Introducing Academically Low-Performing Young Science Students to Practices of Science

Toi Sin Arvidsson, Columbia University, tya2102@tc.columbia.edu
Deanna Kuhn, Columbia University, dk100@columbia.edu

Abstract: Our objective was to engage low-achieving young adolescents in activities introducing them to practices of science. These extended beyond control of variables to include attribution and prediction in coordination of multiple variables affecting an outcome, as well as argument and counterargument in advancing and challenging claims. Social science content was used to help students see the purpose and value of scientific practices. The objective was largely met, as evidenced by two 6th grade classes (n = 49), both outperforming a control group (n = 23). Although students engaged successfully in argument and counterargument, less successful was meta-level reasoning about argumentation and nature of science. Importantly in its practical implications, one intervention group showed less gain in 10 45-min whole-class sessions than did the other group who engaged as pairs in the same sequence of activities over an average of six 24-min individualized sessions, suggesting the greater efficacy of individualized engagement.

Keywords: scientific thinking, scientific practice, low-achieving populations, multivariable reasoning, argumentation

Introduction

Attracting a broad range of students to the study of science has become an objective of increasing concern. To become educated in the practice of science, students need to engage in communities of scientific inquiry in which they develop shared goals, methods, norms, and, ultimately, it is hoped, values. In so doing they become participants in shared practices that are part of larger disciplinary norms and traditions they come to recognize and appreciate (Sandoval, 2014).

Implementing these objectives of science education are a tall order. *Science Education* editor John Rudolph notes that such objectives in fact go back as far as Dewey and that the problem lies not with these educational goals but with schools' seeming inability to devise workable school experiences to achieve them. Adding to the challenge is the restricted way in which scientific practice has been defined, both in the K-12 classroom and in educational research. Science practice typically has been taken to mean use of the scientific method, which in turn has been regarded as the design and analysis of a controlled experiment. Moreover, the experiment is a univariable one, its essence being the control (by holding constant) of variables (COV) in order to identify the effect of a single variable on an outcome. In the real world, in contrast, outcomes are most often caused not by a single cause but by multiple factors acting in concert, a fact that practicing scientists are well aware of and take into account in both their theoretical models and empirical investigations. The univariable logic and execution of COV represents at most one narrow slice of authentic scientific inquiry, and the most recent writing on developing children's competency in science emphasizes involving students not in acquisition of a tool kit of discrete skills such as COV but rather in the practice of science as an authentic, integrated whole.

This is the approach we have sought to implement in the work described here. Our approach is focused on getting students to experience that the methods of science have purpose that makes sense to them and hence are of value. Most important to the approach implemented here, then, is that students' activity be situated in the context of what students will see as a meaningful purpose and goal. In two initial studies, urban middle-school students addressed, for example, the topic of juvenile delinquency, among other social science topics. The reason for employing social science topics such as teen crime goes beyond the fact that such topics are ones our student population are familiar with. Although students will feel they already know something about them, they likely will not know that such topics are the stuff of science. What better way, then, to get them to see its power and relevance? In the course of such activities, students come to see how their (and others') beliefs about a phenomenon like teen crime are subject to influence by means of application of a scientific method.

In this study, we worked with students from a similar population over an extended period and hence were able to engage them in a way that captured and integrated multiple strands of scientific practice. We included COV but went beyond it, making it clear that multiple factors were likely contributors to the outcomes of concern and hence needed to be examined, taken into account, and their effects coordinated. A data analysis tool for K-12 students, InspireData, was integral to this objective as it allows students to visually represent the effects of multiple

factors. Students could then use this multivariable understanding to predict outcomes based on evidence across multiple variables and thereby achieve the larger goal of the activity – drawing on evidence, rather than only their own beliefs, as a source in seeking to understand the relationships being examined. Finally, we engaged them in scientific writing in the form of reports to the sponsoring foundation. This activity included addressing challenges to their claims, thus exercising skills of both argument and counterargument.

