Chapter 8:
Synthesizing the Effects of Prompts, Reflection, Cites, and Beliefs

Introduction and Rationale

Earlier chapters illuminated ways in which directed and generic prompts elicit different kinds of reflection. We saw that students who receive the two kinds of prompts achieve different levels of success on measures of project quality. I described students’ beliefs about science and learning and how those beliefs are related to students’ reflection and performance. Those analyses looked at each variable in isolation. What happens when we combine the effects of prompt type, reflection focus, cites made, and beliefs to investigate how they jointly affect students’ performance? This chapter relates these factors on the basis of the earlier findings.

We can use multiple regression to predict the quality of students’ projects using what we know about them as students and about other aspects of their work. Multiple regression accounts for the fact that some of the input variables are related to one another. For example, analyses in previous chapters identified a negative correlation between elaboration of reflection prompts and lack of reflection in response to reflection prompts. Multiple regression allows us to characterize the effects of both variables, providing in particular a way to tease apart the effects of each to identify which (if either) plays the more important role in predicting each outcome measure. Multiple regression also allows us to see how predictive our models are, by quantifying the amount of variance accounted for by the variables in the model.

Analyses of variance, or ANOVAs, provide another avenue for assessing what plays an important role in students’ performance, and for whom. Once we have identified what variables appear particularly salient for predicting a particular quality measure, we can group students with similar characteristics to see how the groups differ in their performance on the outcome measures. This allows investigation of interactions between condition and other variables, as well. For example, we may find, as implied by some of the results in Chapter 6, that certain kinds of reflection are highly useful for students in one condition but negatively related to success for those in the other condition.

The next, more difficult question is, How can we draw causal conclusions about these relationships? The analyses in Chapters 5, 6, and 7 were mainly
correlative. Though neither multiple regressions nor ANOVAs allow us to make causal links directly, they do provide sufficient evidence to hypothesize about which relationships may be causal and which are more likely to be correlative. The end of this chapter provides more discussion of the causal-versus-correlative issue. For now, we turn first to a more complete discussion of the methods employed in these synthesizing analyses.

**Methods**

No new data sources or outcome measures come into play in these analyses. Rather, the analyses discussed in this chapter combine measures used in Chapters 5, 6, and 7 in new ways. The three outcome measures are the three main project quality measures used earlier: overall critique quality, coherence of scientific ideas, and guidelines quality. These three scores measure the main cognitive goals of the All The News project.

The multiple regressions use variables relating to students’ condition, reflection prompt response characteristics, cites made, and beliefs as predictors of the three outcome measures. Because of the subtle difference in directed and generic prompts—that responses to directed prompts can and should be further broken down into responses to Thinking Ahead prompts and to Checking Our Understanding prompts, whereas for generic prompts this split is inappropriate—the multiple regressions were run separately for each condition, rather than using interaction variables to assess the interaction effects of condition and the other variables.

Reflection prompt response characteristics include *degree of elaboration* and *reflection focus*. For directed prompts, these are broken down further based on the prompt placement (Thinking Ahead and Checking Our Understanding) as well. For generic prompts, the overall values are used. Particular aspects of elaboration and focuses of reflection were chosen based on the correlative analyses in Chapter 6. Not all possible variables were tried in a multiple regression model; only those that I had reason to believe were promising were introduced at all. Because both the number of words and the number of comments were highly related to success, I considered both of these as possible predictors. For reflection focus, variables with strong correlations with at least one outcome measure were tried. Most reflection focuses were tried in at least one model. Schoolish, socially-oriented, and critique-oriented focuses did not have strong correlations with any outcome measures and thus were not used in any of the multiple regression analyses.

The *cites* students made in doing their claim notes can also be used to predict the quality of their projects. Specifically, these variables included the number of principles, classroom experiences, evidence, labs, and everyday experiences
cited, as well as the total number of cites. Recall that we have information on students’ cites for claim 1 (energy conversion) alone, claim 3 (thermal equilibrium) alone, and the two claims combined; most of the cite types were tried in at least one of their forms.

As discussed in Chapter 7, students’ beliefs were coded along the dimensions of autonomy, strategy, and process. The scores of each student in a pair were summed for a measure of the pair’s beliefs in that dimension. The analyses presented here use those pairwise sums as predictors of each outcome measure. I also use a measure of students’ overall beliefs, which sums their scores on each dimension. The overall beliefs score and all three dimension scores were tried as possible predictors of each of the quality measures.

Multiple regressions using these variables were run iteratively, based on principled and empirical expectations of productive predictors. I first identified the best model using students’ focuses of reflection and the kinds of cites they made, choosing variables to include based on strong correlations identified earlier. Once the best model for reflection focus and cites was identified, I added the overall beliefs score to that model. Sometimes the effect was clear: The beliefs variable either clearly reduced the predictive power of the model or clearly enhanced it. Quite often, though, it pointed to variables relating to reflection focus or citing that fell out of the model because they were overshadowed by the beliefs variable, in which case, a new model was identified. I then ran multiple regressions using the three beliefs dimensions as predictors of each quality measure, to identify which dimension or dimensions played the most important role for that quality measure. The next step was to take both the original models (without beliefs score) and the new models (including overall beliefs), and fold in the specific beliefs dimensions identified as particularly important. For the models that included the overall beliefs score, I replaced that variable with the dimension or dimensions that had been identified as very predictive. Again at this stage, I identified whether adding the specific dimensions strengthened or weakened the model. Through this principled iterative process, I identified, for each condition, a single best model for predicting each of the three quality measures.

Looking at where the variables are and are not predictive provides a window into conceptualizing the specific effects of reflection, citing, and beliefs in the contexts of the two kinds of reflection prompts. It also illuminates characteristics—of students, or of their reflection—that can provide criteria for categorizing students into groups with similar, salient characteristics, to see how their performance differs from that of other students. ANOVAs provide the vehicle for these analyses; the variables identified as most predictive were used for grouping students. For example, the results in
Chapter 6 identified lack of reflection as a possibly interesting characteristic, and the multiple regressions confirmed lack of reflection as an important predictor of one quality measure in particular. An analysis of variance, then, would compare the performance on the project of students who were highly unreflective with those who were reflective.

The two types of analyses together give us important insights into what kinds of beliefs and orientations toward reflection and citing are important and in what contexts that is the case. As has been the case in the earlier chapters, qualitative data are interspersed with the quantitative results to provide rich examples of the ideas being discussed.

Results

I break the results into two main groups: those that involve predicting the values of the project quality measures, and those that investigate groups' performance on the project. Before the analysis of the groups I include a synthesis of the predictive findings, which points to potentially interesting groups of students with similar characteristics.

Predicting the Quality Measures

In this section, I discuss each of the project quality measures in turn, starting in each case with the best model identified for the directed prompt condition, and then turning to the best model for the generic prompt condition. To preview my findings, I show that the reflection elicited by directed and generic prompts has very different effects on students' performance—never does a particular focus of reflection span both models of a quality measure. Furthermore, we will see that characteristics of students are at least as important as characteristics of their reflection, especially for the generic prompt condition and for predicting the coherence of students' ideas.

Predicting Overall Critique Quality

We turn first to predicting the quality of students' critiques.

Directed Prompt Condition

The variables most useful in predicting overall critique quality for the directed prompt condition included (a) the number of words in Thinking Ahead responses, (b) the proportion of cryptic (uncodable) responses (to all directed prompts), (c) the proportion of Thinking Ahead responses focused on
specific project goals, and (d) the proportion of non-reflective Checking Our Understanding responses. These variables accounted for 36% of the variance ($F[4, 38] = 6.914$, $R^2 = .421$, adjusted $R^2 = .360$, $p = .0003$; see Table 8–1 for regression coefficients).

