Comparing the Epistemological Underpinnings of Students’ and Scientists’ Reasoning about Conclusions

Kathleen Hogan, Mark Maglienti

Institute of Ecosystem Studies, Box R, Millbrook, New York 12545

Received 24 January 2000; accepted 7 March 2001

Abstract: This study examined the criteria that middle school students, nonscientist adults, technicians, and scientists used to rate the validity of conclusions drawn by hypothetical students from a set of evidence. The groups’ criteria for evaluating conclusions were considered to be dimensions of their epistemological frameworks regarding how knowledge claims are justified, and as such are integral to their scientific reasoning. Quantitative and qualitative analyses revealed that the responses of students and nonscientists differed from the responses of technicians and scientists, with the major difference being the groups’ relative emphasis on criteria of empirical consistency or plausibility of the conclusions. We argue that the sources of the groups’ differing epistemic criteria rest in their different spheres of cultural practice, and explore implications of this perspective for science teaching and learning. © 2001 John Wiley & Sons, Inc. J Res Sci Teach 38: 663–687, 2001

One of the central tasks of science education is to develop students’ abilities to think scientifically. Some researchers conceptualize this task as helping students bridge gaps between everyday and scientific ways of thinking (Hawkins & Pea, 1987; Reif & Larkin, 1991). One gap is constituted by differences in the substantive content of everyday and scientific explanations, which are well documented in an extensive body of literature on students’ alternative conceptions of natural phenomena (Pfundt & Duit, 1998). Another important gap is between the processes of reasoning that scientists and nonscientists use to build new knowledge. Although scientists and lay people use similar thinking processes—building inferences, making arguments, critiquing claims—they coordinate the subcomponents of everyday thought processes to different levels of specificity and rigor (Klahr & Simon, 1999). Such differences in the two groups’ reasoning processes are evident even after accounting for the effect that domain-specific knowledge has on reasoning (Schunn & Anderson, 1999).

Understanding the nature and also the source of differences in how scientists and nonscientists think can inform efforts in science education to engage students in authentic scientific practices, including reasoning practices. To that end, this article presents results of a study that compared how science professionals (scientists and technicians) and nonscientists

Correspondence to: K. Hogan; E-mail: hogank@ecostudies.org

© 2001 John Wiley & Sons, Inc.
(adults and middle school students) reasoned about a set of conclusions drawn from a given body of observations. This study extends prior research on scientific reasoning in two ways. First, it focuses on the epistemological criteria that actual (not idealized) scientists and nonscientists use to make judgments about the validity of scientific conclusions. Second, it links cognitive perspectives on what reasoning processes look like to sociocultural perspectives on where reasoning practices come from—originating in the cultural practices of various groups. We will argue that it is not so much developmental capabilities as it is cultural preferences that lie at the root of reasoning differences between groups such as students and scientists.

From Cognitive to Sociocultural Perspectives on Scientific Reasoning

Over the past few decades, cognitive scientists doing laboratory studies of scientific reasoning have uncovered some characteristic shortcomings in nonscientists’ reasoning when they undertake scientific tasks such as designing experiments and interpreting data—the kinds of tasks that are central to inquiry-based science instruction. These include failing to generate or test hypotheses systematically, ignoring critical variables, trying to produce effects rather than explain them, and interpreting data in a biased manner to support prior beliefs (Klahr, Fay, & Dunbar, 1993; Kuhn, Amsel, & O’Loughlin, 1988; Kuhn, Garcia-Mila, Zohar, & Andersen, 1995; Schauble, 1990). Although there is a general trend of improvement in nonscientists’ reasoning within scientific discovery contexts from preadolescence through adulthood, performance at all ages remains less than optimal (Kuhn et al., 1988).

Kuhn proposed that weaknesses in lay people’s scientific reasoning stem largely from their failure to perceive their current beliefs as hypotheses that can or should be modified in light of new evidence (Kuhn et al., 1995). When people think only with their theories instead of about them, they cannot adequately consider alternative theories. She attributed this central weakness in reasoning to a lack of procedural knowledge about rules for generating and interpreting evidence, inadequate conditional knowledge about evidence evaluation strategies, and limitations in reflective thought.

Other researchers, however, have criticized Kuhn’s conclusions as being largely a product of her knowledge-lean laboratory tasks that present tricky multivariate complexities, and as stemming from using an outmoded empiricist perspective on the nature of science to construct tasks and interpret results (Fay & Klahr, 1996; Koslowski, 1996; Schauble, 1996; Sodian, Zaitchik, & Carey, 1991). These researchers explain shortcomings in scientific reasoning as having less to do with domain-general processes of hypothetico-deductive thought and associated developmental constraints on reflective abilities, and more to do with the presence or absence of domain-specific knowledge that guides experimental explorations (Carey, 1985; Koslowski, 1996; Sodian et al., 1991).

Both of these cognitive research traditions construct deficit models to explain nonscientists’ reasoning. An emerging alternative to emphasizing what nonscientists cannot do or do not know centers on recognizing that limitations in reasoning processes may be due as much to what people choose not to do. For instance, adults and adolescents who demonstrate use of sophisticated reasoning skills in one context do not necessarily display them in other contexts, depending on their motivations, goals, and the task specifications (Klaczynski, Gordon, & Fauth, 1997; Klaczynski & Narasimham, 1998). People seem to possess both informal heuristics for thinking and more formal rules of logic, and use whichever suits a particular purpose. They are pragmatic in their deployment of skills from different cognitive repertoires, depending on perceived contextual demands.
This points to a need to broaden frameworks for interpreting the results of scientific reasoning studies from skills- and knowledge-based explanations to explanatory frameworks that include contextual and motivational dimensions of cognition. Theorists who have proposed such multifaceted cognitive models (e.g., Brown, Collins, & Duguid, 1989; Pintrich, Marx, & Boyle, 1993; Schoenfeld, 1983) recognize the social roots of cognition (Levine, Resnick, & Higgins, 1993). These enhanced cognitive models resonate with depictions in history, philosophy, sociology, and cognitive studies of science that portray the complex interplay of world phenomena, social context, motivation, knowledge, beliefs, values, methods, and commitments in scientific work (e.g., Boyd, Gaspar, & Trout, 1990; Cole, 1992; Dunbar, 1995; Klahr & Simon, 1999; Longino, 1990; Thagard, 1994; Tweney & Chitwood, 1995). As a socioculturally defined group, scientists have ways of thinking that are shaped by the norms and traditions of their community of practice (Gee, 1992; Wenger, 1998). Thus, to understand differences in how people reason it is necessary to understand differences in their contexts for reasoning, recognizing, for instance, how scientists’ thinking benefits from their membership in the cultural institution of science in ways that nonscientists’ thinking does not (Brewer & Samarapungavan, 1991).