Our pedagogical method can be characterized as one of guided inquiry. The phases of the investigation were segmented for students into a sequence of component tasks, with care taken to make clear the purpose and goal of each one and its purpose within the larger task. Students are not given direct instruction as to strategies to apply to the component tasks; rather, attention is focused on the task goal and on their coming to recognize the weaknesses of inferior strategies they use in not achieving this goal. As a culminating activity, they reflect on how their final conclusions differ from their initially solicited beliefs about the roles of each of the factors. Doing so leads to reflection on the task as a whole and on how their evidence-based conclusions provide knowledge central to achieving the best task solution.

Participants came from the low SES, low-achieving middle-school population in which several researchers have reported it difficult to develop rudimentary scientific thinking skills, compared to success in doing so among more privileged groups (Siler et al., 2010; Lorch et al., 2010). Some of these low-performing students, for example, in seeking to design an experiment fail to manipulate the focal variable. This failure can be attributed at least in part to absence of more basic understanding of the purpose of scientific investigation as a) seeking to answer questions whose answers are not already known, and b) engaging in causal analysis, rather than seeking only to optimize outcomes. Furthermore, it is widely observed that students in such populations typically show little interest in or disposition to study science. This population, then, seemed to us an especially important one to reach and achieve success with, remaining mindful of the practicality of the methods examined for large-scale classroom use. We therefore included here a comparison of two parallel methods, identical except that one is administered to pairs of students by a researcher while the other is administered to a whole class by the classroom teacher.

Methods

Student and school sample

Participants were 72 students (38 females) from three 6th-grade and one 7th-grade science classes, all taught by the same teacher. Participating in pairs in a pair intervention condition were 25 students (12 females) from one 6th-grade classroom. An equivalent 6th-grade class of 24 students (12 females) participated in the same intervention as a whole class. Twenty-three students (14 females) drawn equally from another 6th-grade class and a 7th-grade class served as a control group and received only the post-intervention assessments.

A 10-item written multiple-choice task of a type commonly used for this purpose and administered at the beginning of the school year confirmed that students showed little mastery of the control of variables (COV) strategy, with a majority of students scoring no more than 50% correct and scores of 100% correct rare, a finding consistent with others for this population. Group comparisons on this test were non-significant.

Intervention procedure

The content of the intervention was identical across conditions except that the classroom group participated as a whole class led by the classroom teacher and assisted by the first author (whose presence enabled her to confirm fidelity of implementation) and assistant, while in the pair condition the pair worked in a corner of the classroom with a facilitator (the first author) present throughout the intervention. In the classroom condition, students worked with a partner for most activities. In both conditions, the adult scaffolded all activities using a planned protocol of prompts that did not provide direct instruction (unlike tutoring) or hints, but drew attention to the task goal and challenged the weaknesses of inferior strategies. The intervention was administered to the classroom group over 10 45-min class sessions over a period of 16 days. The number of sessions in the pair condition varied as it was tailored to students' progress. Of 13 pairs, one completed the intervention in four sessions and one completed it in five, while most of the rest took six sessions and four pairs required seven sessions. Sessions averaged about 24 minutes (allowing two per class period). These sessions took place over an average of 32 days, with a range from 14 to 59.

At the first session, the activity was introduced about an organization trying to recruit astronauts and therefore want to know what factors (fitness, parents' health, family size, and education, with family size being the only non-effective factor) matter to applicants' performance. Following this introduction, pairs were asked to record on a form which of the four factors they thought would and would not matter. In the classroom condition,

a tally across the class was shown, and in both groups it was noted that opinions differ. Students were then told to work as a team and that they must come to an agreement for everything they do.

Control of variables phase

In the first phase, students were given a set of records of a list of applicants. Each record showed the levels of a list of factors being investigated for each applicant. Students were told that if they studied the records carefully they could determine which factors make a difference to performance and which don't. Students may select any records and request the performance level of the applicant by first explaining what they planned to find out by evaluating the records. Students were then provided with the performance levels of the records.

Once a pair was certain they had reached a conclusion about a factor's status, they could enter it on a "Draft Memo" to the foundation director. In the pair condition, if the conclusion was a valid one based on a controlled comparison, the pair proceeded to choosing another factor to examine. If no controlled comparison existed allowing a valid inference, the adult embarked on a sequence of probe questions whose purpose was to support recognition of the limitation of the students' investigative approach in not yielding a definitive conclusion (e.g., "Couldn't it also be the difference in education that's leading to the different outcomes?"). No superior approaches were suggested, and scaffolding did not go beyond highlighting failure to achieve the goal (a definitive conclusion). In the case of valid conclusions, challenging probes were introduced, e.g., "Suppose someone disagrees with you and doesn't think that this factor makes a difference; what could you tell them to convince them?" In the classroom condition, this probing could not be conducted individually, but once per class session (typically at the end of the session) a whole-class discussion occurred that followed this model, using one pair's work as an example.

