Chapter 6: Conclusion

Revisiting Our Objectives

We set out with this work to examine the relationship of hands-on science with student achievement. Our work arises from the current intersection of two science education policies. The first is the ongoing promotion of hands-on science for the past 40 years and the second a decade-long science reform effort that would reduce the use of hands-on science. Our concern arises from the mixed results on past research regarding the relationship between hands-on science and student achievement. Without a conclusive finding on this relationship, it is not clear which of the two is the better policy to implement.

Our fundamental research question is whether there is a positive link between hands-on science and student achievement. Past research on this topic has examined the relationship of the level of hands-on science in the classroom with student scores on standardized tests. We have identified four issues that may contribute to the inconclusiveness of this past work as well as have implications for current policy. These issues include: 1) the need to control for variables that may be linked to both hands-on science and test scores, 2) the need to examine the link between hands-on science and performance test scores as well as the more traditional multiple choice test scores - the importance of which has increased with the current debate over the adequacy of using multiple choice tests to measure student achievement, 3) the need to investigate a potential differential relationship of hands-on science and test score due to student ability, and 4) the need to consider the multiple facets of hands-on science including the quantity, quality and instructional approach used with it.
With guidance from the theoretical and empirical literature, we developed three hypotheses from our research question and the above four issues. Specifically, we hypothesized that:

1. Higher levels of hands-on science are associated with higher test scores, be they multiple choice or performance tests, all else being equal.

2. This association is stronger with performance tests than with multiple choice tests.

3. This association is weaker for higher ability students.

To test these hypotheses, we analyzed two data sets allowing us to perform complimentary analyses, giving us greater confidence in our results, as well as addressing the first three issues. Unfortunately, we were not able to address the fourth issue as data limitations did not allow us to discern the quality of nor the instructional approach used with the hands-on science. Our primary data set, the RAND data, has a large 8th grade student sample from the Southern California region but a small teacher sample. It is used to test all three hypotheses. The advantage of the RAND data is that it contains both multiple choice and performance test scores for the same students. A disadvantage is the lack of some covariates we would like to see included. The NELS is nationally representative and contains the covariates linked to the level of hands-on science and student test scores. It has both a large student and teacher sample but lacks performance test data and an 8th grade student survey of hands-on science. We use it to test Hypotheses 1 and 3. NELS also contains data on students in 10th and 12th grades allowing us to examine the relationship in higher grades as to its similarity or difference (as predicted by developmental theory) with 8th grade results.
Overall Findings

The table below lays out the overall findings from the analysis. Column 1 lists the survey and the grade level of the sample. Column 2 identifies the source of the hands-on science measure: reported by student or teacher. Columns 3 & 4 are concerned with the evidence for Hypothesis 1 and note whether positive relationships between the hands-on science scale and the multiple choice test scores or performance test scores were found. Column 5 lists whether evidence was found in support of Hypothesis 2 that the relationship would be stronger with performance test scores. Columns 6 & 7 concern Hypothesis 3 and show whether higher ability students have a less positive relationship between hands-on work and multiple choice or performance test scores. These results should be considered with the strengths and weaknesses of the two data sets kept in mind (a topic we will return to in the next section).

Table 6.1: Overall Results

<table>
<thead>
<tr>
<th>Survey</th>
<th>Hands-on Measure</th>
<th>Hypothesis 1</th>
<th>Hypothesis 2</th>
<th>Hypothesis 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MC Test</td>
<td>PA Test</td>
<td>MC Test</td>
</tr>
<tr>
<td>RAND</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8th Grade</td>
<td>Student</td>
<td>Yes¹</td>
<td>Yes</td>
<td>Yes²</td>
</tr>
<tr>
<td></td>
<td>Teacher</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>NELS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8th Grade</td>
<td>Teacher</td>
<td>Yes</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>10th Grade</td>
<td>Student</td>
<td>No³</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Teacher</td>
<td>Yes</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12th Grade</td>
<td>Student</td>
<td>No</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Teacher</td>
<td>No</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

¹ For students of lower Ability Ranks  
² For students of higher Ability Ranks  
³ Not found with full scale but found when using between-class variation in the scale.  
⁴ Incomplete scale
Let us start by summarizing the evidence regarding Hypothesis 1 using the 8th grade results as these are the focus of our primary data set. The RAND results support Hypothesis 1 in regards to students of lower Ability Ranks when using the student reports but provide no supporting evidence when using the teacher data. The NELS 8th grade results, based on teacher reports and multiple choice test scores, also support Hypothesis 1 but without regard to student Ability Rank. As the NELS 8th grade student survey did not include items on hands-on science, we have no results using a student scale. Hence we are faced with the seemingly inconsistent result that teacher reports from one data set support our hypothesis while teacher reports from another do not. This inconsistency is not a total surprise for the RAND data has a low variation in the teacher scale due to the small number of teachers taking part making it difficult to find a relationship.