**Linking Social Practices, Epistemology, and Reasoning**

A central tenet of sociocultural theory (Vygotsky, 1978) that we adopt for this study is that cultural activity occurring on the social plane becomes internalized as tools for cognition on the personal plane. We are particularly interested in how the norms and intellectual standards of a community of practice such as science are manifest cognitively as epistemological criteria. Epistemological criteria are one’s philosophical assumptions, beliefs, and theories about the nature and limits of knowledge and its acquisition (Kitchener, 1983). These criteria include both domain-specific and domain-general norms for deciding what assertions to believe, accept, reject, or modify.

Epistemological criteria are important because they can influence reasoning by guiding the executive control processes of planning, monitoring, and evaluating (Hogan, 2000a; Metcalfe & Shimamura, 1994). For instance, epistemological criteria can function as standards for evaluating the validity or plausibility of scientific knowledge claims (Hewson, 1985), which in turn can influence how people respond to anomalous data and undergo conceptual change (Chinn & Brewer 1993; Strike & Posner, 1992). In a study of how 9- to 11-year-old students selected and evaluated information about evolution (Samarapungavan, 1997), their beliefs about the boundaries of knowledge (e.g., that certain kinds of questions are beyond the scope of biological inquiry) functioned as a roadblock at the earliest stages of cognition by preventing them from noticing, receiving, and considering new information that could have helped to change their ideas about evolution to more canonical views.

Other cognitive performances that are influenced by epistemic knowledge and commitments include providing evidentiary backings or rationales for knowledge claims, and proposing tests of claims (Perkins & Simmons, 1988). One study showed a correlation between elementary school students’ use of controlled experimentation strategies and their epistemological beliefs about the nature of scientific knowledge building (Sodian & Schrempp, 1997). Students who believed that scientific activities give investigators direct, unproblematic access to the way the world works were less likely to use controls in their own experiments than were students who understood science as an interpretive enterprise that necessitated designing experiments to rule out alternative interpretations of results.
Such studies indicate that investigating the epistemological dimensions of students’ thinking is a fruitful line of inquiry with potentially important implications for achieving core goals of science education, such as using inquiry processes and building conceptual understanding. Furthermore, epistemic criteria could account for some of the essential differences between everyday and scientific thinking (Reif & Larkin, 1991). For instance, it has been proposed that scientists have exacting epistemic standards for a need for coherence between theoretical models and observations of phenomena, whereas nonscientists do not have the same level of epistemological sensitivity to a need for such coherence (Perkins & Simmons, 1988). Although some of the criteria that scientists and children use to judge explanations (e.g., empirical accuracy, scope) are similar, scientists seem to have both a broader repertoire of criteria and apply them more rigorously (Brewer, Chinn, & Samarapungavan, 1998; Samarapungavan, 1992). However, these tentative claims about differences in epistemological standards are based on comparing nonscientists’ reasoning with scientists’ reasoning as it is depicted in literature on the history and philosophy of science. In the study reported here, we extend this prior research by comparing the actual on-line reasoning of scientists and nonscientists to investigate the role of epistemological standards in scientific thought.

Study Rationale and Overview

We investigated the epistemological criteria that underlie how four types of participants—scientists, technicians, nonscientist adults, and middle school students—judged the validity of conclusions drawn from a set of scientific observations. The study was designed to preserve some of the advantages of traditional cognitive studies, such as the ability to control conditions so that all participants receive the same prompts and the ability to study a larger size and range of populations than would be possible in a naturalistic study, while bringing a broader sociocultural framework to bear on interpreting the results. In the larger scheme of literature development in our field, fine-grained laboratory-based studies such as this can complement naturalistic studies of how students’ (e.g., Hogan & Corey, 2001) and scientists’ (e.g., Bowen, 2000) norms and standards influence their reasoning practices as their work unfolds in context.

During one-to-one interviews, each participant judged 10 conclusions and explained why they rated each as valid or invalid. This created an implicit task demand for them to articulate the standards and criteria that underlie what they value in a scientific conclusion. Our assumption is that these criteria have a role in a person’s own construction of conclusions from evidence, but by asking participants to rate others’ conclusions we directly tapped their epistemic criteria rather than having to infer their criteria from conclusions they drew themselves.