Once a pair in the pair condition and the majority of pairs in the classroom condition had achieved at least three controlled comparisons (showing fitness, a two-level factor, and education, the only three-level factor, effective and family size, a two-level factor, ineffective), pairs completed their final memo to the foundation director, indicating which factors applicants should be asked about and which they should not and justifying their recommendations with results from their investigations.

Multivariable coordination phase

Students at this point were ready to transition to the next phase in which they represented and reasoned about the influence of all of the factors operating at once. Students were then introduced to the representation of their data using charts generated by the program InspireData and told, "All of the cases that you and your classmates have looked at before are here." It was explained that a chart shows the performance levels of all applicants with each diamond representing a case and its identity seen by hovering over the diamond (Figure 1).

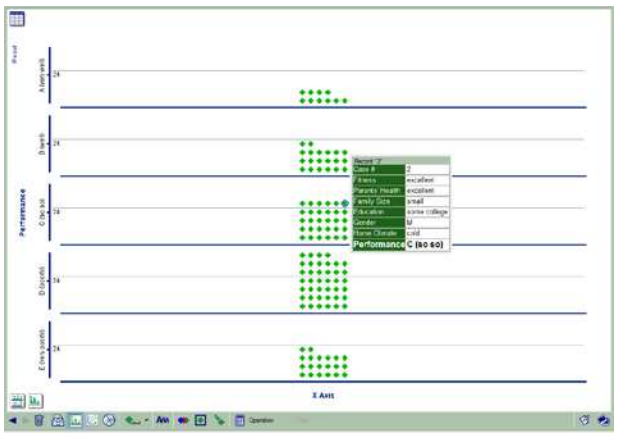

Figure 1. InspireData chart showing all cases.

It was then illustrated that charts can be generated to separate cases into different categories. A new chart with only those cases in which the applicant's fitness was average rather than excellent was shown. To highlight that other factors also contribute to the outcomes, students were then asked why it was that these applicants all of the same fitness level showed a range of performance outcomes.

Next, students were shown a third display in which all levels of the fitness variable are included. They were asked to draw conclusions about whether the factor makes a difference to the applicants' performance. Given the ability to see more data at once, students were asked to see if they reached the same conclusions as they did earlier when comparing individual cases presented on cards. Students were then provided InspireData charts for

each of five factors, four introduced previously and one new one (home climate, a non-effective factor), each of the same form, showing outcomes for all levels of the factor. In their pairs, they did this and then wrote memos to the foundation director confirming their earlier conclusions based on a larger sample or revising their conclusions if they thought necessary.

As in the previous phase of the intervention, prompts were introduced in the case of both correct and incorrect conclusions, e.g., "Suppose someone disagrees with you and doesn't think that this factor makes a difference; what could you tell them to convince them?" In the classroom condition, this probing could not be conducted individually, but once per class session (typically at the end of the session) a whole-class discussion occurred that followed this model, using one pair's work as an example.

Prediction phase

Students were told that now that they had reached final conclusions, they could try using them to evaluate a new set of applicants. They would then be able to select five to be chosen for the program and compare their choices to those of their classmates. Students were told that they could select up to four factors about the new applicants that they could receive information on. As students selected the factors, the adult reminded them to review the InspireData chart and consider whether knowledge of status on this factor would be informative as to outcome. In the classroom condition, a similar process took place at the whole-class level. Information about 10 new applicants on four factors (including one non-effective one, whether or not it was asked for) was presented one at a time. In addition to making each prediction, they were asked for each one, "Which of the four factors you have data on mattered to your prediction?" Students were encouraged to review the InspireData charts to double check their decisions or when there were disagreements between the student pair. In a final discussion pairs made selections of the five top-rated candidates and, in the classroom condition, shared these with the whole class. This discussion included remembering the beliefs they had initially held about the factors and noting that they would not have chosen the same applicants before and after the analysis they had conducted.

