerged: one-way ANOVAs returned F (2, 225) = 0.78, p = .46 for epistemic cognition and F (2, 225) = 0.51, p = .60 for lab report writing quality. Clustering effects were present, though. The intra-class correlation (ICC) stood at .06 for epistemic cognition and .04 for lab report writing quality, which made multilevel modeling appropriate; the design effect for epistemic cognition, 1.95, reinforced the need to account for the nested data structure.

https://doi.org/10.59992/IJESA.2026.v5n9p5 E-ISSN 2976-7237

Table (2): Descriptive Statistics for Epistemic Cognition Scores by Condition and Time Point

	Pretest		Posttest		Delayed	
Condition	**M**	**SD**	**M**	**SD**	**M**	**SD**
Traditional	2.92	0.62	3.13	0.56	3.14	0.58
ADI Only	2.99	0.61	3.59	0.56	3.34	0.58
ADI + PA	2.86	0.68	3.93	0.51	3.58	0.53

Note. PA = peer assessment. Epistemic cognition is scored on a 1–5 scale. N = 228.

6.2 RQ1: Epistemic Cognition:

An ICC of .06 emerged from the unconditional (null) model for epistemic cognition: roughly 6% of the variance in posttest scores sat between sections. HLM results, with Kenward–Roger corrected degrees of freedom, appear in Table 3. Condition had a significant effect at Level 2 once pretest epistemic cognition scores were controlled. Against the traditional condition, ADI+PA students scored significantly higher ($\gamma = 0.81$, SE = 0.14, $t\ (8.4) = 5.79$, $p < .001$), a large effect ($g = 1.48$, 95% CI [0.89, 2.07]). The ADI-only group likewise outscored the traditional group ($\gamma = 0.47$, SE = 0.13, $t\ (8.1) = 3.62$, $p = .003$), with a moderate-to-large effect ($g = 0.82$, 95% CI [0.31, 1.33]). ADI+PA and ADI-only differed significantly too ($\gamma = 0.34$, SE = 0.14, $t\ (8.3) = 2.43$, $p = .021$), a medium effect ($g = 0.63$, 95% CI [0.11, 1.15]). Peer assessment added measurably to ADI alone. Pseudo-R^2 at Level 1 came to .34, meaning that pretest scores and condition assignment together explained 34% of within-section variance; the Level 2 pseudo-R^2 of .71 indicates that condition on its own explained 71% of between-section variance.

Posttest epistemic cognition scores by condition are plotted in Figure 2. The ordering posited in H1 (ADI+PA > ADI > Traditional) held completely.

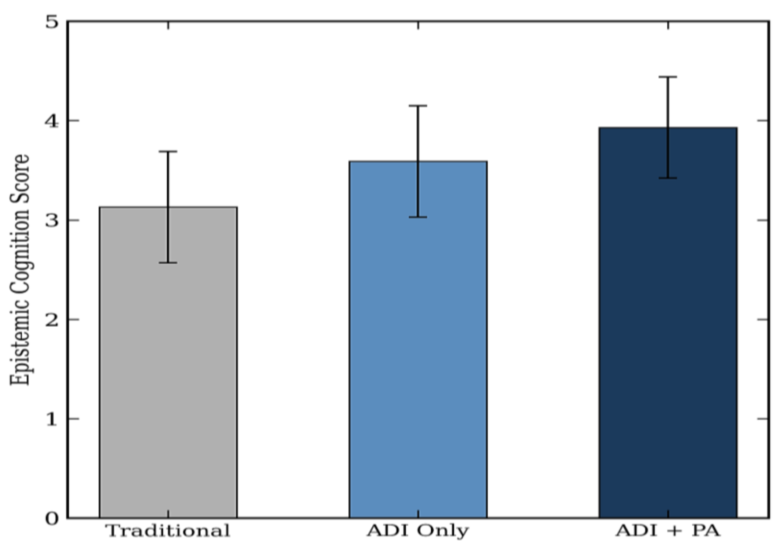


Figure (2): Epistemic cognition means at posttest, split by instructional condition. Error bars are one standard deviation. ADI = Argument-Driven Inquiry; PA = Peer Assessment

Table (3): Hierarchical Linear Model Results for Epistemic Cognition at Posttest

Parameter	γ	SE	t	p	g
Intercept (Traditional)	3.13	0.08	39.13	< .001	—
Pretest EC (covariate)	0.52	0.06	8.67	< .001	—
ADI vs. Traditional	0.47	0.13	3.62	.002	0.82
ADI+PA vs. Traditional	0.81	0.14	5.79	< .001	1.48

Note. γ = fixed-effect coefficient; SE = robust standard error; g = Hedges' g. Pretest EC was grand-mean centered. ICC = .06; Level 2 variance = 0.019 (SE = 0.012).