Although epistemic criteria potentially influence all stages of scientific inquiry and knowledge building, the conclusions phase was chosen as the task context because drawing inferences from evidence is one of the most central pursuits of scientific thinking (Koslowski, 1996). Also, judging others’ conclusions is an authentic scientific activity, manifest for instance in peer review (Berkenkotter & Huckin, 1995). Both drawing and evaluating conclusions are also common thinking activities in everyday contexts (Perkins, Faraday, & Bushey, 1991). Our central research questions, then, were: What epistemic criteria do adolescents, nonscientist adults, technicians, and scientists use when reasoning about the validity of conclusions, and what might account for any differences in their criteria?
Methods

Participants

The 45 volunteer participants in this study were 24 eighth graders (9 female, 15 male from working- and middle-class families) drawn from six science classrooms in rural and suburban schools in upstate New York that were participating in a 2-year watershed ecology program, and 21 adults drawn from one workplace in the same general region. The adult sample was composed of 16 (5 male, 11 female) science professionals who were subdivided according to education and experience level into one group of six research scientists with doctorates in ecology or related fields, and one group of 10 technicians with bachelors’ or masters’ degrees in science; and five non-science professionals (all female) with bachelors’ or masters’ degrees in the humanities or business.1 Two of the students were racial minorities (1 African American, 1 Asian Indian), reflecting the low racial diversity of the schools; one nonscientist adult and one technician were Asian American; and all other participants were European American. All of the participants were volunteers who agreed to participate in the study during their free time within the school or work day.

Materials and Procedure

The two authors collected the data for this study through interviews. Before completing the conclusions rating interview task that is the focus of this study, we assessed students’ domain knowledge of ecology (i.e., knowledge about ecological relationships and dynamics) with an aquatic ecosystem food web interview task (for details, see Hogan, 2000b) resulting in categorization of their prior knowledge on a five-level scale from low to high knowledge. In addition, ratings of students’ overall science achievement level was provided by their teachers on a five-level scale from very low to very high. The adults’ relative background in ecology was inferred from their academic and job specializations which they documented on a brief demographic questionnaire.

In the conclusions rating interview task, which lasted 30–40 min, each participant evaluated 10 conclusions that hypothetical students (HS) made based on a given body of evidence. The evidence was a set of observations that a fictitious biology teacher named Mr. Tesser made during his summer work as an environmental research assistant investigating a marsh where purple loosestrife, an exotic species, was growing.

After being introduced to a story scenario about how each fall Mr. Tesser shares his summer data with his students and asks them to write conclusions based on his observations, the participants saw a photograph of purple loosestrife and read a bulleted summary of Mr. Tesser’s observations, which was kept on the table for them to refer to throughout the task. His observations included seeing a wilted brown orchid with sprouts of purple loosestrife growing around it, fewer cattails than the previous summer, a lot of spiders and hopping bugs on the purple loosestrife, and more loosestrife and less open water in the middle of the marsh than were present the previous summer.

Participants then read 10 of Mr. Tesser’s HS’s conclusions, presented one at a time on separate index cards. We crafted the conclusions so that they contained a variety of appropriate and inappropriate inferences from the evidence. Nine of the 10 conclusions contained major weaknesses in reasoning according to contemporary philosophy of science and cognitive perspectives on basic principles and standards for making valid inferences and arguments from evidence (Boyd et al., 1990; Cole, 1992; Giere, 1984; Schauble, 1996)—the
same perspectives and principles that underlie national standards for students’ scientific reasoning [National Research Council (NRC), 1996]. For instance, Conclusion 1 had several major weaknesses: It failed to take into account the full range of Mr. Tesser’s evidence, it extrapolated beyond the evidence, and it made a claim that lacked empirical consistency with the evidence:

Conclusion 1: Purple loosestrife is really pretty, so overall it makes marshes better places. Lots of different kinds of plants will come to live there once purple loosestrife starts growing.

In contrast, Conclusion 5 had no glaring weaknesses because it took all of the empirical evidence into account and made appropriately tentative claims:

Conclusions 5: Fewer native plants, less water habitat, but more spiders and small bugs. This means that purple loosestrife might create conditions that cause some things to die out, but other things to be more abundant than they were before.

Examples of the other conclusions used in the task are presented in the Results.

After reading each conclusion card, the participants were asked, “How valid is this conclusion, given Mr. Tesser’s evidence?” The term validity was discussed with the participants as a rating of whether the conclusion could be justified by Mr. Tesser’s evidence, and then participants were asked to explain in their own words what the term meant before they began the rating task. They then rated each conclusion’s validity on a scale of 1 (low validity) to 5 (high validity), explained their rating, and listed the strengths and weaknesses of the conclusion. The interviewer filled out a separate “Mr. Tesser’s Scoring Sheet” for each conclusion as the participant spoke, by circling the numeric rating the participant gave the card, and then jotting down the strengths and weaknesses the participant mentioned. Participants’ responses were also tape recorded and later transcribed. In addition, the scientists were asked at the end of their interview to summarize the criteria they used to judge the conclusions, and to reflect on why and how these came to be their personal criteria for judging conclusions.

Analyses
Analyses focused primarily on how participants explained and justified their ratings of each conclusion, rather than on the actual numeric value they assigned to the conclusions. Coding of participants’ explanations for why they rated each conclusion as having high, medium, or low validity occurred in two phases: the first qualitative and the second quantitative. For the qualitative phase, we first extracted from the transcripts a list of all of the strengths and weaknesses each participant mentioned for each conclusion. This list had a total of 906 strengths and weaknesses across the 10 conclusions rated by the 45 participants.

A coding scheme for categorizing these strengths and weaknesses by type was developed inductively from close examination of the nature of the responses, and refined through practice applications of early versions of the scheme to the data until it could fully describe the range of participants’ responses. Each strength and weakness was then coded into categories using the final coding scheme (Table 1). While categorizing the strengths and weaknesses we noticed that a prominent feature of some responses was that they included personal inferences or views as a basis for judging the conclusions. Therefore, we added a PI or a PV code next to each strength or weakness that included a personal inference or a personal view. For instance, the statement “It
does do damage to other living things because the orchid died” was coded as S2/PI because the person based a judgment of the strength of a conclusion on the fact that the conclusion agreed with his or her own inference that purple loosestrife caused the orchid to die. The statement “It’s not good that there’s a lot of that plant in the water and they think it’s good” was coded as WO/PV because it pointed out a weakness in the conclusion based on the rater’s own view about whether the plant is good. To ensure coding reliability, the two authors each coded the entire data set separately, with one coder blind to the identities of the participants, achieving 94% interrater reliability. We then met to finalize codes for the 53 items (6% of the data) that we had scored differently.