Looking to the NELS results using higher grades for further evidence leads to further mixed evidence. The NELS teacher reports support the hypothesis for 10th grade students as they did for 8th grade. However, the 10th grade student reports do not and we again face inconsistent results when using reports from the same source (this time students) in different data sets. Here again, this inconsistency is not unexpected as the NELS 10th grade student scale is incomplete and may not be an adequate measure of hands-on science. For the 12th grade results, the results based on either the student and teacher data do not support Hypothesis 1. Further consideration must be given to the appropriateness of comparing 12th graders with 8th graders before making conclusions based on this finding.

Concerning Hypothesis 2, we can draw upon our RAND results which show a stronger relationship for hands-on science with one type of performance test than with the
multiple choice test for higher ability students but no difference for lower ability students. Our finding for higher ability students is not really in the spirit of Hypothesis 2 which was based on the idea that hands-on science would better prepare students for performance tests. Instead, our finding occurs because for higher ability students hands-on science is negatively linked to multiple choice test scores rather than associated with higher performance test scores.

For Hypothesis 3, the RAND student results provide evidence that hands-on science is positively linked to test scores for lower ability students for both types of test while this relationship does not occur for higher ability students. For performance test scores, we see these findings after breaking out the between-class variation in the scale. However, the NELS results show a uniform positive relationship between hands-on science and test score, when using the 8th and 10th grade teacher data, for all students regardless of student ability. Because the NELS model includes additional relevant covariates we have greater confidence in them and overall we do not find enough evidence to support Hypothesis 3.

Findings for the major covariates are similar for all the analyses. We see positive relationships between test scores and higher SES and higher student ability (individually in both data sets as well as in high class track and high class achievement level in the NELS data), and negative relationships between test scores and Hispanic, Black and classroom percentage minority.

The findings were further examined through a series of sensitivity analyses which confirmed their robustness. For the RAND data, we broke the student scale down into its between-class and within-class variation as the former may be a better measure of the
actual differences in hands-on science. Additionally we checked if the results were maintained for different student characteristics and considered the need for non-linearities in the relationship of hands-on science to test score. For the NELS data, we examined alternative teacher scales (based on uniformity between the three waves). In addition we performed a detailed analysis of the impact of course taking on the relationship including current enrollment in a science course and the pattern of courses taken.

In conclusion, we find little evidence to support Hypotheses 2 & 3. After a first glance at the data for Hypothesis 1, the results from the two data sets appear contradictory. The analysis of the RAND data found a relationship between test scores and the student hands-on science scale but not with the teacher scale. With NELS, we found a relationship between the 8th and 10th grade teacher scale and test scores but little relationship when using the student scale in 10th grade. No relationship was found when using NELS 12th grade data, teacher and student.

However, our findings in regard to Hypothesis 1 need to be further evaluated. We must consider the quality of the data and its appropriateness before concluding that the mixed findings do not allow us to make a conclusion regarding Hypothesis 1.

Evaluating the RAND and NELS:88 Results Based on the Strengths and Weaknesses of the Data

The RAND and NELS data sets differ in their strengths and weaknesses and this is what makes them complimentary. When considering the inconsistent evidence provided by the results from both data sets we wish to place greater emphasis on the results derived from the strengths of each. Here we examine the strength of the data sets
in regards to three issues: 1) the teacher scale results, 2) the student scale results, and 3) the NELS 12th grade results.

Regarding the teacher scale, NELS has the better data on teacher reports of hands-on science. The RAND data contains a small sample of teachers with little variation in their reports of hands-on science. For this reason, it is not unexpected that we would not find a relationship between the teacher scale and test scores when using the RAND data. NELS has a large teacher sample with greater variation in teacher reports. If we had failed to find the positive relationship using the NELS, it would be strong evidence that the relationship does not exist. However, we did find such a relationship when using NELS (8th and 10th grade data) and we can attribute this, in part, to the larger teacher sample available. For this reason, we have greater confidence in our results for the teacher scale when using the NELS data than the RAND data. Furthermore, our confidence is greater in the NELS results because of the additional covariates that have been controlled for.