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Standard Coefficient</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>10.074</td>
<td>4.874</td>
<td>10.074</td>
<td>2.067</td>
<td>.0456</td>
</tr>
<tr>
<td>Elaboration (TA)</td>
<td>0.289</td>
<td>0.115</td>
<td>0.321</td>
<td>2.508</td>
<td>.0165</td>
</tr>
<tr>
<td>Uncodable responses</td>
<td>-135.548</td>
<td>43.874</td>
<td>-0.390</td>
<td>-3.089</td>
<td>.0037</td>
</tr>
<tr>
<td>Project-specific resp. (TA)</td>
<td>14.890</td>
<td>5.900</td>
<td>0.314</td>
<td>2.524</td>
<td>.0159</td>
</tr>
<tr>
<td>Non-reflective resp. (COU)</td>
<td>-9.213</td>
<td>4.569</td>
<td>-0.255</td>
<td>-2.016</td>
<td>.0509</td>
</tr>
</tbody>
</table>

Table 8–1: Regression Coefficients for Predicting Overall Critique Quality for Directed Prompt Condition

Elaborated plans, especially if the ideas had to do with specific project goals, were linked to high quality critiques overall. However, writing uncodable responses to any prompt or non-reflective responses to Checking Our Understanding prompts was linked to poor critiques. (The number of words in responses to Checking Our Understanding prompts and the proportion of non-reflective responses to those prompts were very highly correlated with one another; length of Checking Our Understanding response was also a predictor of critique quality but fell out of the model when proportion of non-reflective responses was included.)

All of the findings fit the same pattern: Students who take the opportunities given to them to reflect are likely to be more successful than students who (by not elaborating, being cryptic, or being actively non-reflective) eschew these opportunities. The more successful students develop plans and monitor their understanding, allowing them to identify problems with their own understanding and with the evidence and claims, as well.

A Thinking Ahead response from a pair that exemplifies the model is given here:

S107 & S121, Thinking Ahead response: When we critique evidence, we need to look at all the evidence shown and decide what are the main ideas or facts that are important to support the claim. Then, we need to decide whether the claims in each piece of evidence are credible, and what proof there is that one piece is credible. We need to state our own opinions, and rate the evidence in categories of science, methods, credibility, and usefulness.

This pair scored in the top quartile of overall critique quality, and exhibited high autonomy, high elaboration, no uncodable responses, high project-specific plans, and no non-reflective responses to Checking Our Understanding prompts. In planning for critiquing evidence, they laid out a
specific plan, delineating important areas of concentration. One of these students, S107, was interviewed for this research. In Interview 2, students were asked to rank order some student critiques of the "Rooms" evidence (see Appendix K for the interview script; see Appendix F for the evidence itself). S107 used good criteria in ordering the critiques, putting critique #5 at the top:

S107: "... not only because of the quantity but they explained it more thoroughly, and they related their evidence to everyday life, and they considered all the variables, all the different things that could've happened while this experiment was being done."

S107 seemed to see the importance of controlling variables (the most salient aspect of the "Rooms" evidence) and ordered the critiques using that as one of several good criteria.

Another pair, who did not respond at all to the Thinking Ahead prompt for critiquing evidence, made the following response to a Thinking Ahead prompt for critiquing claims:

S312 & 329, Thinking Ahead response: To do a good job on our claim critiques, we need to make sure we thoroughly read and understand the claims, and we also need to make sure that we critique them fairly.

Their response, in contrast to S107 and S121's, is much less specific. This pair also had high autonomy, but lower elaboration, some uncodable and non-reflective responses, and fewer project-specific plans. They achieved an overall critique quality score in the bottom quartile. S312 was also an interviewee; this student, in contrast to S107, did not seem concerned with the idea of controlling variables. This student considered critique #4 the best, but rather than presenting it as pointing out variables to be controlled, S312 instead said it was "basically what I wrote" and that it gave enough "information" without being too long. (In fact, their critique of the methods used in this evidence was, "[T]he methods for this evidence are not very good, because they don't give very much info.")

**Generic Prompt Condition**

In predicting overall critique quality for the generic prompt condition, four variables account for 42% of the variance ($F[4, 37] = 8.389$, $R^2 = .690$, adjusted $R^2 = .419$, $p < .0001$; see Table 8–2 for regression coefficients). The three variables most predictive of critique quality in the generic prompt condition were (a) the autonomy of the pair, (b) the degree to which they focused on content, and (c) the total number of cites they made for claim 3.
<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Standard Coefficient</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>11.913</td>
<td>3.362</td>
<td>11.913</td>
<td>3.544</td>
<td>.0011</td>
</tr>
<tr>
<td>Elaboration</td>
<td>0.082</td>
<td>0.095</td>
<td>0.117</td>
<td>0.868</td>
<td>.3907</td>
</tr>
<tr>
<td>Content-focused responses</td>
<td>Ð22.502</td>
<td>9.594</td>
<td>Ð0.286</td>
<td>Ð2.345</td>
<td>.0245</td>
</tr>
<tr>
<td>Cites (claim 3)</td>
<td>2.096</td>
<td>0.743</td>
<td>0.368</td>
<td>2.819</td>
<td>.0077</td>
</tr>
<tr>
<td>Pair Autonomy</td>
<td>1.006</td>
<td>0.366</td>
<td>0.346</td>
<td>2.751</td>
<td>.0091</td>
</tr>
</tbody>
</table>

Table 8–2: Regression Coefficients for Predicting Overall Critique Quality for Generic Prompt Condition

Focusing on content in response to generic prompts was related to poor critiques of the evidence and claims, perhaps because generic prompts lack an evaluative context. Citing many ideas for claim 3 (the thermal equilibrium claim) was linked positively to overall critique quality, however. Since claim 3 was the harder of the two claims, citing many cites for it may help students create better overall critiques—or, it may have just helped them critique claim 3 itself better. (Surprisingly, no aspect of claim notes cites were found to be very predictive of critique quality for students in the directed prompt condition.) Pairs’ beliefs, in particular about autonomy, were positively related to overall critique quality for the generic prompt condition, unlike for the directed prompt condition, where the type of reflection was more important than beliefs in predicting overall critique quality. Autonomous students may make sense of generic prompts more easily than do less autonomous students. They may use the prompts as an opportunity to reflect on a broad range of ideas and to identify problems with their own thinking. The role of specific types of reflection in response to generic prompts was less important in predicting overall critique quality, except in the case of focusing on content, which had a strong negative effect.

A fourth variable, the average number of words in response to generic prompts, was somewhat positively related to overall critique quality. This variable was highly correlated with claim 3 cites, so it is unclear which, if either, played a causal role. (Note that the number of words in response to Thinking Ahead prompts was a significant predictor of overall critique quality for students in the directed prompt condition, and that the number of words in Checking Our Understanding prompts would also have been had it not been so closely aligned with the amount of lack of reflection in response to Checking Our Understanding prompts.)

Summary of Predictors

Table 8–3 summarizes the variables most predictive of overall critique quality for the two conditions. We see that the models for the two conditions have
little in common. Degree of elaboration is important in both cases—but it is far less important for the generic prompt condition than for the directed prompt condition.