For the quantitative phase of analysis, we made a holistic summary rating of each participant’s evaluation of each conclusion on a scale of 0–4 (Table 2), according to whether they noted the major strengths and weaknesses we had embedded in the conclusion cards and whether they based their judgments of strengths and weaknesses on personal inferences and views or on evidence. The coding scheme thus prioritized a basic tenet of scientific reasoning (Giere, 1984) that is socially sanctioned by the scientific community (Longino, 1990) as well as reflected in science education standards (NRC, 1996): that explanations and conclusions should be consistent with experimental and observational evidence, not based merely on one’s prior views.2

In total, each participant received 10 numeric scores—one for each of the 10 conclusions they judged. The two authors scored these independently, reaching 86% agreement, and then discussed initial differences in ratings to reach consensus on final scores. Finally, group means

---

**Table 1**

*Categories of conclusion strengths and weaknesses noted by participants*

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1—Conclusion takes a range of evidence into account</td>
<td>W1—Conclusion ignores some important observations and evidence</td>
</tr>
<tr>
<td>Example: “She used almost everything that he had there . . . more bugs, less water, very good.”</td>
<td>Example: “The student didn’t consider all the information, like the spiders, hopping bugs, and the water.”</td>
</tr>
<tr>
<td>S2—Conclusion makes appropriately qualified assertions</td>
<td>W2—Conclusion makes unsupported claims or inferences</td>
</tr>
<tr>
<td>Example: “He says ‘might’ create conditions, so he’s not saying it’s true, he’s saying this might be true because of these observations.”</td>
<td>Example: “It has nothing to do with the observations.”</td>
</tr>
<tr>
<td>S3—Conclusion points out that certain inferences cannot be drawn</td>
<td>W3—Conclusion uses vague language</td>
</tr>
<tr>
<td>Example: “They’re right, you can’t necessarily assume that the purple loosestrife caused it to die.”</td>
<td>Example: “They’re being very unclear and not specific.”</td>
</tr>
<tr>
<td>SO—Other strengths</td>
<td>W4—Conclusion just reiterates observations</td>
</tr>
<tr>
<td>Example: “They think different about the situation.”</td>
<td>Example: “The person just exactly repeated what he said.”</td>
</tr>
<tr>
<td>WO—Other weaknesses</td>
<td></td>
</tr>
<tr>
<td>Example: “They didn’t say how or why the bugs can live off the plant.”</td>
<td></td>
</tr>
</tbody>
</table>
Results

This section first presents statistical analyses of how the scientists, technicians, nonscientist adults, and students differed in their performance on the conclusions rating interview task. We then qualitatively explore the most interesting and educationally relevant aspect of these results—differences in the epistemological criteria that scientists and students used to rate the conclusions. We present excerpts from interviews with scientists and students to illustrate contrasts in how they dealt with three types of criteria for judging conclusions: coherence with personal inferences from the data; coherence with prior knowledge, beliefs, or values; and specificity of the conclusions. We then present the scientists’ own summations of the criteria they used to judge the conclusions, which included empirical support and consistency, logical consistency, and scope of the conclusions. Finally, we present the scientists’ reflections that attribute the source of their epistemological criteria to their socio-intellectual interactions within the scientific community.

Group Differences in Task Performance

The pattern of task performance based on the average scores per group was that scientists performed highest when rating the validity of conclusions, followed by technicians, nonscientist adults, and students (Table 3). In other words, adults with more formal scientific training, experience, and knowledge outscores adults with nonscience backgrounds, and adults received higher scores than students regardless of their background. There was a significant main effect for these differences, $F(3, 41) = 15.95, p < .001$, and a large effect size ($\eta^2 = .54$). However, in post-hoc Tukey tests, differences between the students’ and nonscientists’ scores were not significant ($p = .57$) and differences between the technicians’ and scientists’ scores were not significant ($p = .98$), but differences between the scores of students and technicians ($p < .001$), students and scientists ($p < .001$), nonscientists and technicians ($p = .04$), and nonscientists and scientists ($p = .03$) were significant. Thus, students and nonscientists were a relatively homogeneous group that performed lower than scientists and technicians as a homogeneous group.

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Does not mention any relevant strengths and weaknesses of the conclusion.</td>
</tr>
<tr>
<td>1</td>
<td>Mentions some relevant strengths and weaknesses of the conclusion, but not the major ones. Also uses agreement with personal inferences or views as a basis for judging the conclusion.</td>
</tr>
<tr>
<td>2</td>
<td>Mentions some strengths and weaknesses of the conclusion, but not the major ones. Does not base judgments on agreement with personal inferences or views.</td>
</tr>
<tr>
<td>3</td>
<td>Mentions the major strengths and weaknesses of the conclusion, but also uses agreement with personal inferences or views as a basis for judging the conclusion.</td>
</tr>
<tr>
<td>4</td>
<td>Mentions the major strengths and weaknesses of the conclusion. Does not base judgments on agreement with personal inferences or views.</td>
</tr>
</tbody>
</table>

(for the scores achieved by students, nonscientists, technicians, and scientists) were compared using one-way analysis of variance.)
Within the student group, students’ performance did not differ according to their prior knowledge of ecology, $F(4, 23) = 1.27$, $p = .32$, or their science achievement level, $F(4, 23) = .56$, $p = .70$. This means that higher achieving and more ecologically knowledgeable science students did not demonstrate superior reasoning about conclusions.