Post-intervention assessment

All post-intervention assessments except one were conducted individually and all were delayed in order to assess maintenance of achievements. Among students in the pair condition these assessments occurred an average of 26 days following completion of the intervention (range 14 to 42 days). Among students in the classroom condition they occurred an average of 32 days following completion of the intervention (range 18 to 46 days). Assessments for students in the control condition occurred during this same time period.

Findings

Designing experiments and making inferences

Final post-intervention achievement, maintenance and near transfer

The first component of the delayed post-intervention assessment was administered individually to students in both intervention conditions to assess the extent to which they maintained and transferred experimental design and inference skills (and specifically COV), The two items, each introduced a new variable (height and then strength) and asked students to select up to two cases to test its effect. Performance is summarized in Table 1. The large majority of both groups consistently designed comparisons that varied the focal variable. A majority in the pair condition also consistently designed controlled comparisons, as did half of those in the classroom condition, with the remainder doing so only inconsistently. The percentage of students who consistently designed controlled comparisons was higher in the pair than the classroom condition, $X^2(1) = 4.86$, $p = .027$.

Table 1: Comparison of intervention groups on maintenance and near transfer of design and inference skills

Skill levels	Pair condition	Classroom condition
Never varied focal variable	1 (4%)	3 (13%)
Varied focal variable only sometimes	2 (8%)	3 (12%)
Consistently varied focal variable but inconsistent control of other variables	2 (8%)	6 (25%)
Consistent controlled comparison	20 (80%)	12 (50%)

Note. $n = 25$ for the pair condition and 24 for the classroom condition.

Far transfer

The far transfer task, consisting of three written items of two levels of complexity, involving new content but assessing the same skills as the near transfer task. Performance by condition is summarized in Table 2. Thus, comparing tables 1 and 2, to an approximately equal extent across the two intervention conditions, fewer students maintained controlled comparison consistently in the far transfer context (table 2), when content was new and represented only in a traditional paper-and-pencil format. Results differed, however, by item complexity. For the two-variable items, the percentage of students who consistently chose controlled comparisons was significantly higher in the pair group than in the control group, $X^2(1) = 5.60$, $p = .018$, and marginally higher in the pair group than the classroom group, $X^2(1) = 3.50$, $p = .062$. The difference between classroom and control groups was non-significant for the two-variable items. However, for the more complex three-variable item, significantly higher percentages of students in both pair and classroom groups showed controlled comparison compared to the control group, $X^2(1) = 16.33$, $p < .001$ and $X^2(1) = 3.85$, $p = .050$, respectively. In addition, the percentage was higher in the pair than the classroom group, $X^2(1) = 5.13$, $p = .024$.

Table 2: Comparison of groups on far transfer of design and inference skills

Skill levels	Pair condition	Classroom condition	Control condition
Two-variable items			
Did not consistently vary focal variable	1 (4%)	12 (50%)	10 (43%)
Consistently varied focal variable but inconsistent control of other variables	9 (36%)	4 (17%)	7 (30%)
Consistent controlled comparison	15 (60%)	8 (33%)	6 (26%)
Three-variable item			
Did not vary focal variable	3 (12%)	4 (17%)	7 (30%)
Varied focal variable but inconsistent control of other variables	1 (4%)	7 (29%)	10 (43%)
Controlled comparison	21 (84%)	13 (54%)	6 (26%)

Note. $n = 25$ for the pair condition, 24 for the classroom condition and 23 for the control condition.

Multivariable analysis and prediction

This component, administered to all students, has previously been reported on (Kuhn, Ramsey, & Arvidsson, 2015). It presents (authentic although simplified) data about factors (Employment, Family size, Education, Home Climate) having an effect on average life expectancy across different countries and one non-contributing factor (Country size). Students were asked to predict life expectancy of additional countries based on information on their status on the identified factors. The task also asks respondents to indicate which factors they considered in their prediction.