<table>
<thead>
<tr>
<th></th>
<th>Directed Prompt Condition (standard coefficients)</th>
<th>Generic Prompt Condition (standard coefficients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elaboration</td>
<td>.321*</td>
<td>.117</td>
</tr>
<tr>
<td>Uncodable</td>
<td>-.390</td>
<td>no role</td>
</tr>
<tr>
<td>Project-specific plans</td>
<td>.314</td>
<td>no role</td>
</tr>
<tr>
<td>Non-reflective monitoring</td>
<td>-.255</td>
<td>no role</td>
</tr>
<tr>
<td>Content focus</td>
<td>no role</td>
<td>-.286</td>
</tr>
<tr>
<td>Cites (claim 3)</td>
<td>no role</td>
<td>.368</td>
</tr>
<tr>
<td>Pair Autonomy</td>
<td>no role</td>
<td>.346</td>
</tr>
<tr>
<td><strong>Adjusted R²</strong></td>
<td><strong>.360</strong></td>
<td><strong>.419</strong></td>
</tr>
</tbody>
</table>

Table 8-3: Predicting Overall Critique Quality for Directed and Generic Prompts

* = number of words in responses to Thinking Ahead prompts only

Otherwise, we see that success with the two types of reflection prompts has quite different predictors, in that kinds of responses that are highly predictive in one case play no role at all in the other case. Furthermore, the two most predictive variables for the generic prompt condition (claim 3 cites and pair autonomy) are not directly related to prompt type at all. While students may make more cites as an indirect result of the kind of prompts they received, their autonomy is a characteristic of the students. Autonomy appears to play a large role in how students interpret—and benefit from—generic prompts.

**Predicting Coherence**

We turn next to predicting the coherence of students’ ideas in the two conditions.

**Directed Prompt Condition**

For the directed prompt condition, the four variables identified as most important were (a) autonomy, (b) proportion of uncodable responses, (c) proportion of Thinking Ahead responses focused on general plans, and (d) number of principles cited for the two claims, combined. These four variables accounted for 36% of the variance of students’ coherence scores ($F[4, 34] = 6.271$, $R^2 = .425$, adjusted $R^2 = .357$, $p = .0007$; see Table 8-4 for regression coefficients).
Table 8-4: Regression Coefficients for Predicting Coherence for Directed Prompt Condition

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Standard Coefficient</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.007</td>
<td>0.466</td>
<td>0.007</td>
<td>0.015</td>
<td>.9880</td>
</tr>
<tr>
<td>Pair Autonomy</td>
<td>0.144</td>
<td>0.050</td>
<td>0.381</td>
<td>2.865</td>
<td>.0071</td>
</tr>
<tr>
<td>Principle cites (both claims)</td>
<td>0.162</td>
<td>0.080</td>
<td>0.267</td>
<td>2.028</td>
<td>.0507</td>
</tr>
<tr>
<td>Uncodable responses</td>
<td>-12.355</td>
<td>4.685</td>
<td>-0.348</td>
<td>-2.637</td>
<td>.0125</td>
</tr>
<tr>
<td>General goal focus resp. (TA)</td>
<td>-2.711</td>
<td>1.154</td>
<td>-0.309</td>
<td>-2.349</td>
<td>.0248</td>
</tr>
</tbody>
</table>

Writing uncodable (cryptic and/or irrelevant) responses to any prompt and writing plans focusing on very general goals were both linked to poorly integrated ideas for directed prompts. (Students who write general plans may not identify knowledge integration as a goal, or, they may simply have a propensity toward taking the path of least resistance.) As expected, citing many principles for claim notes was positively linked to coherent ideas. Beliefs also played a role; pairs who exhibited highly productive beliefs at the beginning of the semester about the autonomy appropriate for learning science and who received directed prompts were more likely to develop coherent ideas than were students with less productive beliefs. Interestingly, before beliefs were added to the model, focusing on content in response to Checking Our Understanding prompts appeared as a positive predictor of coherence of ideas. However, its effects are overshadowed by the role beliefs play. Here, unlike for predicting overall critique quality for the directed prompt condition, we see that characteristics of the students are at least as important as characteristics of the focuses of their reflection.

Part of the letter from a pair who matches some of the characteristics of the predictive model exemplifies the coherence students with these characteristics can achieve:

S107 & S121, letter: For the claim "some materials are naturally cold" we thought this claim was semi-valid. It didn't really explain what makes marble and metal benches colder. It also didn't really prove that if you stayed in a room with lots of marble and metal in it, you would stay cooler. When improving this claim, we think that you should include evidence proving how the marble and metal benches would be colder and how it would make the room cooler.

The scientific ideas for this evidence involve the fact that even though the room temperature feels the same, the material of the object affects and determines how it functions (absorption, attraction, etc.) in a certain climate. The methods of this evidence show that even if some surfaces feel different than others, their temperature remains the same. It sometimes depends on how the object will react to different amounts of temperature. The credibility for this evidence is very high because you can relate to it very easily, as with the use of park benches of different materials in the same climate. This is also very believable because the same thing happens to me at the beach (the sand is hot and the grass and towel are cool) as well as with benches of different materials. We think that the evidence is very useful because it shows how materials have different feeling temperatures even in the same climate and
temperature. Picnic benches made of wood feel much more different or warmer than a steel bench, even if they are in the same climate and temperature.

These students, whose work was discussed earlier as well, achieved a high degree of coherence. Though their understanding of the phenomena is not 100% accurate, they attempted to integrate their ideas and did so in productive directions. They were in the top quartile of autonomy, and cited some principles in their claim notes. They exhibited no uncodable reflection prompt responses and did not write many general plans. In Interview 3, toward the end of the semester, S107 demonstrated a fairly well-integrated (though tentative) understanding of energy conversion. The student initially talked about black objects absorbing heat, but later changed the statement to scientifically normative ideas—and unlike some students, appeared to be changing conceptions thoughtfully and purposefully. Other interviewees waffled between "attract" and "absorb" and between "heat" and "light" without recognizing that they were doing so. S107 appears to have knowledge integration as a goal, often making statements like "That doesn't make sense" or "That makes a little more sense."

Generic Prompt Condition

Four variables were very predictive of coherence for the generic prompt condition. Together, the four variables accounted for 34% of the variance (F[4, 34] = 6.326, R² = .400, adjusted R² = .337, p = .0005; see Table 8–5 for regression coefficients). The most predictive variables were (a) the proportion of responses focused on content, (b) the proportion of responses saying ideas were the same, (c) the number of principles cited for the two claims, and (d) autonomy. Note that two of these variables are the same as for the directed prompt condition.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>Standard Coefficient</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.465</td>
<td>0.441</td>
<td>0.465</td>
<td>1.052</td>
<td>.2993</td>
</tr>
<tr>
<td>Pair Autonomy</td>
<td>0.088</td>
<td>0.051</td>
<td>0.239</td>
<td>1.740</td>
<td>.0899</td>
</tr>
<tr>
<td>Content-focused responses</td>
<td>–2.559</td>
<td>1.339</td>
<td>–0.249</td>
<td>–1.912</td>
<td>.0635</td>
</tr>
<tr>
<td>&quot;Same&quot; responses</td>
<td>8.097</td>
<td>4.078</td>
<td>0.254</td>
<td>1.986</td>
<td>.0543</td>
</tr>
<tr>
<td>Principle cites (both claims)</td>
<td>0.195</td>
<td>0.080</td>
<td>0.330</td>
<td>2.444</td>
<td>.0193</td>
</tr>
</tbody>
</table>

Table 8–5: Regression Coefficients for Predicting Coherence for Generic Prompt Condition

Focusing on content in response to generic prompts was negatively related to coherence of ideas, as it was for critique quality, as well. Focusing on content may not be a productive activity if a context and purpose are not made clear. (Remember that focusing on content in response to Checking Our Understanding prompts was somewhat positively related to coherence for
directed prompts, lending credence to the hypothesis that context plays an important role.) As expected, focusing on principles and being autonomous were both positively related to coherence; this was also the case for directed prompts. Surprisingly, saying one's ideas were the same as before was positively related to coherence. However, it should be noted that a very small percentage of responses were coded in this category.