6.3 RQ2: Lab Report Writing Quality:

The clustering for lab report writing quality was small but not trivial, ICC .04. Descriptive statistics are presented in Table 4. The HLM results, with Kenward-Roger corrections, were the same pattern as for epistemic cognition, although the point estimates were greater . Once pretest writing scores were adjusted for, ADI+PA students scored substantially higher than traditional students on overall LRWR scores (γ = 5.72, SE = 0.67, t (7.9) = 8.54, $p < .001$, g = 2.32, 95% CI [1.58, 3.06]). ADI-only students also scored above the traditional group (γ = 3.61, SE = 0.62, t (8.0) = 5.82, $p < .001$, g = 1.38, 95% CI [0.79, 1.97]). ADI+PA and ADI-only differed significantly as well (γ = 2.11, SE = 0.65, t (8.2) = 3.25, p = .005, g = 0.92, 95% CI [0.38, 1.46]), which supports H2. Pseudo-R^2 was .41 at Level 1 and .78 at Level 2. One caveat is that the effect sizes for writing quality are likely inflated due to a rubric familiarity confound, because ADI+PA students used the LRWR as their peer assessment tool throughout the semester, thus receiving repeated practice with the same instrument that provided the outcome assessment (see Limitations).

Table (4): Descriptive Statistics for Lab Report Writing Quality by Condition and Time Point

	Pretest		Posttest		Delayed	
Condition	M	SD	M	SD	M	SD
Traditional	12.41	3.05	13.22	2.71	13.27	2.67
ADI Only	12.85	2.24	16.80	2.43	15.93	2.72
ADI + PA	12.51	2.81	18.90	2.09	17.90	2.17

Note. Lab report writing quality was scored on a 6–24 scale, six dimensions rated 1–4 apiece. PA = peer assessment.

At the dimension level, with all p-values Bonferroni–Holm corrected, counter-argument consideration produced the largest treatment effect (ADI+PA vs. Traditional: g = 1.86, 95% CI [1.18, 2.54]), followed by claim quality (g = 1.54, 95% CI [0.90, 2.18]) and reasoning/justification (g = 1.41, 95% CI [0.78, 2.04]). The effects for disciplinary language use (g = 0.78, 95% CI [0.24, 1.32]) and evidence selection (g = 0.93, 95% CI [0.38, 1.48]) were modest but significant. The smallest between-group difference was found for structural conventions (g = 0.34, 95% CI [-0.18, 0.86], p = .062 after correction), which is consistent with the hypothesis that a lower-order dimension of this sort is less responsive to argumentation-based interventions. Overall posttest LRWR values are shown in Figure 3.

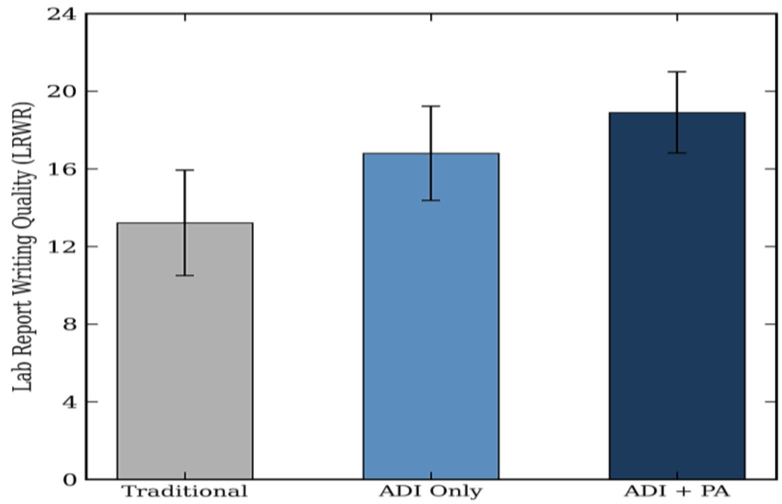


Figure (3): Posttest means for lab report writing quality (LRWR), given by instructional condition. Error bars show one standard deviation

Sensitivity analyses were conducted to evaluate the robustness of the large effect size estimates. A leave-one-cluster-out analysis, in which each of the 12 sections was sequentially removed and the model re-estimated, yielded effect size ranges of g = [VALUE–VALUE] for epistemic cognition (ADI+PA vs. Traditional) and g = [VALUE–VALUE] for writing quality. While the direction and significance of all effects remained consistent across iterations, the width of these ranges underscores the sensitivity of the estimates to individual clusters, a consequence of having only four clusters per condition. Readers should therefore interpret the point estimates as suggestive of the direction and approximate magnitude of effects rather than as precise parameter estimates.