There were two ways in which responses to the conclusion-rating task received low scores (see the coding scheme, Table 2). The first was when the responses did not identify any major or minor strengths or weaknesses of a conclusion (thus receiving a score of 0) or mentioned only minor rather than major strengths and weaknesses (thus receiving scores of 1 or 2). The second was when the responses included judgments of the conclusions that were based in part on personal inferences and values, as opposed to using strictly evidence-based criteria for judging the conclusions. A sizable percentage of students’ (31%) and nonscientists’ (27%) responses, but a low percentage of technicians’ (5%) and scientists’ (1%) responses, included personal inferences and values as a criterion for judging the validity of the conclusions. Thus, given the group differences in performance shown in Table 3, it follows that more of students’ and nonscientists’ responses than technicians’ and scientists’ responses failed to identify major strengths and weaknesses in the conclusions, and/or contained personal inferences. The next section begins where the statistical analyses leave off by presenting and comparing actual samples of students’ and scientists’ responses.

Contrasting Students’ and Scientists’ Epistemic Criteria for Rating Conclusions

The following sections present contrasts in the responses of the two groups whose average scores differed the most from one another, and whose performances have the most potential to inform science education: students and scientists. Although some students’ evaluations of some of the conclusions were similar to the scientists’ evaluations—for instance, 38% of the students’ 231 conclusion ratings were coded as Level 4 and did not differ substantively from the scientists’ responses that were coded as Level 4—the purpose of highlighting contrasts in students’ and scientists’ reasoning is to examine closely the ways in which the two groups’ reasoning can differ. This section thus explores qualitatively what accounted for the statistical differences in the two groups’ performances. These contrasts will provide a basis for reflections in the discussion on the origins of the epistemic criteria that underlie differences in students’ and scientists’ reasoning.

The Criterion of Coherence with Personal Inferences from the Data. One difference between students’ and scientists’ reasoning emerged when students judged the conclusions based on the degree to which they were consistent with their own inferences from Mr. Tesser’s
data. These students formed their own conclusions about causal connections after reading Mr. Tesser’s evidence, and then rated the given conclusions high if they agreed with their own conclusions, and low if they did not. In contrast, scientists were careful to note that it was impossible to assert any causal connections based on Mr. Tesser’s observations alone.

This difference in criteria for rating the conclusions emerged, for instance, in judgments of Conclusion 2, which read: “Orchids die all the time anyway—that is why they are endangered. You can’t assume that the purple loosestrife caused it to die.” This conclusion recognized that it is unjustified to assume a causal connection between orchid death and purple loosestrife growth, yet it also contained an unsupported tautological statement about orchids being endangered because they “die all the time.” Interestingly, students such as the one quoted below did not criticize the tautology, but did fault the conclusion for failing to assert an obvious causal connection between purple loosestrife growth and orchid death:

Well, let me see, well, it’s kind of valid, I’d give it a 3. Because that’s true that, I don’t know about orchids or anything, they do die and stuff, but as soon as that stuff (purple loosestrife) came there, it (the orchid) was wilted and brown and stuff, and he could have said that they do die often, but it (purple loosestrife) seemed like the cause if it, the way he (Mr. Tesser) wrote it, it was because of the purple loosestrife. It’s like, “you can’t assume,” and it’s right in there and you can read it; basically, he was basing it on the purple loosestrife is killing it.

Thus, this student believed that Mr. Tesser’s evidence pointed so clearly to the fact that purple loosestrife was killing the orchids that Conclusion 2 was weak for not recognizing this causal linkage. A scientist, on the other hand, praised the conclusion for pointing out that it is impossible to draw a causal claim from the evidence, and also commented on the weakness in the tautological argument about why orchids are endangered:

I think that you can’t assume the purple loosestrife caused it to die is correct, although I don’t know if it’s “very valid,” but it’s almost a 4 on that scale. You might infer that purple loosestrife had something to do with the death of the orchid, but based on that alone you can’t conclude that it’s sort of outgrowing it or something. Orchids die all the time anyway; well, that’s true, but it’s like saying smokers die all the time anyway, too. So that bit of logic doesn’t help them. I wouldn’t frame the argument that way. And because things die all the time, that’s not why things are endangered, that’s false. Everything dies all the time. The second half (of the conclusion) seemed okay, but the first half didn’t.

Whereas the scientist valued the conclusion for failing to make a causal claim, the student faulted it for that reason. The judgment of another scientist fell in between these two examples. This scientist agreed that reluctance to claim a causal connection was a strength of the conclusion, but went on to fault the conclusion for not at least acknowledging that the evidence did suggest that there could be a causal link between loosestrife growth and orchid death:

. . . . There is certainly some suggestion that purple loosestrife, just from these observations, is crowding out at least some other plants. So there is this observation, I would want to go look for more orchids. I wouldn’t want to reject the possibility that purple loosestrife could have outcompeted that orchid. Although it is possible that it died from other causes, it is possible that it outcompeted it.

Although this scientist thought that a more robust conclusion should acknowledge that competition from the loosestrife could be responsible for orchid death, she did not go so far as
the student did to claim that loosestrife was the obvious cause of the orchids’ death. Overall, then, the two scientists demonstrated more appreciation for conservative claims and a need for more evidence than did the student.

The contrast between scientists’ appreciation for not overextending claims and some students’ disdain for conclusions that failed to assert seemingly obvious causal connections also emerged in judgments of Conclusion 8, which read: “I think that orchids die out because purple loosestrife has some kind of toxic stuff in its leaves and roots that does damage to other living things. But it also has sugars and nutrients in its tissues that small bugs can live off of.” One student rated this conclusion as follows:

Okay, I will give this student a 5 ’cause the student explained that how it made the orchids die out and how the bugs live there more. How more bugs live there. Weakness is that he just said “I think,” which meant that, “I think” means that he’s not really sure. It might be a weakness. Like if you say “I think that” if you were doing an essay about some energy or something, and you say “I think” that means you’re not sure, and that means that could, like, ruin your essay.