Students overall did well on this task, indicating they understood the task and were capable of performing it, yet there were significant differences in performance across groups. Eighty-three percent of students in the pair condition, 63% of those in the classroom condition, and 22% of those in the control condition had modal performance of zero error (a correct prediction). Of remaining students who did not attain a modal performance level of zero error, all but one student showed a modal level of 1, with the remaining student showing a modal level of 2. When the mean prediction error scores over the six items were compared, the pair condition showed significantly less error than the classroom condition, $t = 2.78$, $df = 37.70$, $p = .009$. The classroom condition showed significantly less error than the control group, $t = 2.52$, $df = 42.05$, $p = .016$.

In attributing factors as having influenced their prediction, students again showed good performance overall but significant group differences. The pair group most often attributed influence to the four effective factors and least often to the ineffective factor. The control group were less likely than the pair group to attribute influence to each effective factor and more likely to attribute influence to the ineffective factor, with the classroom group intermediate but closer in performance to the pair group than to the control group.

Students were scored based on the consistency of their attributions as follows:

1-Chose only one but inconsistent factor across 6 countries

2-Chose only one consistent causal factor across 6 countries

3-Chose multiple but inconsistent factors across 6 countries

4-Chose multiple consistent (but not all four effective) factors across 6 countries

5-Chose four effective factors completely consistently across 6 countries

The mean difference in score between pair and classroom groups was significant, t =3.32, df = 42.38, p = .002. The classroom group significantly outperformed the control group, t = 2.86, df = 37.87, p = .007, as did the pair group, t = 7.73, df = 42.12, p < .001. With respect to individual patterns of consistency in attributions, the same pattern of group differences appears. Only one student in the pair group attributed influence to the non-effective factor one time. In the classroom group, one student did so one time, while seven (29%) did so multiple times. In the control group, six (26%) students did so one time while 11 (47%) did so multiple times.

In the pair group, 18 students (72%) consistently attributed influence to all four effective factors and never to the ineffective factor. In the classroom group, only nine students (38%) showed this pattern, while 13 (54%) showed inconsistency in attribution across cases (i.e., a factor is in some cases claimed to have influenced a prediction and in other cases not). In the control group only two students (09%) showed the consistent pattern with the remaining 21 (91%) showing inconsistency.

With regard to number of factors to which influence was attributed, 83% (19) of the pair group, 54% (13) of the classroom group and 17% (4) of the control group most frequently correctly attributed influence to four factors. Only students in the control group–22% (5)–most frequently identified only a single factor as responsible for the outcome. (Remaining students most often chose 2 or 3 factors.). Pair and classroom groups performed significantly better than the control group in most frequently attributing to four factors with p <. 001 and p = .015, respectively (Fisher's Exact test).

Argumentation

The final component of the delayed post-intervention assessment is an elaboration of the cancer task used by Kuhn et al. (2015). One of its parts pertains to types of counterargument and the other to reconciling divergent claims. Both are far transfer tasks as they bear no surface similarity to and make no reference to the intervention content. Furthermore, both ask students to reason about argument rather than only engage in it.

Evidence and counterargument

In the first part of the argumentation task, students were told:

> The Public Health department of Portland, Ohio has noticed that the percentage of residents diagnosed with cancer is much higher in the inner city than in the outlying neighborhoods. The department is undertaking a study to find out why there are more people getting sick with cancer in the inner city than the outlying area.

The student was then asked to choose among four options that would constitute the strongest evidence to show someone was wrong who claimed that the difference was due to the fact that city people more often go to tanning salons. Results for students in the three conditions appear in Table 3. Shown are percentages of respondents choosing each option as providing the strongest counterargument to the claim that tanning salon use was a causal factor with respect to cancer rates. As seen in Table 3, most respondents chose B or C with a smaller proportion favoring A. All students appeared to recognize that D was irrelevant to the claim and none of them chose that option. Among the three relevant options, A, B, and C, Chi-squared goodness-of-fit tests showed that distribution of choices of students in the classroom and control conditions did not differ significantly from chance, $X^2(2) = 1.75$, p = .417 for the classroom condition, and $X^2(2) = 0.61$, p = .738 for the control condition. For the pair condition, significance was borderline, $X^2(2) = 5.83$, p = .054.