One pair from the generic prompt condition exemplifies this model: They were high in autonomy, never made a "same" response, cited several principles in their claim notes, and were in the middle two quartiles of content-focus for reflection prompts. These students achieved a very high level of coherence:

S729 & S707, letter: Now after reading the third claim, about how materials are naturally cold, we have really begun to question your writers. You should either fire them, or send them to Mr. K's science class for a little lecture. No material in naturally cold. The reason the material is colder then the rest of the room is because heat flows faster into some objects than others. This is because of there different characteristics. Two objects at different tempatures combined there tempatures to make one tempature that is in the middle is a good principle for this claim. We learned this in the equilibrium lab. We would change the claim to say Some materials feel colder then room tempature. The is because that your hand maybe is not as accurate as you think it is. The object might be a room tempature but your hand maybe isn't picking it up, we aren't perfect.

These students link the ideas of thermal equilibrium and rate of heat flow, and add the idea of humans being limited detectors of temperature. S707 was interviewed for this research, and demonstrated a well-integrated understanding of energy conversion in Interview 3, as well.

Summary of Predictors

Table 8–6 synthesizes the variables most predictive of the coherence of students' ideas for the two conditions. Here we see greater similarity between the conditions than we saw for overall critique quality; the autonomy of the pair and the degree to which they cite principles play important roles in predicting coherence for both conditions. Recall that students in the generic prompt condition cited significantly more principles than did students who received directed prompts. Not surprisingly, those students in either condition who cited principles were likely to develop a coherent, integrated understanding of science; in citing more principles, students in the generic prompt condition may have expanded their repertoire of ideas and as a result developed a more coherent understanding of those principles.

11
Table 8–6: Predicting Coherence for Directed and Generic Prompts

Autonomy was only important for the generic prompt condition in predicting overall critique quality, but here it is important for both conditions. Highly autonomous students are likely to develop a coherent understanding regardless of what kind of prompt they receive.

**Predicting Guidelines Quality**

Next, we review predictors of students' ability to develop guidelines for critiquing.

**Directed Prompt Condition**

The two variables most useful in predicting the quality of students' guidelines quality for the directed prompt condition were (a) proportion of uncodable responses and (b) proportion of responses focused on content. These two variables accounted for almost 30% of the variance ($F_{[2, 40]} = 9.842$, $R^2 = .330$, adjusted $R^2 = .296$, $p = .0003$; see Table 8–7 for regression coefficients). Both were linked to poor guidelines.
The three variables most predictive of *guidelines quality* for the generic prompt condition were (a) the proportion of responses focused on project-specific goals, (b) the number of principles cited for the two claims, and (c) the number of everyday experiences cited for the two claims. These variables account for more than 50% of the variance of guidelines quality scores for generic prompts ($F[3, 42] = 17.464, R^2 = .555$, adjusted $R^2 = .523$, $p < .0001$; see Table 8–8 for regression coefficients).

Focusing on project goals was positively related to guidelines quality. Logically, if a student understood the criteria for critiquing emphasized by the project, the student would be likely to create better guidelines. Citing principles was also positively related to this measure of abstraction; principles are basically abstractions of one sort. Emphasizing everyday experiences to the exclusion of more abstract ideas (like principles) was linked to poor abstraction for generic prompts.
Writing guidelines for critiquing essentially represents students' ability to abstract back from the act of critiquing to developing criteria for critiquing. Citing principles and citing everyday experiences may be opposite sides of the same coin. Some students are comfortable with the abstraction required by principles, whereas others desire the concrete nature of everyday experience. Interviews with students made this distinction obvious, in that some students discussed principles and others turned to concrete examples when the conversation became too abstract. Students' propensities toward and abilities to engage in abstraction appear to play an important role for students in the generic prompt condition. A surprising finding is that this is not the case for students in the directed prompt condition.

One pair, whose letter was quoted earlier as a good example of high coherence, also wrote good guidelines and matches fairly closely this model, as well. Their guidelines received a high score; they mention the importance of controlling variables and gathering much data and comparing it with other scientists' results.

S729 & S707, letter: We feel that your writers have a lot of thinking to do after reading our review. There are some guidelines that we think that your writers should follow when writing and collecting data. First, when conducting an experiment you should always use the same environment for the whole experiment to get accurate results, and especially when comparing two things. Second, after collecting data they should always compare the data with other results or other scientists, so they can compare their ideas and maybe find a flaw in your evidence. Third and most important, is that should never make a claim or print evidence in the paper without checking it, or comparing evidence, and they should always perform the experiment twice.

The students wrote some project-specific plans (they fell in the middle two quartiles of this score), they cited many principles, and cited no everyday experiences in their claim notes. They appear to be comfortable with the idea of abstracting, which might help explain their high coherence, as well. On the other hand, S707, in Interview 2, did not develop very good criteria for critiquing, though the student developed very coherent ideas in the interview and in the project, and did show a propensity toward abstraction in the interview, as well. In fact, this pair's overall critique quality score was in the middle two quartiles, rather than being high. S707's understanding of what is entailed in critiquing may have been tenuous; the pair may have together been able to use their abstraction capabilities to develop good guidelines, but their ability to actually critique evidence may have suffered. This may indicate a competence-performance issue. Students may have the ability to develop criteria for critiquing but may lack competence in the act of critiquing.
Summary of Predictors

Table 8–9 synthesizes the predictors of guidelines quality for the two conditions.

<table>
<thead>
<tr>
<th></th>
<th>Directed Prompt Condition (standard coefficients)</th>
<th>Generic Prompt Condition (standard coefficients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content focus</td>
<td>-.250 indirect negative role</td>
<td></td>
</tr>
<tr>
<td>Uncodable</td>
<td>-.538 no role</td>
<td></td>
</tr>
<tr>
<td>Project-specific goals</td>
<td>no role</td>
<td>.602</td>
</tr>
<tr>
<td>Principle cites (both claims)</td>
<td>no role</td>
<td>.563</td>
</tr>
<tr>
<td>Everyday experience cites</td>
<td>no role</td>
<td>-.176</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>.296</td>
<td>.523</td>
</tr>
</tbody>
</table>

Table 8–9: Predicting Guidelines Quality for Directed and Generic Prompts

Here again, we find evidence that the two types of reflection prompts have very different predictors of their success. No variables span both models. However, focusing on content, for the generic prompt condition, was positively related to citing everyday experiences and negatively related to focusing on project goals, so content focus may have an indirect negative effect in both conditions. In neither condition was any aspect of beliefs a predictor.

Identifying Groups of Students

The multiple regressions, combined with the individual correlative analyses presented in the previous chapters, can be synthesized to help identify promising groups of students for analysis.

First, we might ask, what is the role of beliefs? Autonomous students who received generic prompts created better critiques, and those who received either kind of prompts tended to develop more coherent ideas than did less autonomous students. For the generic prompt condition, there was a much more obvious effect of beliefs than there was for the directed prompt condition. One way to group students, then, is according to their autonomy. This allows comparison of the groups’ performance on the project.