This student appreciated that the conclusion accounted for a range of observations, including that there were fewer orchids yet more bugs, and also appreciated the explanatory nature of the conclusion. The student faulted the conclusion, however, for hedging with “I think,” which indicated that the person who drew the conclusion did not have sufficiently solid knowledge to state ideas with assuredness. This behavior, the student notes, could be detrimental if displayed within the context of a typical school task such as writing an essay.

The following scientist, on the other hand, appreciated that the conclusion was stated like a tentative hypothesis, but was uncomfortable with the lack of empirical data supporting the assertions:

That’s another conclusion that’s gone way beyond any of the observations that were made. But again, it’s a good hypothesis to explain the observations. You’d need a lot more testing to support the observations, so in terms of a valid conclusion based on the observations we have, there’s no basis for it. So that is [rated as] a 1. But it is a good place to start.

Just as in the prior example, in this case the scientist felt that the conclusion strayed too far from the evidence, whereas a student showed more appreciation for what seemed like good insights in the conclusion.

*The Criterion of Coherence with Prior Knowledge, Beliefs, or Values.* In addition to judging conclusions based on the criterion of coherence with their own interpretations of Mr. Tesser’s data, some students also judged conclusions based on the criterion of coherence with their own prior ideas about causal mechanisms. This occurred for instance, in judgments of Conclusion 3, which read: “I think purple loosestrife causes native plants to decline because it shades them out, and it just physically crowds them out, and probably sucks up all the nutrients in the soil.”

The following student rated this as a highly valid conclusion because it was congruent with his own knowledge about biology and ecology:

Oh, wow! [Rate this as a] 5. This is still based on Observation Number 1, umm, and the actual quote is [he repeats Conclusion 3 aloud]. All valid. Because the orchid, if it was
growing and it was dying with the loosestrife dying all around it? Then, based on my knowledge of biology, which is limited I’ll grant you, but still, I would say that, yes, there are less nutrients there. And physical crowding, plants have, if you look at them, have a characteristic distance they like to grow from each other and orchids don’t really like to be growing right next to other plants, they like to have a distance . . . and shading them out; yeah, I can accept that because loosestrife is a pretty tall plant, from what I’m seeing here.

If you can’t get to the, umm, if you can’t get the sunlight for photosynthesis, of course it’s going to die.

This student thought about each mechanism proposed in the conclusion, checked each against his own prior knowledge, and judged each to be plausible. He did not worry about the evidentiary backings for the claims because they made such good sense to him. In contrast, a scientist who also acknowledged the plausibility of the proposed mechanisms rated this conclusion less favorably:

I’d give this about a 3, because again the person is drawing conjecture as to why something has happened. But to the student’s credit, he or she is very much sort of qualified, you know, “I think” purple loosestrife causes this, and “It probably sucks up all the nutrients.” And there are a couple of sort of logical mechanisms, both shading and physical crowding, so he’s being specific, he or she, is being specific as to what they think is going on, and making it sound like an opinion rather than fact. So I think they’re doing a pretty good job. Again it doesn’t get a “very valid” score because again there’s no real evidence here that that’s what’s going on, but they’re being clear about, ah, setting up some, some ideas. They’re making it clear that it’s an opinion.

What the scientist appreciated most about this conclusion was that it was stated more as a set of hypotheses than as conclusive claims. This was a feature of the conclusion that the student did not even mention, and in fact, as was evident in the previous quotations, students sometimes faulted conclusions that were couched in tentative terms.

The two conclusions that evoked the largest number of personal inferences and values (together comprising 66 of a total of 181 responses in the data set coded as PI or PV) as the basis for judging the conclusions’ validity referred to topics that were completely tangential to Mr. Tesser’s evidence. Conclusion 7 focused on how purple loosestrife threatens fish coming into the marsh, although fish were not mentioned at all by Mr. Tesser: “Probably the greatest threat of purple loosestrife is to the fish that come into the marsh from the river. Fish are really important food for other animals, and for humans, too.” The student quoted below evaluated the conclusion based on the criterion of how true the statements seemed rather than on their relationship to Mr. Tesser’s evidence:

All right, I’ll give this one a 5 because it’s true; I mean, fish are really important food for other animals, like birds and us—a lot of people like to fish. So it’s, like, the purple loosestrife kind of does hinder other animals and our ability to fish. [There are] not really any weaknesses.

In the following response, a scientist also acknowledged that the conclusion contained some interesting insights, but then criticized the conclusion for not being tied to Mr. Tesser’s evidence:

Okay, this is pretty interesting, because the student is thinking of the larger system and that’s interesting. They are seeing purple loosestrife as a threat, so they are seeing it as
something that’s a hindrance and they’re putting a value on it. They probably think that things are knocking things out of the system. So it’s good that they’re thinking about how this is going to reverberate through the rest of the system. But there is no real basis for them to think that fish come into the marsh from the river. The connection with the fish is interesting, but not valid based on the information they’ve been given. So I would give this a 1.

Whereas the student gave Conclusion 7 the highest possible rating, the scientist gave it the lowest rating. Although both of them appreciated the value of the ideas within the conclusion, the scientist applied stringent criteria for the need for the conclusion to be backed by evidence. The student rated the conclusion based on a personal estimation of the truth of the ideas.