Table 3: Percentages of students making different response choices regarding counterargument

	Pair condition	**Classroom condition**	**Control condition**
A. Air pollution is a more likely cause of cancer in the city	4 (17%)	5 (21%)	6 (26%)
B. Many people outside the city also go to tanning salons and don't get cancer	6 (26%)	10 (42%)	9 (39%)
C. Many people who don't go to tanning salons also get cancer.	13 (57%)	9 (38%)	8 (35%)
D. There are more tanning salons outside the city than in the city.	0 (0%)	0 (0%)	0 (0%)

Note. n = 23 for the pair condition, 24 for the classroom condition and 23 for the control condition.

Reconciling claims

The final task continued the topic theme but was presented in writing in the classroom on an occasion about one and a half months after the other posttests, reducing possible influence of students' particular responses to the

previous task having the same theme. Two potentially conflicting causal claims are now explicitly presented and the question asks the participant how to interpret this discrepancy:

> You were hired by the Health Department to find out why people living in the city of Logan, Georgia are getting cancer more often than people who live outside the city. You tested and found out that air pollution was worse inside the city than outside. You wrote a report of your findings to the Health Department director, telling her that air pollution was a likely cause of the increase in cancer.
>
> She also got a report from another person she hired. This report said that a likely cause of the cancer increase was not enough stores in the city for people to buy healthy fruit and vegetables that lower risk of cancer.
>
> The director isn't sure what to conclude and she has written you asking for advice. What would you write back? Give her the best advice you can.

This question, in contrast to the previous one, solicits reasoning not about the claims themselves but rather meta-level reasoning about their status in relation to one another and how the discrepancy between them is to be understood – a form of reasoning that is epistemological in nature and central to scientific practice. We expected that answers to this question would give us insight into students' epistemological understanding regarding the nature of science, more specifically the extent to which they understood it as an enterprise involving competing claims whose status evolves as more evidence is brought to bear on them.

Contrary to our expectations, a very large majority of participants did not address the question as one of how the divergence in claims is best understood and reconciled. Their answers thus did not bear directly on their understandings regarding the nature of the scientific enterprise, except in the negative sense of their not seeing the divergence as warranting attention or interpretation. Instead, students' dominant stance was one of how to use toward practical ends the information that had become available. Their understanding of the nature and objectives of science, in other words, remained one common among this age group – producing good outcomes rather than analysis of causes and effect. Thus, the objective students identified in the context of this question was to reduce cancer. None of the responses raise questions about whether the validity of the causal claims should be evaluated, rather than their being simply put into action.

Conclusions and implications

In light of a history of difficulty in effecting advances in higher-order thinking skills in the disadvantaged, low performing population studied here, the goal of this study was largely met. With 80% of students in the pair condition and 50% in the classroom condition consistently showing controlling across multiple tasks, these results compare favorably to previous efforts with similar populations devoted only to the COV strategy (Lorch et al., 2014; Siler et al., 2010). With unfamiliar material, the majority (60%) of the pair group maintained consistent control, while a third (33%) of the classroom group did so. These findings are consistent with the view that continued and varied experience across a succession of domains is necessary if consolidation of higher-level cognitive strategies (as assessed in tests of maintenance and far transfer) is to be achieved.

With respect to the less studied skills entailed in coordination of effects of multiple variables, both intervention groups displayed considerable mastery, maintenance, and transfer, especially in relation to the far from optimal performance shown by adults (Kuhn et al., 2015) as well as the present control group. Playing a critical role in this success, we believe, is the InspireData tool, as it affords students a visual representation of relations among variables (vanAmelsvoort, Andriessen, & Kanselaar, 2007). The charts allowed students to see and interpret a representation of the more common and realistic case of multiple variables in action together – a portrayal fundamental to scientific practice. A further contributing factor, we believe, is the social science frame of both intervention and transfer tasks that attached a readily understandable purpose and goal to the reasoning students were asked to do. Also, the frame of the present intervention task – optimizing astronaut selection – has the engineering focus of producing good outcomes, but it suggests a way that the two orientations, engineering and analysis, can be coordinated. The typical conception has been that the orientation of science students needs to progress away from the engineering focus and toward an analysis focus. Conceived differently, science students can come to appreciate that analysis is an essential tool in the achievement of engineering goals.

Results are more mixed with regard to argumentation and to developing meta-level thinking about argument and related understandings of nature of science. The intervention activities involved not only a repeated requirement to justify one's claims with appropriate evidence, but also repeated engagement with the probe, "Suppose someone disagreed with you…" which we expected to afford exercise in defending and supporting alternative claims using evidence-based arguments. This expectation was largely met within the intervention.