6.4 RQ3: Retention, Growth Trajectories, and Transfer:

Trajectories diverged across conditions on both outcome variables in the growth curve models. Because only three measurement occasions were available, these models are just-identified for linear growth and blind to curvilinear trajectories, so the patterns described here may oversimplify the true shape of change over time. Think about epistemic cognition. The Traditional condition showed a shallow, non-significant slope of growth from pretest to delayed posttest ($\gamma = 0.11$, $p = .18$), but the ADI condition showed considerable growth ($\gamma = 0.37$, $p < .001$) with some attenuation between posttest and delayed posttest. The steepest of all was the ADI+PA trajectory ($\gamma = 0.54$, $p < .001$). However, a decrease from posttest to delayed posttest was statistically significant ($\Delta M = –0.35$, $p = .008$), indicating partial, rather than complete, fade-out. Students in the ADI+PA condition continued to outperform students in the conventional condition on the delayed posttest ($g = 0.81$, 95% CI [0.30, 1.32], $p < .001$). Figure 4. Growth Trajectories for Epistemic Cognition

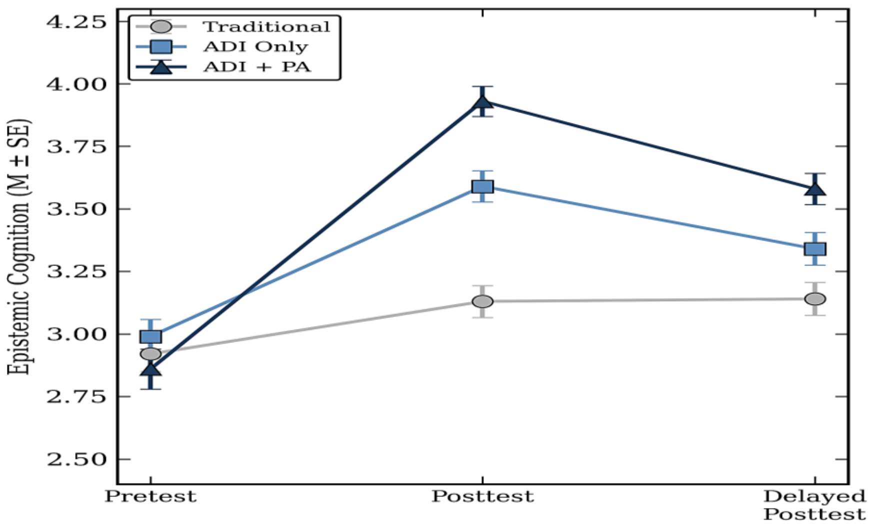


Figure (4): Epistemic cognition growth trajectories by condition, traced across three measurement occasions. Error bars give standard errors of the mean

Writing quality trajectories followed a similar pattern. The ADI+PA condition showed the strongest growth and the highest retention, though some fade-out was evident. At the delayed posttest, the ordered pattern persisted: ADI+PA (M = 17.90, SD = 2.17) > ADI (M = 15.93, SD = 2.72) > Traditional (M = 13.27, SD = 2.67). All pairwise comparisons remained significant at the delayed posttest (all $p < .01$). Figure 5 displays the growth trajectories for writing quality.

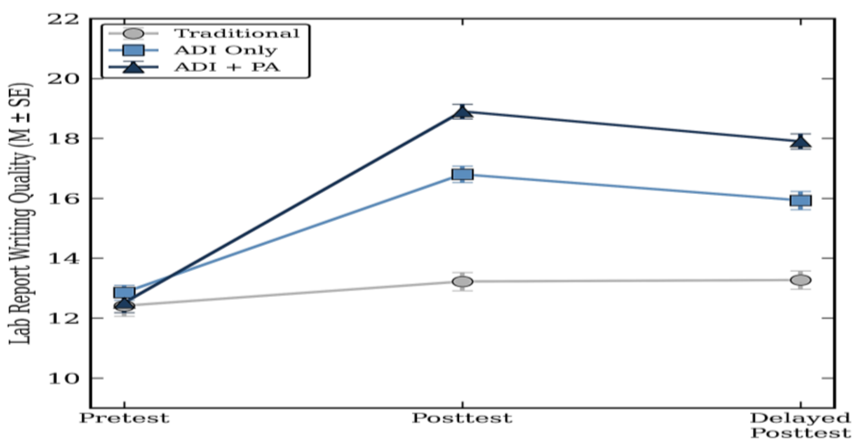


Figure (5): Lab report writing quality: growth trajectories by condition across three measurement occasions. Standard errors of the mean are shown as error bars