Conclusion 10 also went far beyond Mr. Tesser’s evidence to make value judgments about the “problem” of purple loosestrife: “Purple loosestrife is definitely a major problem that needs to be dealt with. It should be removed from our wetlands immediately. Maybe town workers could pull it up, since herbicides might damage the other plants, too.” Students such as the following reacted strongly to the HS’s proposal to remove purple loosestrife from the marsh and the suggestion of a technique for doing so:

Well, I’d probably give her a 4. ’Cause, yes, they’re saying that, umm, there should be something done about it, but just pulling it up may not take care of it all. ’Cause there may be some left, like a seed or something that produces, umm, down in the water, and it would just keep growing, so there’s really not much like, like, that probably wouldn’t take care of it. But she is agreeing with what he’s [Mr. Tesser’s] saying, it needs to be taken care of and there are things missing, like the cattails are missing and the other flowers, err, plants, that made it look nice. Now it’s taking over and it’s all one big color. So, and not many people, I mean it’s pretty, but it would be better to have a colorful thing. Umm, if she wants, or if they want it removed from the wetlands they should come up with an idea to get all of it instead of just pulling it up, so what she’s saying to pull it up, it’s not really you know, it’s not like it’s going to take all of it, so, what would be a better answer is to take, to create something to totally clear it out, but wouldn’t hurt any of the other plants.

This student critically evaluated the HS’s proposal to remove purple loosestrife by pulling it up, and provided a rationale for why that method might be ineffective. However, she supported the HS’s sentiment that loosestrife should be eradicated, which she felt followed from Mr. Tesser’s observations about the missing cattails and other flowers, and from her own perspectives on the relative aesthetic value of different plant species.

It is exactly the kind of opinionated reasoning demonstrated by the student quoted above that a scientist criticized in the HS’s reasoning within Conclusion 10:

[Jokingly] You’re sure this person is not from the [name of a conservation organization]? Umm, there really is no evidence here that it’s a major problem, so it, this one doesn’t fit the facts, it’s not valid to draw these conclusions or statements from these facts. They’re carrying it way too far in suggesting that immediate, fairly severe action, even though it’s not herbicides, should be applied to this plant. So the biggest problem is it doesn’t fit what we’ve been told here. They seem to have a bit of a hidden agenda. They have a bias, it seems to me. That may be my bias, because I’ve heard this sort of thing many times before. But. I’d give it a 1. No strengths—I mean, they have strong opinions and they have some plans of implementing their opinions if they want town workers instead of herbicides, but it, you can’t get here [points to the conclusion card], from here [points to the list of Mr. Tesser’s evidence]. I mean, there are lots of people who have opinions. You need some basis.
While acknowledging that it is common to have opinions and biases, the scientist thought that it is necessary to have evidence to back them up. The student, on the other hand, appreciated the opinions and ideas for action in part because she felt that the evidence pointed directly to the conclusion that purple loosestrife is bad for the marsh.

The Criterion of the Specificity of Conclusions. The main weakness of Conclusion 6 was that it lacked precision and specificity of language: “All that stuff is really bad for the plants. It makes it hard to grow. And they need everything to be just right. You never know exactly what a plant will need.” Yet, students such as the one quoted below were not fazed by this conclusion’s vague language. They seemed simply to read their own meanings into the vague statements, and then moved on to evaluating what they took the language to mean. In the following statement, the student does not mention that vague language is a weakness of Conclusion 6, but in fact uses the same vague language, talking about “it” without stating what “it” is. The student does, however, perceptively note two other weaknesses in the conclusion: its failure to account for the entire range of evidence, and its internal inconsistency:

Okay, this student I would give a 4. Because it is also ignoring the fact that it could be good, too, but he said you never know exactly what a plant will need, so umm... I would consider that a strength, because, umm, if you never know what exactly a plant will need then, then, umm, you can’t... Wait. I’m sorry, that should be a weakness, because, umm, he said you never know exactly what a plant will need, and if he said that, that means you can’t be sure that it is bad. So concluding what I just thought of now, I wouldn’t give this student a 4 now, I would give it like, umm, a 2.

In contrast, to the student’s careful scrutiny of the claims within the conclusion, scientists such as the one quoted below focused right away on the lack of specificity of the language, which they felt made it impossible to evaluate the substance of the conclusion:

Well, there’s not much here, so it’s hard to evaluate. The weakness is just vagueness, that it’s not clear what “stuff” means or what’s bad, what plants they’re referring to. So it’s not specific enough; none of this is specific enough to evaluate very well... and it’s hard to find a strength based on the observations, again, there’s just not that much specific there. In terms of validity, there’s nothing there, so I guess I’d give it a 1, ’cause there are no valid conclusions drawn from the actual data.

Whereas the student was willing to infer what the HS was talking about, the scientist was not. Therefore, the student got past the language and thought critically about what he assumed the HS was claiming. The scientist, however, thought that nothing about the conclusion could be evaluated because of its vague language. The contrast in a student’s and a scientist’s evaluation of this conclusion again points to an example of a student being more willing than a scientist to read his or her own interpretations into the conclusions as a basis for evaluating them.

Scientists’ Reflections on Their Epistemic Criteria for Judging Conclusions

At the end of each interview with the scientists, they were asked if they were aware of the criteria they used to judge the conclusions. Although the participants’ criteria were inherent in their comments about each conclusion, we thought that it would be useful to hear the scientists articulate their standards for judging conclusions. This also provided a check on the
summaries of their criteria that we formed directly from their responses during the conclusion judging task.

**Scientists’ Epistemological Criteria for Judging Conclusions.** The criteria that emerged from the scientists’ comments about how they judged the conclusions are grouped into three main categories: empirical support and consistency, logical consistency, and scope. Each of these is described in turn, and illuminated with quotations from the scientists.

**EMPIRICAL SUPPORT AND CONSISTENCY**

All of the scientists mentioned first and foremost that they looked for whether each conclusion was supported by and consistent with Mr. Tesser’s evidence and did not inappropriately extend beyond the evidence. For instance, one scientist preferred conclusions that connected observations to a logical argument that did not go beyond the observations: “Basically, I was asking for some strict interpretations of the observations that were made, and any logical conclusions that would be drawn from them, based on them directly, as opposed to extrapolating either about the effects of the purple loosestrife or the necessity of removing it.” Although the scientists recognized that conclusions should propose an argument or make a claim rather than just restating observations, they were critical of conclusions that did not take into account equally plausible alternative explanations, as is evident in the following scientist’s statement: “So what I saw the [hypothetical] students doing was occasionally wildly overstating the observations, or coming up with a halfway decent mechanism, but without much, without recognizing that that was probably one of many possibilities.”