Students with practice became successful in meeting these demands and most often did so confidently by means of direct appeal to evidence, e.g., "I'd show them the results on the chart."

Where students demonstrated less success was in extending these competencies to meta-level reflection about argumentation in new, far transfer contexts unrelated to the intervention. No group differences clearly appeared, despite the intervention students' strong performance in taking into account the effects of multiple variables on an outcome, both within the intervention and in a transfer task with new content. However, they commonly did not recognize absence of an antecedent in the presence of the outcome (option C) as similarly weak in being consistent with an alternative factor having produced the outcome, choosing this option as often as the correct option B. Students' understanding of the strength of various kinds of evidence in weakening (as opposed to supporting) a claim is thus fragile and in need of development, a finding consistent with research showing the use of evidence to weaken claims more difficult than its use to support claims (Kuhn et al., 2015).

Finally, despite their mastery of counterargument within the intervention, students' performance was weak in the assessment of meta-level reasoning about argumentation in the final task, asking them to account for divergent claims. The large majority of students did not treat divergent causal claims as a cause for attention, examination, or attempted reconciliation. Instead, without acknowledging the divergence, they focused on one or the other or both imputed factors as causes worthy of action, without further investigation. This conception stands in stark contrast to recognizing diverging claims as needing to be examined and evaluated as an undertaking central to the practice of science.

It is now commonly emphasized how critical it is that students develop their understandings of the nature of science by engaging in it. The nature of science, most science educators now agree, cannot be taught in a deep way through passive direct instruction and rather must be experienced through students' own activities. As students engage in a larger number of purposeful, goal-directed activities that involve science practices, over a range of content, they are in the best position to extract general attributes of these practices and to appreciate their value. This experience can only accrue gradually in a facilitative context. As well as the skill components involved in coordinating evidence and claims in the service of argument, scientific practice encompasses values and norms that come from participation with others in a community that upholds shared standards of knowing.

We turn finally to the comparison of our two intervention groups. Overall it was the individually instructed group who consistently showed superior performance, despite the lesser time (less than a third as may sessions on average) invested. This outcome, we believe, is one that has important practical implications, in speaking to the value of individualized instruction, in particular for the population we studied. Constant interaction with each other and the adult and required to think about and justify whatever they said at just the time they said it may have explained some of the differences.

The conclusion to be drawn, we believe, is that chronically disadvantaged, underserved, and underperforming students like the ones we worked with have the potential to be successful in developing higher-order intellectual skills, given sustained purposeful activity in an advantageous setting that engages students in dense exercise of reasoning. In current work, we are therefore exploring ways to make the protocol used here automated in a way that could make it practical for large-scale use. This is of course a sizeable step from personal conversation between peers and a more knowledgeable facilitator, but we believe it may be one worth pursuing, especially in seeking to reverse the persistent lack of success among low achieving students.

References

Kuhn, D., Ramsey, S., & Arvidsson, T. S. (2015). Developing multivariable thinkers. *Cognitive Development, 35, 92-110.*

Lorch Jr, R. F., Lorch, E. P., Calderhead, W. J., Dunlap, E. E., Hodell, E. C., & Freer, B. D. (2010). Learning the control of variables strategy in higher and lower achieving classrooms: Contributions of explicit instruction and experimentation. *Journal of Educational Psychology*, *102(1),* 90-101.

Lorch Jr, R. F., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., & Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *Journal of Educational Psychology*, *106*(1), 18-35.

Sandoval, W. (2014). Science education's need for a theory of epistemological development. *Science Education, 98,* 383-387.

Siler, S., Klahr, D., Magaro, C., Willows, K., & Mowery, D. (2010, January). Predictors of transfer of experimental design skills in elementary and middle school children. In *Intelligent Tutoring Systems* (pp. 198-208). Springer Berlin Heidelberg.

van Amelsvoort, M., Andriessen, J., & Kanselaar, G. (2007) Representational tools in computer-supported collaborative argumentation-based learning: How dyads work with constructed and inspected argumentative diagrams. *Journal of the Learning Sciences, 16,* 485-521.