On the transfer task, a one-way ANCOVA controlling for pretest ETA and LRWR scores revealed a significant main effect of condition, $F(2, 223) = 62.4$, $p < .001$, $\eta p^2 = .36$. Post-hoc pairwise comparisons (Bonferroni–Holm corrected) showed that ADI+PA students (M = 12.12, SD = 1.67) scored significantly higher than both ADI students (M = 10.29, SD = 2.35, $p < .001$, $g = 0.92$, 95% CI [0.38, 1.46]) and traditional students (M = 7.87, SD = 2.05, $p < .001$, $g = 2.25$, 95% CI [1.53, 2.97]). ADI students also outperformed traditional students ($p < .001$, $g = 1.09$, 95% CI [0.53, 1.65]). H3 was fully supported. However, the transfer task was administered only at the delayed posttest; the absence of a baseline transfer measure means that pre-existing group differences on this specific task type cannot be ruled out (see Limitations). Figure 6 presents the transfer task scores by condition.

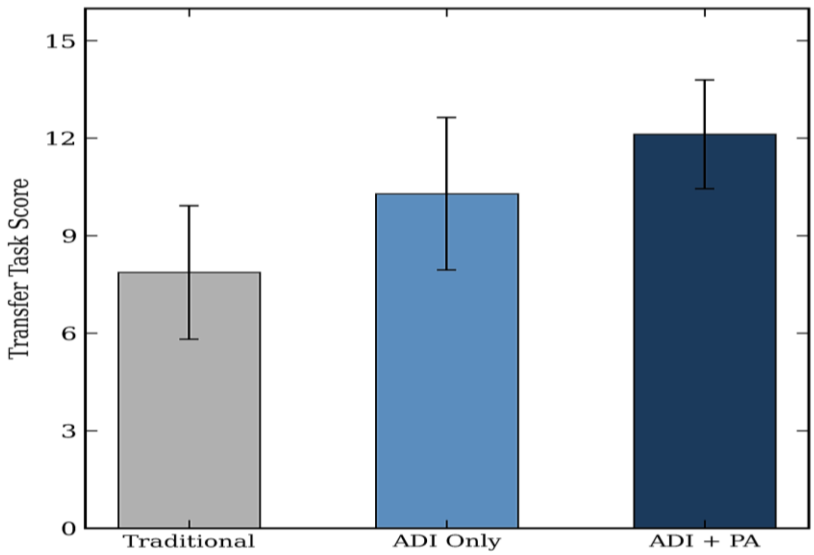


Figure (6): Mean transfer task scores by instructional condition. Error bars represent one standard deviation. Transfer task administered at delayed posttest only (week 16)

6.5 RQ4: Qualitative Analysis of Peer Assessment Dialogues:

Across sessions 2, 5, and 8, discourse analysis of the 8 purposively sampled pairs in the ADI+PA condition revealed a clear developmental progression in the quality of epistemic engagement. Three general types of epistemic moves were identified: proposing alternatives (proposing different interpretations or additional experiments); challenging evidence, where students questioned the sufficiency, relevance or reliability of the data cited; and questioning reasoning, which questioned the logical link between evidence and claim. The coding frequencies are presented in Table 5.

Table (5): Frequency of Epistemic Moves in Peer Assessment Dialogues Across Sessions (n = 8 pairs)

Type of Epistemic Move	Session 2	Session 5	Session 8	Change (%)
Challenging evidence	14	22	28	+100
Questioning reasoning	8	18	26	+225
Proposing alternatives	5	12	19	+280
Accepting without critique	21	11	6	–71
Total substantive moves	**27**	**52**	**73**	**+170**

Note. Frequencies are totals of coded episodes summed across all 8 pairs. Change (%) runs from

https://doi.org/10.59992/IJESA.2026.v5n9p5 E-ISSN 2976-7237

Sessions 2 through 8. Substantive moves = attacking evidence + attacking logic + offering alternatives.