What is the role of citing principles? For both conditions, citing principles was positively related to coherence. Thinking about scientific principles appears to help students develop a coherent, integrated understanding of science through helping them expand their repertoire of ideas. It may also be
that being able to use principles gives students more to draw on in terms of making criticisms of the evidence and claims. For the generic prompt condition, citing principles was also positively related to good guidelines. We might hypothesize that students who can use abstractions (principles) would also be able to make abstractions. Students who seem facile with citing principles represent another group of students to investigate more closely.

What is the role of elaboration? In general, elaborating in response to reflection prompts was a positive activity. However, it is not yet clear whether the act of elaborating one’s reflection is causally related to success, or if it is a correlation that people who elaborate also do other good things. To help us shed more light on this issue, highly elaborative students could be compared with more terse students for another interesting analysis.

What is the role of writing uncodable responses? For the directed prompt condition only, being cryptic or saying irrelevant things was a very strong predictor of poor performance on the quality measures. This may be explained by students being either inarticulate or unable to make sensible links—and in either case, they would be likely to perform poorly on the quality measures. Cryptic students may bear further investigation, even though they represent a very small part of the overall student population.

Finally, what is the role of being unreflective? Students who were non-reflective in response to directed prompts—who said there was "no problem" with their level of understanding—were likely to develop poor critiques, whereas students who were non-reflective in response to generic prompts—who said their ideas were the same as they had been earlier—were likely to develop coherent ideas. However, students were far more likely to be unreflective in response to Checking Our Understanding prompts than in response to generic prompts; about 21% of all responses to Checking Our Understanding prompts were unreflective, whereas less than 1% of all generic prompt responses were coded as "same." Non-reflectors represent another interesting group for further analysis.

Investigating Groups of Students

In this section, I investigate how groups of students with similar salient characteristics performed on the project. The groups include autonomous learners, principle-citers, elaborators, cryptic students, and non-reflectors. These groups are not intended to be mutually exclusive. The only groups
with a notable degree of overlap are the low autonomy students and the low principle-citers.¹

To preview the findings, we will see first that autonomous learners developed significantly more coherent ideas than did low autonomy students. Second, principle-citers developed significantly more coherent ideas and wrote significantly better guidelines than did students who cited few principles. Third, elaborators did significantly better at critiquing than did those who were terse in response to reflection prompts. Fourth, cryptic students did significantly worse at critiquing and guidelines. Those who wrote uncodable responses to directed prompts in particular were significantly more likely to do poorly on their guidelines than were those who did not write uncodable responses or did so in response to generic prompts. Finally, non-reflectors responding to directed prompts did significantly worse at critiquing than did reflective people or people who were non-reflective in response to generic prompts.

At a higher level, we see evidence for the hypothesis that directed prompts can have quite positive effects when certain kinds of actions are taken, but strong negative effects with other kinds of actions. Generic prompts, on the other hand, tend to be more robustly positive. Generic prompts also have a few quite strong positive effects, in particular for autonomous students.

The Autonomous Learners

In investigating the autonomous learners, I compared students in the top quartile, middle two quartiles, and bottom quartile of autonomy scores. The multiple regressions lead us to expect that highly autonomous students may develop coherent ideas regardless of prompts, and may develop better critiques when they receive generic prompts.

The ANOVAs indicated no significant differences in the overall critique quality scores of students of different levels of autonomy, regardless of condition. However, students of low autonomy developed significantly less coherent ideas than did students of middle or high autonomy (F[2, 80] = 4.604, p = .0128). The effect sizes are quite large: 74% for the low autonomy versus middle autonomy comparison, and 86% for the low/high comparison. That is, the middle and high groups did 74% and 86% of a standard deviation, respectively, better at developing coherent ideas than did the lower group.

¹Of the 19 low autonomy pairs, 11 were also in the bottom quartile of principle citers. (Of these 11, four pairs were in the directed prompt condition, and seven were in the generic prompt condition.)
Table 8–10 gives the mean differences and effect sizes, and Figure 8–1 shows the means.

<table>
<thead>
<tr>
<th>Groups Compared</th>
<th>Mean Difference</th>
<th>p-value (Fisher’s PLSD)</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low versus Middle</td>
<td>−.611</td>
<td>.0097</td>
<td>74%</td>
</tr>
<tr>
<td>Low versus High</td>
<td>−.711</td>
<td>.0065</td>
<td>86%</td>
</tr>
<tr>
<td>Middle versus High</td>
<td>−.100</td>
<td>NS</td>
<td>12%</td>
</tr>
</tbody>
</table>

Table 8–10: Mean Differences in Coherence for Autonomy Groups
\((\sigma^2 = 0.684; N: \text{Low} = 19, \text{Middle} = 40, \text{High} = 24)\)

Though the interaction between autonomy and condition was not significant, the effect looks especially pronounced for the generic prompt condition. Generic prompts may help middle autonomy students, in particular, perform at higher levels of coherence than directed prompts do. The effect size between the two middle-autonomy groups is 60%—students of mid-range autonomy who received generic prompts did about 60% of a standard deviation better than those who received directed prompts. When we compare the two high autonomy groups, the effect size is 31%. Figure 8–2 shows the means for the interaction.
8 cases were omitted due to missing values.

Figure 8-2: Mean Values of Coherence for Autonomy Groups by Condition (1 = bottom quartile [N: DP = 9; GP = 10], 2 = middle 50% [N: DP = 17; GP = 23], 3 = top quartile[N: DP = 13; GP = 11])

It appears that prompt condition makes no difference for students with very low autonomy. (The effect size was only 3%.) However, students who have a propensity toward autonomy may be better served by generic prompts; they may develop a more coherent, integrated understanding when allowed to take responsibility for directing their own reflection.

One middle-autonomy pair who achieved high coherence wrote particularly nice responses to generic prompts. For example:

S510 & 524, generic prompt response: Our thoughts now are that there is no way that the anti-heat shirt is able to keep you cooler. Wearing lighter clothes and staying away from generally hot objects is a claim that we, as students who have been researching this sort of evidence, are able to easily believe.

These students begin questioning the idea that materials can be naturally cold in this reflection prompt response. In their letter, they achieve a high degree of coherence in explaining what is happening instead:

S510 & 524, letter: With your claim that some materials are naturally colder, this is not true also. The reason that (for example) metal seems colder than wood, is not because it is colder, because it isn’t, but because it is a better energy conductor. When you touch metal with your hand, the heat energy that is collected in your hand flows very rapidly into the metal, and in that way, cools down your hand very quickly. A wood object, when you touch it with your hand, does the same thing as the metal, but much less rapidly. This is also a good claim to include in your article about heat in the summertime. But I would suggest an experiment about how fast the energy flows through the metal and wood.

Generic prompts may have afforded these students, who had some propensity toward autonomy, the opportunity to reflect more broadly on the phenomena and to begin identifying ways in which their ideas might differ from the
claims being made. We see that the effects of reflection in response to prompts propagate through the students' work on the rest of the project.

The Principle-Citers

In investigating the principle-citers, I compared the students in the top quartile, middle two quartiles, and bottom quartile in the number of principle cites in the claim notes. The multiple regressions lead us to expect that principle-citers will do well at coherence (both conditions) and guidelines quality (especially those who received generic prompts). As noted, there is a relatively high degree of overlap for the low autonomy and low principle-citer groups; 11 of the 29 pairs in the bottom quartile of principle-citers were in the bottom quartile of autonomy scores, as well. It is thus difficult to disentangle the effects of being autonomous and citing principles.

The ANOVAs indicated that there was a significant difference in the coherence of ideas developed by students who cited principles differently. In particular, students who cited very few principles developed significantly less coherent ideas than did other students (F[2, 86] = 5.541, p = .0055). Again, the effect sizes are large: 64% for the low/medium group comparison, and 93% for the low/high comparison. Table 8–11 gives the mean differences and effect sizes, and Figure 8–3 shows the means.