For the student scale, the RAND data is stronger. RAND’s student survey asked questions well related to the level of hands-on science in the classroom including key ones on how often experiments were done and scientific materials were used in class. NELS did not ask students in 8th grade about the level of hands-on science so no comparison with the RAND data can be done for the 8th grade. In 10th grade students were asked several questions but the key one on how often experiments were done was not included. Therefore the 10th grade student scale is lacking a crucial item and we are not surprised that it shows little relationship with test score. For this reason, we have greater confidence in our results for the student scale when using the RAND data.
Both the 12th grade student and teacher scales showed no relationship with test score. The teacher scale was similar in construction to the scales of the earlier grades and the student scales contains the crucial item on how often experiments were done. Therefore the quality of the items making up the scale does not reduce our confidence in the results.

However, there are two other concerns regarding the 12th grade data. The first is the reliability of the 12th grade (and 10th grade) student scale. As noted in Chapter 5, our descriptive analysis found opposite results for the 10th grade versus 12th grade student scale when students were broken down into racial/ethnic and student ability groups. Specifically, Black reported the lowest scale in 10th grade but highest in 12th. The student scale also flipped from high to low in 10th versus 12th grade for the highest ability rank and class achievement. This concern over the student scale reduces our confidence in the 12th grade results based upon it.

Second, 12th grade students are very different from 8th grade students. Abstract thinking should be further developed by 12th graders and the concrete illustrations of hands-on science may be less beneficial for their learning. For this reason, they may not respond to hands-on science the same way that 8th graders do and we would expect the relationship of hands-on science to test score to be weaker for older students. For this reason, we have less confidence that the results using the 12th grade data make a proper comparison with those of 8th grade.

Our evaluation leads us to differentiate among the results providing us with more confidence in those based upon the RAND 8th grade student scale and the NELS 8th and 10th grade teacher scales. From these results, we argue that hands-on science is positively
related to test scores. The evidence is stronger for a relationship with multiple choice test scores as it comes from two separate surveys using different multiple choice tests. The results from the NELS data makes us more confident that the relationship between hands-on science and multiple choice test scores does indeed remain when a fuller array of covariates are controlled.

The magnitude of the relationship was found to be substantial for both data sets. An increase of 1 point in the 5 point hands-on scale leads to an increase of almost 0.2 of a standard deviation of the test score for the multiple choice test in the RAND analysis and 0.1 in the NELS 8th grade analysis. If we consider a shift from the lowest to the highest level of hands-on science (from 1 to 5 points), the coefficients would convert to four times their value, a magnitude equivalent (though opposite in the sign) to what we saw for the negative relationships of Hispanic or Black to test score.

In conclusion, our analysis provides three important findings. First, hands-on science is positively related to test scores for upper middle and lower high school students. Second, this relationship persists after controlling for additional variables that are closely related to science achievement and after carrying out a series of sensitivity analysis. Third, the relationship is substantial in that a full implementation of hands-on science is equivalent to the negative association of minority status (Black and Hispanic) with test score. Not enough evidence was found in support of the two other hypotheses that hands-on science would be more strongly associated with performance test scores versus multiple choice test scores or that its association with test score would depend upon student ability rank. Further research, discussed below, could further confirm these findings and address the weaknesses of our data.
Policy Implications

After using the strengths of the data to evaluate the results, we find a positive relationship between hands-on science and test score. This finding supports a continued emphasis on the promotion of hands-on science that began in the 1960s, with an exception for upper high school. The finding of a relationship with both multiple choice and performance test scores should make this promotion more attractive. In those states or districts that intend to continue to rely on multiple choice tests, the use of hands-on science can support efforts to increase scores without the fear that the increased time required for hands-on science will harm student test scores. States or districts that adopt performance assessments often promote increase hands-on science under an assumption that the two are linked. Our results provide evidence to confirm this assumption. States and districts using a combination of test types may have been torn in the approach to take toward increasing scores. Our results show that an emphasis on hands-on science can support gains in both types of test scores.

A continued emphasis on promoting hands-on science will require greater attention to the practical constraints to its use (discussed in Chapter 2) including the need for logistical support and adequate training whose lack helped block the adoption of the curricula developed in the 1960s (Chapter 3). It is clearer today that supporting the teaching of hands-on science requires more than just developing new curricula. Others issues need to be addressed including the recurrent needs in training faculty and providing materials in a timely and affordable basis, addressing faculty turnover, offsetting deficits in faculty content knowledge and classroom management skills, and ensuring long-term administration support for this work. In addition, the requirement for
greater time per topic under hands-on science (discussed in Chapter 2) must be taken into account. This can be partly addressed by reducing the repetitive busy work aspect of hands-on science that has been criticized throughout its history. But, consideration will also have to be given to a reduction in content coverage both in the curriculum and in the standardized tests.