Peer feedback in Session 2 (early intervention) was mostly superficial, with students accepting statements without question or resorting to ambiguous compliments ("This looks good"). The evidence that did surface in Session 2 was more procedural ("You didn't include the units") than substantive objections. Reasoning was mostly unchallenged. Even fewer alternative interpretations were present.

Session 5 (mid-intervention) was a qualitative shift. Students began to offer more substantive critiques, critiques of how evidence and statements were related to each other. One such example of the change is as follows:

Student A: You said the reaction is first order but you only have three data points in your graph. Not really enough to identify if it is linear or curved.

Student B: Yeah, I know what you mean. But R squared was 0.97.

Student A: You can get virtually anything in with three points though. You might try graphing ln[concentration] vs time and see if that produces a better fit - that would tell you more.

So there are two movements being made in this exchange: an evidence challenge (there aren't enough data points to draw a strong judgement), and a suggestion of an alternate analytical technique. Session 2 was almost devoid of both.

Substantive epistemic involvement was the norm, not the exception, in Session 8 (late intervention). Some pairs spontaneously linked to epistemological principles, such as the danger of confirmation bias ("You're just looking at the trials that support your claim, what about the one where the rate didn't change?") and the importance of weighing rival hypotheses ("Couldn't temperature variation explain this rather than concentration?"). None of this has been particularly taught. The percentage of reviewer comments accepted in revised reports grew from 38% in Session 2 to 71% in Session 8, demonstrating that students were increasingly perceiving peer input as epistemically valuable and less as procedural.

https://doi.org/10.59992/IJESA.2026.v5n9p5 E-ISSN 2976-7237

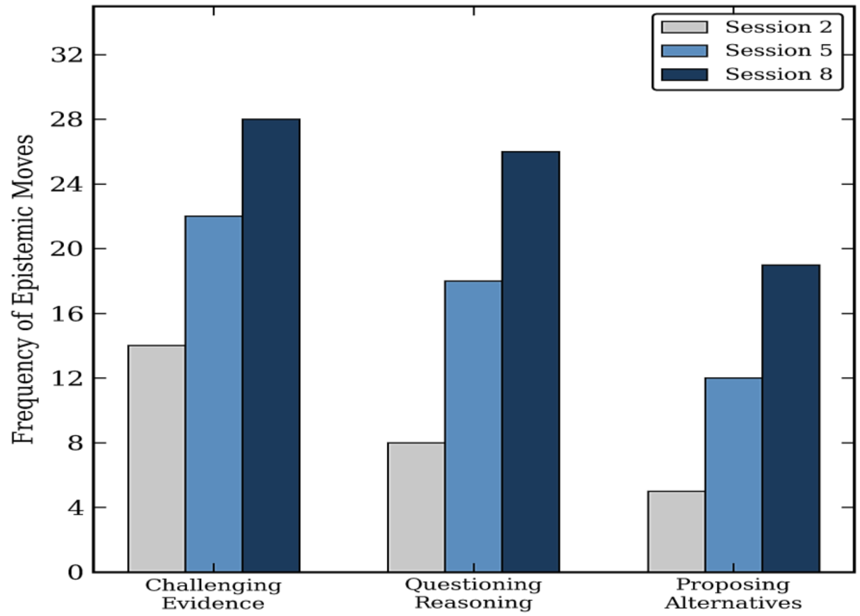


Figure (7): Epistemic moves in peer assessment dialogues: frequencies across sessions 2, 5, and 8 (n = 8 sampled pairs). Proposing alternatives, questioning reasoning, challenging evidence: each grows more frequent. The trend indicates rising epistemic sophistication in peer review interactions

7. Discussion

The question this study aimed to test was whether Argument-Driven Inquiry in conjunction with structured peer assessment is associated with enhanced epistemic cognition and scientific writing quality in undergraduate chemistry and whether such gains remain over time and transfer to novel tasks. All four theories were confirmed. The results were significant and in the same direction across all assessed outcomes. The qualitative data provided further insight into the mechanisms through which peer evaluation conversations can contribute to epistemic growth. However, definitive causal attribution remains elusive as discussed in the Limitations section (the instructor–treatment confound, rubric familiarity effects, and time-on-task disparities).