<table>
<thead>
<tr>
<th>Groups Compared</th>
<th>Mean Difference</th>
<th>p-value (Fisher's PLSD)</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low versus Middle</td>
<td>−.524</td>
<td>.0089</td>
<td>64%</td>
</tr>
<tr>
<td>Low versus High</td>
<td>−.763</td>
<td>.0036</td>
<td>93%</td>
</tr>
<tr>
<td>Middle versus High</td>
<td>−.239</td>
<td>NS</td>
<td>29%</td>
</tr>
</tbody>
</table>

Table 8–11: Mean Differences in Coherence for Principle-Citing Groups
\((\sigma^2 = 0.670; N: \text{Low} = 29, \text{Middle} = 44, \text{High} = 16)\)

Since students work in pairs, students who cite principles expand their own repertoire of ideas or at least make their existing repertoire "visible" and available for use. As a result, they can integrate those ideas with others, especially if they take advantage of opportunities to identify places where they could distinguish, link, and reorganize their ideas.

---

2Though citing principles and developing coherent ideas seem conceptually tied, the measures are independent since the number of principles cited refers to the students’ claim notes, and the coherence refers to their letters.
Additionally, students who cited very few principles were significantly worse at writing guidelines than were other students, as measured by their guidelines quality scores ($F[2, 86] = 3.217, p = .0449$). The effect size comparing the low and high principle-citing groups was 72%, and comparing the low and middle groups, the effect size was 48%. Table 8–12 presents the mean differences and effect sizes.

<table>
<thead>
<tr>
<th>Groups Compared</th>
<th>Mean Difference</th>
<th>p-value (Fisher's PLSD)</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low versus Middle</td>
<td>-.362</td>
<td>.0500</td>
<td>48%</td>
</tr>
<tr>
<td>Low versus High</td>
<td>-.550</td>
<td>.0229</td>
<td>72%</td>
</tr>
<tr>
<td>Middle versus High</td>
<td>-.188</td>
<td>NS</td>
<td>25%</td>
</tr>
</tbody>
</table>

Table 8–12: Mean Differences in Guidelines Quality for Principle-Citing Groups
($\sigma^2 = .580$; N: Low = 29, Middle = 44, High = 16)

In neither case was the interaction with condition significant, perhaps because there were only 5 pairs in the high principle-citing, directed prompt cell. (See Appendix O for the interaction analyses.)

One pair from the generic prompt condition provides an example of these links between principle-citing and both coherence and guidelines quality. This pair's letter was quoted earlier in this chapter as showing high levels of

Figure 8–3: Mean Values of Coherence for Principle-Citing Groups
(1 = bottom quartile [N = 29], 2 = middle 50% [N = 44], 3 = top quartile [N = 16])

2 cases were omitted due to missing values.
coherence and guidelines quality. Here is one of their claim notes, in which they cite a principle:

**S729 & S707, claim 1, validity note:** In thinking about the validity of this claim, we think that this claim is wrong, because in the claim it says that black attracts heat, but black actually absorbs light more than any other color and changes it to heat energy.

These students may have the ability to abstract and may apply it both to creating guidelines for critiquing and to using principles in explaining scientific phenomena.

One of these students, S707, was interviewed. The student demonstrated a coherent understanding of energy conversion in Interview 3. (See Appendix L for Interview 3 script.) The student also showed an interesting ability and propensity to use concrete examples as demonstrations of the abstractions being made (rather than relying on concrete examples to the exclusion of scientific principles). For example, when asked for a sentence using the words black, white, and reflect, the student said "The black paint on the house did not reflect as much light as the white paint did." Later, when asked to explain what was meant by a certain sentence debunking the "attract" idea, the student made a concrete visual representation, linking the sun's rays to raindrops and pointing out that neither one could change direction. This student—and perhaps the student's partner—seems to have the ability to employ principles as abstractions but to keep the meaning associated with them by linking them (appropriately) to concrete examples. Furthermore, they expanded their repertoire of ideas by citing principles and other ideas. They distinguished between ideas appropriately—between attract and absorb, for example—and made links between the ideas as well—here, through an analogy between sunlight and raindrops.

The Elaborators

In investigating the elaborators, I compared students in the top quartile, middle two quartiles, and bottom quartile of the number of words in responses to prompts. The multiple regressions lead us to expect that students who elaborate their responses may do better at critiquing, regardless of prompt type.

The post-hoc analyses comparing each group indicate that students who elaborated their responses did, in fact, achieve significantly higher overall critique quality scores than did those who were less elaborative (Fisher's PLSD p = .0311 for low elaboration versus high elaboration; p = .0459 for middle versus high). Table 8–13 shows the mean differences and effect sizes.
<table>
<thead>
<tr>
<th>Groups Compared</th>
<th>Mean Difference</th>
<th>p-value (Fisher's PLSD)</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low versus Middle</td>
<td>−1.067</td>
<td>NS</td>
<td>14%</td>
</tr>
<tr>
<td>Low versus High</td>
<td>−4.957</td>
<td>.0311</td>
<td>66%</td>
</tr>
<tr>
<td>Middle versus High</td>
<td>−3.890</td>
<td>.0459</td>
<td>52%</td>
</tr>
</tbody>
</table>

Table 8–13: Mean Differences in Overall Critique Quality for Elaboration Groups
\((\sigma^2 = 56.137; N: \text{Low} = 21, \text{Middle} = 45, \text{High} = 23)\)

The interaction with condition was not significant, but elaboration may be the biggest boon for those in the directed prompt condition. Table 8–14 and Figure 8–4 show the means for the interaction.

<table>
<thead>
<tr>
<th></th>
<th>Directed Prompt Condition</th>
<th>Generic Prompt Condition</th>
<th>Group Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1: Low elaborators</td>
<td>20.7 (N = 11)</td>
<td>21.3 (N = 10)</td>
<td>21.0</td>
</tr>
<tr>
<td>Group 2: Medium elaborators</td>
<td>20.5 (N = 22)</td>
<td>23.6 (N = 23)</td>
<td>22.1</td>
</tr>
<tr>
<td>Group 3: High elaborators</td>
<td>27.9 (N = 10)</td>
<td>24.5 (N = 13)</td>
<td>26.0</td>
</tr>
<tr>
<td><strong>Condition Mean</strong></td>
<td><strong>22.3</strong></td>
<td><strong>23.3</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 8–14: Mean Values of Overall Critique Quality for Elaboration Groups by Condition
\((\sigma^2 = 55.992)\)
Interaction Bar Plot for OverallCritiqueQual.
Effect: Words %ile Group * Cond.
Error Bars: ± 1 Standard Deviation(s)

2 cases were omitted due to missing values.

Figure 8–4: Mean Values of Overall Critique Quality for Elaboration Groups by Condition
(1 = bottom quartile [N: DP = 11; GP = 10], 2 = middle 50% [N: DP = 22; GP = 23],
3 = top quartile [N: DP = 10; GP = 13])

We see that highly elaborative students responding to directed prompts tended to write better critiques than did any other group of students, with an effect size of 46% of a standard deviation above the highly elaborative generic prompt students, and a whole standard deviation above the middle group of directed prompt students. This is an indication that elaborating responses to directed prompts can be a very productive activity. We will see how this finding fits into a larger pattern demonstrating the importance of engaging in reflection.