The finding of a positive relationship does not support the current emphasis of science reform to temper the use of hands-on science. Instead, it suggests that further research should first be done on the instructional methods proposed by reform to determine if there is evidence for their reduced emphasis on hands-on science.

Additionally, current efforts to focus on inquiry as the primary instructional approach for hands-on science need to be reconsidered. We were not able to separate out the relationship by instructional approach so our results are based on the average effects of the various instructional approaches. As inquiry has only been promoted for a relatively short period of time it is unlikely to be the predominant instructional approach used with hands-on science and so is unlikely to be the primary source of the relationship of hands-on science and test scores. Further research should be done on all available instructional approaches before we discard some of them that may be contributing to positive student outcomes.

The finding that the positive relationship of hands-on science and test score does not differ by type of test has implications for testing programs. If multiple choice tests capture the benefits of hands-on science as well as performance tests, then the promotion of the increased use of performance tests cannot be justified on these grounds. An alternate view of our results could be that both the performance and multiple choice tests
used in this analysis were poor at capturing process skills. For this reason, further research using other versions of both types of tests should be done in an attempt to extend our analysis.

Our inconsistent findings on whether ability rank affects the link of hands-on science to achievement should reduce criticisms of its widespread application among all students. That the evidence does not clearly show an adverse relationship for high ability students should reduce resistance from academic oriented students and their parents who fear that hands-on science might reduce the breadth of knowledge needed to succeed in standardized tests.

In considering these policy implications, several caveats should be kept in mind. First, our conclusion of a positive relationship of hands-on science and test score is based on our evaluation of the mixed results using the strengths of the data sets. Others might evaluate the results differently and give greater weight to our findings that did not show such a relationship. Second, our conclusion concerns a positive relationship but does not prove a causal relationship between hands-on science and test score. If some other non-controlled factor is responsible for both greater levers of hands-on science and higher tests, then policies to promote hands-on science may not lead to higher test scores unless they also raise levels of the unknown factor. Third, our analysis was not able to control for quality of instruction nor instructional approach (the latter was discussed above). However, our inability to consider quality of hands-on instruction leads to conservative results. By definition, higher quality instruction leads to greater student learning and higher test scores. We would then expect to find a stronger relationship between high
quality hands-on science and test score than the relationship we found which was based on an unknown average level of quality.

In sum, our results provide evidence in favor of the above discussed policy options. Additional research will be necessary to confirm our findings, further address some of the inconsistencies we found as well as overcome the caveats we had to place upon our policy options due to research issues we could not resolve.

**Further Research**

Further research needs identified in this work can be broken down into two categories: research to confirm and extend our findings and research to support the better use of hands-on science in the classroom.

Because the two data sets were in many ways complementary rather than confirmatory, there are several findings whose replication would provide additional evidence to support the policies described above. The results from the NELS analysis would be strengthened using similar data that also contained performance test scores. A data set containing both more covariates and performance test scores could also be used to confirm the lack of a differential relationship by type of test. Furthermore, the collection of this type of data longitudinally that also contained consistent items on the level of hands-on science across grades would allow a stronger test of the relationship of hands-on science to test score including some consideration of causality.

We examined two multiple choice tests from separate respected sources and two performance tests developed by RAND. Similar results using different standardized tests
would strengthen confidence in our results. The policy impact of this further research would be greater if the tests chosen were already in wide use.

Confirmation of a more basic issue would be to address the validity of hands-on scales obtained from survey data using such techniques as classroom observation or teacher and student logs. Surveys on hands-on science to both students and teachers could be given and compared to results from these techniques. These comparisons could be used to improve survey items and help develop items to identify quality and instructional approach. This work might also help explain inconsistencies in results from teacher versus student reports on the level of hands-on science.

Research for the better application of hands-on science is primarily concerned with the instructional approaches used in providing hands-on science. The current choices in the approaches are primarily made on theoretical grounds or convenience. Specific studies on the approaches used in the classroom and the relationship with test scores could be done to determine the best approach or mix to use. The inclusion of items in the national surveys on what instructional approaches are used with hands-on science would be a step forward in this direction.

In addition, further analysis on upper high school grades could examine the uniqueness of upper high school science instruction and learning. This could help determine empirically if there is a difference in the relationship of hands-on science and test score between middle school and upper high school students. The lack of evidence from our work to support the use of hands-on science in upper high school requires further research to determine its role there.