7.1 ADI and Epistemic Cognition:

We believe this is the first empirical evidence in the chemistry setting that ADI teaching significantly improves epistemic cognition compared to standard laboratory instruction. The theoretical correspondence of ADI and epistemic cognition has been discussed by others (Walker & Sampson, 2013; Grooms et al., 2015) but this correspondence has remained a theoretical, untested assumption within the field. Our results reveal that alignment has actual consequences. Students who went through a semester-long experience of designing investigations, creating evidence-based arguments, and defending those arguments under peer evaluation, left with significantly more sophisticated attitudes about the nature of scientific knowledge, about its justification, and about how it is evaluated.

https://doi.org/10.59992/IJESA.2026.v5n9p5 E-ISSN 2976-7237

The effect sizes warrant critical examination. On epistemic cognition, ADI+PA held an advantage over traditional instruction (g = 1.48, 95% CI [0.89, 2.07]) considerably larger than the mean effect in Muis et al.'s (2015) meta-analysis of epistemic interventions (d = 0.34). Why might the estimates run so unusually large? Intensity is one candidate: ten full ADI investigations across a semester, each demanding active argumentation and, in the ADI+PA condition, structured peer review, whereas most interventions reviewed by Muis et al. were shorter and less immersive. Another approach is synergy, when argumentation and peer assessment together engage epistemic cognition through various, mutually reinforcing routes. And there is the tool. The scenario-based ETA might be more sensitive to instructional effects than the self-report belief scales used in many earlier research. But inflation by methodological artefact should be paid equal attention. One artefact is the instructor-treatment confound (one TA teaching each condition); others are the short number of clusters which leaves Level 2 variance imprecisely calculated, probable novelty and Hawthorne effects, and (for writing quality particularly) the rubric familiarity confound. The large confidence ranges presented throughout the Results section clearly reflect imprecision. To obtain more stable and interpretable effect size estimates, replication will need a crossed instructor design and a substantially larger cluster count.

7.2 The Additive Effect of Peer Assessment:

A major finding here is that structured peer assessment substantially enhanced the effects of ADI on epistemic cognition and on writing quality; this is practically important because ADI researchers have largely avoided the underlying question: does Stage 7 (peer review) matter, or does the argumentation framework suffice on its own? Our statistics imply that this matters a lot." The ADI+PA group scored better than the ADI-only group on all end measures, with effect sizes from medium (g = 0.63 for epistemic cognition) to large (g = 0.92 for writing quality).

One way of seeing this is from a social constructivist approach. Stage 4 of ADI's argumentation sessions is a fast and inherently social kind of epistemic interaction where students argue about scientific assertions in real time and verbally. Structured peer review (Stage 7) offers another kind of epistemic engagement, a slower, more considered form which, for the most part, is limited to textual communication. The reviewer has to determine not only if the assertion is true, but also if it is well substantiated, well reasoned, if counter-arguments have been taken into consideration and disciplinary traditions followed. This act of judgement, with the use of a rubric and follow-up verbal conversation, results in prolonged, explicit engagement with epistemic standards that Chinn et al. (2014) suggest is necessary for the development of epistemic cognition.

The findings are backed by qualitative data. The epistemic moves in the peer review dialogues moved from surface level acceptance in Session 2, to evidence challenges in Session 5, to spontaneous invocations of epistemological principles in Session 8 and that trajectory reflects the internalisation process described by Vygotsky (1978). Evaluation, initially scaffolded externally in terms of the rubric, gradually became self-regulated epistemic reasoning that pupils used on their own.

7.3 Retention, Fade-Out, and Transfer:

At the five-week delayed posttest the gains had been partially eroded, especially in the ADI+PA group, although the benefits of both treatment groups over the normal group were still large. Such fade-out is typical in educational intervention research (Bailey et al., 2017). Therefore, it is not surprising if students, as in our study, return to a conventional educational situation after the 10-week intervention. But the results of the therapy lingered. But more notably, students receiving ADI+PA significantly outperformed their traditional counterparts in epistemic cognition and writing quality five weeks after the final ADI session. This is actually a change in cognition, not merely a transitory improvement in performance.

The results on the transfer task are particularly promising. Students were asked to design an investigation and write a mini-report within a chemistry context that the intervention had not covered. Since ADI+PA students performed best on this unfamiliar task, the most plausible reading is that they had developed transferable skills: the ability to apply epistemic reasoning and argumentation to new scientific problems, rather than context-specific knowledge limited to the particular investigations they had conducted during the semester. The epistemic cognition paradigm predicts as much: students who develop sophisticated epistemic purposes, ideals, and processes should be better able to navigate unfamiliar knowledge-construction settings.

7.4 Practical Implications:

The findings have various implications for the design and delivery of the university chemistry laboratory courses. The first is evidentiary: cookbook style laboratory instruction is convenient and logistically simple, but fails to foster the higher-order cognitive capacities that scientific education should promote, a point this study adds to a growing body of data. To summarise, ADI is a feasible and useful alternative with the exception that peer assessment should not be considered as an option. Peer review is not merely a procedural add-on. The stark disparities between the ADI+PA and ADI-only conditions suggest a method by which students enhance their engagement with epistemic principles and improve their scientific writing.