These students who elaborate take advantage of the opportunities inherent in the directed prompts to reflect—they identify weaknesses with their understanding and places where distinctions and links can be made. In the multiple regression analyses, in fact, we saw that elaborating in response to both types of directed prompts was highly predictive of successful critiquing, but that elaboration of Checking Our Understanding prompts was also negatively correlated with non-reflective responses to those prompts. Encouraging elaboration in response to directed prompts, if non-reflective responses to Checking Our Understanding prompts could be discouraged, would be an especially appropriate instructional goal. However, as we will see, the non-reflective responses can be particularly problematic.
Compare 2 pairs’ responses to the same Checking Our Understanding prompt:

**S410 & S421, Checking Our Understanding response:** Claims in the article we didn't understand very well included nothing, we understand basically all 3 claims.

**S107 & S121, Checking Our Understanding response:** Claims in the article we didn't understand very well included the fact that if you were in a room with lots of marble and metal in it, you would stay cooler. Perhaps if you were touching something marble or metal, you would stay cooler, but I'm not so sure what the results would be in a room. Also, the anti-heat shirt confuses me. What chemical is this shirt dipped in and how can it affect one's body temperature? How does it work? Does such a thing really exist?

The first pair writes a terse (and non-reflective) response to the call for monitoring, and develops a poor critique. The second pair, by contrast, writes an elaborated response to the prompt, and develops a very high-quality critique. We can see the second pair starting to critique the evidence and claims even here in their reflection prompt response; the prompt primes them to think about the kinds of things they will include in their critique later, and in fact the effects of that reflection are seen later in the project itself.

An interview with S107 was discussed earlier in this chapter in the context of coherence. This student also developed and used good criteria for critiquing after the project. S421 was also interviewed. This student actually had relatively sophisticated criteria for critiquing, pointing out that some people critiquing the evidence just “restated what he [the experimenter] did,” and that that was necessary but not sufficient in critiquing. One high-quality critique, in S421’s opinion, "listed a lot of things that they could do to make it better." The student may have developed these sophisticated ideas about critiquing after the project was over, or may have simply not applied these good ideas about and ability to critique to the critiquing in the project itself.

*The Cryptics*

In investigating the cryptic students, because so few pairs made cryptic responses, I compared students who made any uncodable responses with those who never made uncodable responses to prompts. The multiple regressions lead us to expect that directed prompt students who respond in an uncodable way may do worse on all measures.

The ANOVAs indicated no differences in the coherence of students' ideas when they made various levels of uncodable responses, regardless of condition. However, students who made uncodable responses achieved significantly lower overall critique quality scores than those who did not (F[1, 87] = 5.120, p = .0261), with an effect size of 55%.
Students who wrote uncodable responses also were significantly worse at writing guidelines, as measured by their guidelines quality scores, with an effect size of 49% ($F[1, 88] = 4.171, p = .0441$). The interaction with condition was also significant ($F[1, 1, 1, 86] = 9.933, p = .0022$ for the interaction). Students who wrote uncodable responses to directed prompts were significantly more likely to do poorly on their guidelines than those who (a) did not write uncodable responses (effect size 131%) or (b) did so in response to generic prompts (effect size 130%). Table 8–15 and Figure 8–5 show the means for the interaction.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Directed Prompt Mean</th>
<th>Generic Prompt Mean</th>
<th>Group Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 0: Not cryptic</td>
<td>1.6 (N = 32)</td>
<td>1.4 (N = 35)</td>
<td>1.5</td>
</tr>
<tr>
<td>Group 1: Cryptic</td>
<td>0.6 (N = 11)</td>
<td>1.6 (N = 12)</td>
<td>1.1</td>
</tr>
<tr>
<td>Condition Mean</td>
<td>1.3</td>
<td>1.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 8–15: Mean Values of Guidelines Quality for Uncodable Groups by Condition ($\sigma^2 = 0.532$)

Here, taking a certain action (being cryptic or making irrelevant points) can have a strong negative effect for the directed prompts but no effect for the generic prompts. A similar trend was apparent for the quality of students'
Students who write cryptic, irrelevant, uncodable responses to directed prompts develop a less complete understanding of what it means to critique and may be less able to do critiquing, as well. Writing cryptic responses to generic prompts may play a smaller role in students developing an understanding of and ability to critique because they are reflecting in directions of their own choice. When students receive directed prompts, though, they are directly given the opportunity to make explicit the kinds of things they need to do to critique. Not taking advantage of this opportunity may be more harmful than choosing one’s own focus for reflection, even if that choice is equally cryptic or irrelevant. Again, we see that some students take advantage of opportunities to reflect and benefit from the reflection they engage in, while others disregard those opportunities—and their performance suffers. For these students, the directed prompts appear actually to derail the reflection; again, we see effects (here, the negative effects) of reflection propagating through the project.

The Non-Reflectors

Lastly, in investigating students who were not reflective, I compared students with an above-average proportion of non-reflective responses to Checking Our Understanding prompts or with any "same" responses with the rest of the students. The multiple regressions lead us to expect that these students will do poorly at critiquing (if they received directed prompts) and well at coherence (if they received generic prompts).

The ANOVAs indicated no significant differences in the coherence of ideas developed by reflective and non-reflective students, regardless of condition. It appears that the proportion of the population who said their ideas were the same as before was large enough to cause a significant effect in the multiple regression analysis, but small enough to be negligible in these analyses.

However, highly non-reflective people responding to directed prompts achieved significantly lower overall critique quality scores than did (a) reflective people (effect size 85%, for comparison with reflective directed prompt responders) or (b) people who were non-reflective in response to generic prompts (effect size 78%; F[1, 1, 1, 85] = 4.757, p = .0319 for the interaction between non-reflective responses and condition). Table 8–16 and Figure 8–6 show the interaction means.
Non-reflective students who inappropriately monitored their understanding were less likely to develop high-quality critiques than other students. Why might monitoring poorly play a role in overall critique quality? Students who do not see problems in their own understanding might also be less likely to see problems in evidence and claims they have been asked to critique. These happy-go-lucky students may be relativistic—seeing all answers as equally valid—or they may simply be overly optimistic. An alternative hypothesis, and one consistent with the findings regarding the elaborators and the cryptic students, is that some students do not take opportunities to reflect. These students then have fewer chances to identify weaknesses in their own understanding (of evidence, claims, scientific ideas) than do other (more reflective) students and thus are less successful at
critiquing and at developing a coherent understanding. Since relatively few students in the directed prompt condition actually developed a coherent understanding—only 11 pairs of students attained scores of 2 or 3—further research is necessary to see if students who are particularly poor at self-monitoring tend to develop less coherent understandings.

In any case we should work to help these students develop better monitoring skills, and help them see that monitoring can be applied to both their own understanding and to information "out there" in the world. One example of a pair who might benefit from this emphasis on monitoring is quoted here:

*S323 & S324, Checking Our Understanding response: Pieces of evidence we didn't understand very well included the supplex nylon and how it worked and if it really works. Everything else we completely understood.

These students note one idea they do not understand, but otherwise assess their understanding as flawless. They then go on to develop a critique of the evidence and claims in the bottom quartile of scores.

**Discussion and Implications**

I began this chapter with a caveat that the analyses presented earlier were correlative, rather than causal, in nature. The multiple regressions and ANOVAs do not give us causal answers, either, but they do take us closer. This discussion section is intended to consider reasonable hypotheses about possible causality, and simultaneously to point to times when correlation is a more apt inference.