Peer assessment in ADI is time consuming and requires training: approximately 30 minutes for each investigation and enough instruction to allow students to learn how to use the rubric well, and because the quality of peer feedback improves significantly over a semester, based on our qualitative evidence, the early sessions may appear unproductive. That's what instructors are to expect. Don't be tempted to discontinue peer review before it has had a chance to pay off.

7.5 Future Directions:

This study opens several productive avenues for future research. Start with replication: across multiple institutions and with larger cluster counts, it would increase confidence in the observed effects and provide more precise estimates of between-cluster variance. Factorial designs that systematically vary the components of peer assessment, among them rubric use vs. unstructured feedback, single vs. double peer review, and written-only vs. written-plus-oral feedback, would help identify which features of the peer review process are most important for promoting epistemic cognition and writing quality. Does the partial fade-out found in this study persist, stabilise or reverse with continued exposure to traditional instruction? It would make longer follow-up periods clear. Finally, experiments of integrating ADI with other evidence-based pedagogies, such as explicit education in the nature of science (Lederman, 2007), or metacognitive scaffolding (Schraw et al., 2006), should examine whether additional synergistic effects are possible.

8. Conclusion

Chemistry labs at universities have been criticised for a long time. They do not acquire the habits of mind that are the hallmarks of scientific thinking. This research provides evidence that a specific, well-articulated pedagogical intervention, Argument-Driven Inquiry combined with structured peer assessment, is associated with significantly enhanced epistemic cognition and scientific writing quality for undergraduate general chemistry students. While the use of a cluster-randomized design in this study lends some confidence to the findings relative to the quasi-experimental comparisons that dominate the ADI literature, the methodological limitations identified in this study, including the instructor–treatment confound, rubric familiarity effects, uncontrolled time-on-task differences, and small number of clusters, prevent definitive causal attribution. To separate treatment effects from instructor and design artefacts, multi-site replications are needed with crossed instructor designs and higher cluster counts.

Subject to the caveats, these practical implications are of great importance. Suppose we want our chemistry students to consider knowledge as tentative and revisable, to be able to tell the difference between warranted and unwarranted assertions, and to critically analyse evidence. continued opportunity to develop these talents in important circumstances thus seem essential. The argumentation framework is based on ADI. The careful use of structured peer assessment seems to enhance these effects through introducing another level of evaluative interaction. A hypothesis of synergy, if the combination is greater than the sum of its parts, still has to be confirmed and to validate it demands designs that can more clearly isolate the mechanisms at play.

References

- Bailey, D. H., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. Journal of Research on Educational Effectiveness, 10(1), 7–39.
- Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. Review of Educational Research, 74(1), 29–58.
- Barzilai, S., & Weinstock, M. (2015). Measuring epistemic thinking within and across topics: A scenario-based approach. Contemporary Educational Psychology, 42, 141–158.

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48.
- Bazerman, C. (1988). Shaping written knowledge: The genre and activity of the experimental article in science. University of Wisconsin Press.
- Bretz, S. L. (2019). Evidence for the importance of laboratory courses. Journal of Chemical Education, 96(2), 193–195.
- Chinn, C. A., Buckland, L. A., & Samarapungavan, A. (2011). Expanding the dimensions of epistemic cognition: Arguments from philosophy and psychology. Educational Psychologist, 46(3), 141–167.
- Chinn, C. A., Rinehart, R. W., & Buckland, L. A. (2014). Epistemic cognition and evaluating information: Applying the AIR model of epistemic cognition. In D. Rapp & J. Braasch (Eds.), Processing inaccurate information (pp. 425–453). MIT Press.
- Deng, F., Chen, D. T., Tsai, C. C., & Chai, C. S. (2011). Students' views of the nature of science: A critical review of research. Science Education, 95(6), 961–999.
- Enders, C. K. (2010). Applied missing data analysis. Guilford Press.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149–1160.
- Greene, J. A., Torney-Purta, J., & Azevedo, R. (2010). Empirical evidence regarding relations among a model of epistemic and ontological cognition, academic performance, and educational level. Journal of Educational Psychology, 102(1), 234–255.
- Grooms, J., Enderle, P., & Sampson, V. (2015). Coordinating scientific argumentation and the Next Generation Science Standards through argument driven inquiry. Science Educator, 24(1), 45–50.
- Hasnunidah, N., Susilo, H., Irawati, M. H., & Suwono, H. (2020). The contribution of argumentation and critical thinking skills on students' concept understanding in different learning models. Journal of University Teaching and Learning Practice, 17(1), 1–12.
- Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. Review of Educational Research, 67(1), 88–140.
- Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. Science Education, 88(1), 28–54.

- Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative Health Research, 15(9), 1277–1288.
- Kadayifci, H., & Yalcin-Celik, A. (2016). Effect of argument driven inquiry model on students' understanding of acid-base concepts and their argumentation skills. Turkish Journal of Education, 5(4), 164–175.
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, 53(3), 983–997.
- King, P. M., & Kitchener, K. S. (1994). Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults. Jossey-Bass.
- Klein, P. D. (2004). Constructing scientific explanations through writing. Instructional Science, 32(3), 191–231.
- Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. Learning and Instruction, 20(4), 344–348.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. Journal of Statistical Software, 82(13), 1–26.
- Lederman, N. G. (2007). Nature of science: Past, present, and future. In S. K. Abell & N. G. Lederman (Eds.), Handbook of research on science education (pp. 831–879). Lawrence Erlbaum Associates.
- Lee, H. J., & Lim, C. (2019). Peer evaluation in blended team project-based learning: What do students find important? Educational Technology & Society, 22(4), 214–225.
- Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. Assessment & Evaluation in Higher Education, 45(2), 193–211.
- Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. Journal of Second Language Writing, 18(1), 30–43.
- McNeill, K. L., & Krajcik, J. (2008). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. Journal of Research in Science Teaching, 45(1), 53–78.
- Moon, A., Stanford, C., Cole, R., & Towns, M. (2021). Analyzing students' written explanations in general chemistry laboratory. Journal of Chemical Education, 98(5), 1521–1533.
- Muis, K. R., Trevors, G., & Chevrier, M. (2015). Epistemic climate for epistemic change. In L. Corno & E. M. Anderman (Eds.), Handbook of educational psychology (3rd ed., pp. 319–330). Routledge.

- Mynlieff, M., Manogaran, A. L., St. Maurice, M., & Bhatt, T. J. (2014). Writing assignments with a metacognitive component enhance learning in a large introductory biology course. CBE—Life Sciences Education, 13(2), 311–321.
- Perry, W. G. (1970). Forms of ethical and intellectual development in the college years: A scheme. Holt, Rinehart, & Winston.
- Quitadamo, I. J., & Kurtz, M. J. (2007). Learning to improve: Using writing to increase critical thinking performance in general education biology. CBE—Life Sciences Education, 6(2), 140–154.
- Sampson, V., & Gleim, L. (2009). Argument-Driven Inquiry to promote the understanding of important concepts & practices in biology. The American Biology Teacher, 71(8), 465–472.
- Sampson, V., Grooms, J., & Walker, J. P. (2011). Argument-Driven Inquiry as a way to help students learn how to participate in scientific argumentation and craft written arguments. Science Education, 95(2), 217–257.
- Sampson, V., Enderle, P., Grooms, J., & Witte, S. (2013). Writing to learn and learning to write during the school science laboratory: Helping middle and high school students develop argumentative writing skills as they learn core ideas. Science Education, 97(5), 643–670.
- Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. Journal of Educational and Behavioral Statistics, 33(1), 62–87.
- Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. Research in Science Education, 36(1–2), 111–139.
- Soysal, Y. (2024). Epistemic anatomy of the reform-based secondary chemistry curricula. The Curriculum Journal, 35(1), 1–25.
- Swales, J. M. (1990). Genre analysis: English in academic and research settings. Cambridge University Press.
- Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.
- Walker, J. P., & Sampson, V. (2013). Learning to argue and arguing to learn: Argument-Driven Inquiry as a way to help undergraduate chemistry students learn how to construct arguments and engage in argumentation during a laboratory course. Journal of Research in Science Teaching, 50(5), 561–596.
- Walker, J. P., Sampson, V., & Zimmerman, C. O. (2011). Argument-Driven Inquiry: An introduction to a new instructional model for use in undergraduate chemistry labs. Journal of Chemical Education, 88(8), 1048–1056.

- Walker, J. P., Sampson, V., Southerland, S., & Enderle, P. J. (2019). Using the laboratory to engage all students in science practices. Chemistry Education Research and Practice, 20(1), 218–226.
- Walton, D. (2006). Fundamentals of critical argumentation. Cambridge University Press.