*Overview of Reflection Prompts’ Effects*

We see that there may be more negative types of reflection as a result of directed prompts. For example, students were more likely to be non-reflective for directed prompts than for generic prompts. Also, certain focuses may be more "harmful" when used in response to directed prompts than when they were used for generic prompts. For example, students were cryptic equally often (and very infrequently) for the two types of prompts (about 2% in each case), but being cryptic was related to poor critiquing and guidelines for the directed prompt condition, whereas it played no role for the generic prompt condition. This may indicate that students are better served choosing their own focus for reflection than trying—and failing—to reflect in a direction chosen for them. In both of these instances, when students engaged in a certain type of reflection (or lack thereof) in response to directed prompts, their performance was significantly diminished. When they reflected that way in response to generic prompts, there was no ill effect. Combined with
the fact that directed prompts in fact elicited some of these "bad" focuses more—most notably, the lack of reflection—we have reason to be wary of directed prompts. In particular, directed prompts may actually hinder these non-reflective students in identifying weaknesses in their current knowledge. Without that as a catalyst, the other processes of knowledge integration remain inert.

However, remember that one group of students—those who were highly elaborative in response to directed prompts—performed significantly better than did their counterparts who received generic prompts. Thus we should not discount directed prompts out of hand, but perhaps should work to identify ways to improve the prompts or their delivery or help students to use them in the more productive ways.

Generic prompts may enable students to make productive responses more than directed prompts do. However, the predictors of quality measures for the generic prompt condition were most often beliefs (autonomy in particular) or cites made, rather than reflection focus. Generic prompts may not afford particularly positive or negative reflection. Success with the generic prompts may be more strongly tied to student characteristics. Students with certain characteristics like autonomy may have a greater propensity toward knowledge integration. But, those students are more successful when given generic prompts than when given directed prompts. For these students, generic prompts may allow productive reflection in any direction, priming the students to identify weaknesses in their knowledge and helping them see places where they should distinguish among and link ideas throughout the rest of the project.

Relating Autonomy, Principle-citing, and Elaboration

Why is elaboration only predictive of overall critique quality? Elaboration allows students to reflect both more deeply and more broadly. Students who elaborate on their ideas are likely to identify aspects of the evidence and claims they perceive should be changed, as well as developing a better understanding of why those particular changes are necessary. Elaboration may be more closely linked to the act of critiquing than that of thinking about critiquing or developing coherent ideas because successful critiquing depends more on identification of problems, which is afforded by elaboration. Elaboration does not necessarily entail making links, which is more necessary for success on the other two quality measures.

A related question is, Why is autonomy more predictive of coherence than of the other quality measures? Highly autonomous students, or indeed students with any propensity toward autonomy at all, may develop more coherent
ideas because they are simply more comfortable with the idea of being responsible for their own learning. Two autonomous students paired together may achieve a kind of synergy: Because they both expect to develop an understanding themselves, they may find that collaborating brings twice as many ideas to the fore, that they can make more distinctions and links, and that their understanding benefits as a result. Low autonomy students put the responsibility for their learning on others. Often, in a classroom emphasizing collaboration, this is a successful strategy—low autonomy students can sit back and relax while their partners or groupmates do all the work. When two low autonomy students are paired, though, the strategy fails. An important question, then, becomes, How can we help low autonomy students become more autonomous? The CLP/KIE curriculum significantly improved students’ beliefs about autonomy and represents one approach to solving this problem. Students involved in a curriculum like this may add “personal responsibility” and “sense-making” to their repertoire of models of learning.

Why are autonomous students more successful with generic prompts? This may be because they are more able than other students to identify for themselves productive focuses for their reflection. Autonomous students are used to thinking for themselves and have developed productive strategies for doing so. These students regularly identify weaknesses in their own knowledge, distinguish among ideas, and make links appropriately; generic prompts allow them to do those processes explicitly—visibly—rather than implicitly. These students may even have their own more productive reflection interrupted by directed prompts. Less autonomous students may benefit more from the scaffolding provided in the directed prompts, though these may also constrain their reflection and may even lead some students to reflect in unproductive ways. Lower autonomy students are less likely to identify weaknesses in their own knowledge. They expect others to identify weaknesses for them. Neither kind of prompt is particularly helpful for these low autonomy students because both kinds of reflection prompts require the students to take some responsibility for their own learning; however, modeling regular reflection for these low autonomy students may have important long-term effects. Further research would be required to investigate this hypothesis.

**Considering Prompt Condition and Lack of Reflection**

Why is being non-reflective more problematic for students who received directed prompts? First, directed prompts elicited a lack of reflection significantly more often than did generic prompts. Checking Our Understanding prompts in particular allow a “no problem” response that is simply less likely to occur to students responding to generic prompts. Given the complexity of the task it is unlikely that these students are actually
monitoring their understanding appropriately. In fact, students who are non-reflective in response to directed prompts are significantly less successful at critiquing. These students may, as discussed earlier, lack the propensity toward identifying problems in general, and this characteristic may play itself out in causing the students to perceive both their understanding and the claims and evidence under scrutiny as flawless. These students are particularly unlikely to identify weaknesses in their knowledge or in "knowledge" they are presented with as evidence.

The effect might alternatively be causal. Students who do not reflect on their own understanding may lack the opportunities that other students take to improve their understanding of the claims and evidence—and in the process, identify aspects of the claims and evidence that should be changed. Directed prompts allow students to explicitly claim complete understanding, absolving them of responsibility for furthering that understanding. Generic prompts, by not affording that non-reflective opportunity, at least give students the chance to reflect in some productive way. Future research might be able to determine the causality by intervening when students assess their understanding as perfect to prompt for further reflection.

Investigating the Complex Role of Reflecting on Content

What is the role of focusing on content? For the directed prompt condition, students who focused on content when asked to assess their understanding were slightly more likely to develop a coherent understanding of the science. Students who focused on content for Checking Our Understanding prompts were also less likely to be non-reflective, and being unreflective was related to low scores on overall critique quality, so indirectly, focusing on content for Checking Our Understanding prompts might also be positively related to critique quality. The directed Checking Our Understanding prompts explicitly ask students to identify weaknesses in their current knowledge. Students who take advantage of that opportunity develop a significantly better understanding of the science than do students who do not. These students are more likely to move on to other processes of knowledge integration, namely distinguishing and integrating ideas. For students in the generic prompt condition, focusing on content without assessing their understanding of it was related to less success at both critiquing and developing coherent ideas. By focusing on content without that orientation toward monitoring, students in the generic prompt condition can mimic the non-reflective students of the directed prompt condition: They can answer the prompt without engaging in any true reflection. The context of assessing one's understanding of content appears to play an important role in determining how productive reflecting on content is.
How did focusing on content relate to guidelines quality? Here, content focus represents a particularly tricky issue. For the directed prompt condition, focusing on content in response to prompts in general (Thinking Ahead prompts or Checking Our Understanding prompts) was negatively related to success at developing guidelines. Focusing on content may distract students from critiquing along the multiple criteria we emphasize—in particular, they might focus on critiquing the science to the exclusion of thinking of methods and credibility. Students may then not be able to articulate what makes a good critique.

This chapter has synthesized the effects of prompt type, reflection focus, ideas cited, and student beliefs. We have seen that directed prompts have some strong positive and strong negative effects on students' work on the project, and that generic prompts have mainly neutral or positive effects. In particular, directed prompts allow students to make a non-reflective type of response that hinders students' knowledge integration processes, while generic prompts both help students expand their repertoire of ideas and help them identify weaknesses with their understanding, thus facilitating the knowledge integration processes. We have seen that individual student characteristics play at least as important a role in students' performance as does the focus of their reflection, and that these individual characteristics may play an even larger role when students receive generic prompts. The final chapter will pull back another step, allowing us to consider the implications of this work.