The scientists did, however, tend to give hypothetical students a lot of credit for proposing mechanisms as long as they did not assume causal connections between two observations, but rather made the claims sound more like tentative hypotheses. One scientist said that she preferred conclusions that went out on a limb to propose possible mechanisms to account for the observations to “wishy-washy” conclusions that simply reiterated the observations.

Although the empirical consistency criterion included making sure that claims did not extend inappropriately beyond the evidence, some of the scientists stated that they apply this criterion more or less stringently depending on the stage of research in which a claim is proposed. For instance, one scientist said, “I’d encourage speculation if it was a colleague trying out a hypothesis on me, but not for a conclusion.” Likewise, another scientist implied that applying the logical criterion of strict empirical consistency could inhibit the early stages of research: “I don’t think that I have the same rules for peer review as I do for my own hypotheses; I have my own instincts about where to go, and this may be too vague of a word, but I think of it as instinct and inspiration.” These and other scientists stated that intuition is valuable when deciding which preliminary data to ignore and which to pursue for instance, but that later they hold their ideas up for scrutiny against more logical criteria. Another scientist cautioned that “You have to have a clear sense of the line of when you are setting aside logic in favor of what you know about the system.” This comment indicates that the scientist sees a need for metacognitive monitoring that employs conditional knowledge—knowing when, where, and how to apply epistemological criteria for evaluating ideas.

When asked if scientists would relax their standards for empirical support for a claim when they review another person’s argument that matches their own interpretations of the system being studied, one scientist responded this way:

...What I personally know about purple loosestrife, it seems to me that it isn’t fair to judge the students’ conclusions based on that...I think it’s fair to draw the attention, unlike a courtroom, it’s fair to draw the attention of a student or another scientist to
additional evidence. I could say to an author, “I don’t believe this conclusion because you’ve ignored the work of x and y who show this.” And that seems fair, but it doesn’t seem fair to do the other, to cut them slack because I agree with them.

This scientist thus accepts the criterion of judging someone’s conclusion based on the evidence they cite rather than on its coherence with one’s personal ideas. The scientists, unlike the students, were consistent in applying this criterion throughout the conclusion judging task. This scientist’s comment indicates that he also strives to apply this criterion in his professional practice as an editor of a scientific journal.

LOGICAL CONSISTENCY

In addition to being consistent with the evidence, scientists such as the one quoted next mentioned that an explanation or conclusion should not contain any logical contradictions:

The observations have to be consistent, I don’t know how to say this, consistent with the observations and also with themselves, within themselves. There can’t be any glaring contradiction, and I don’t know formally how to say that, but I see that in scientific papers that I evaluate, too—it’s pretty common for people to have explanations where there are internal inconsistencies.

This criterion refers to rules of logic rather than to rules of evidence. For instance, one of the conclusions that violated a rule of logic claimed that purple loosestrife has “toxic stuff” in it that damages other living things, but also has sugars and nutrients in it that provide sustenance to small bugs. The conclusion did not offer an explanation that resolved the seemingly contradictory claims, and so could be criticized as lacking logical consistency. Interestingly, the scientist quoted above pointed out that professional scientists’ explanations also often have internal inconsistencies that are identified through the peer review process. This comment is telling in that it emphasizes that meeting scientific standards for knowledge claims does not depend just on the reasoning skills of individuals, but is in part a collective enterprise.

SCOPE

Scientists also judged the conclusions based on whether they addressed the range of evidence that had been presented, as opposed to addressing only a portion of the evidence. They also appreciated conclusions that showed that the HS was thinking about possible system-level effects, rather than thinking about each observation in isolation of other information. One scientist described applying the scope of explanation criterion as follows:

And one of the other, my likes, which I guess was obvious, was the use of multiple pieces of evidence to draw a broader conclusion, which to me is more interesting or valid, a broader conclusion, than just saying “spiders like to build here,” which ignores six of the seven observations, and that may not be fair, but that’s kind of, well, no I think it is fair, you think about a scientific paper, a scientific paper is supposed to be important or interesting if it has a more general application, so some of these students clearly have drawn conclusions and some of these are more general and some are more specific.

This scientist describes the issue of the scope of conclusions in terms of the usefulness of a conclusion. The scientist values generalizability—the ability to make a claim that applies beyond the specific situation that was studied—and sees such general applicability as being related to how many different pieces of evidence were taken into consideration. Thus, this scientist’s primary concern with the scope of a conclusion is not so much with the potential for
introducing a selective bias by considering only some of the evidence, but rather with the potential for how interesting and valuable the conclusion will be for science as a whole.

The Origins of Scientists’ Epistemological Criteria. Once the scientists discussed the criteria that had guided their judgments, they were asked where they thought their criteria came from. They all mentioned that they learned to judge conclusions based on evidence through the social processes of doing science. One scientist put it this way:

We get it from the community, by seeing how people react to things, how they judge things, what the standards are. When you are around it enough, you internalize the rules. It is not like law school where there are probably strict rules laid down and you learn them. Nobody teaches us these things or stands up at a blackboard and lists them, but we all know how to do it.

When another scientist was asked if he learned criteria for forming and judging conclusions through explicit instruction in classes as a student, he responded:

No, actually, I’d have to think about that a bit more, but my first impression is that it comes from actually practicing science and trying to make conclusions to other people and they reject them based on another interpretation, then you realize that yes, I can say that and no, I can’t say that based on these data.

A third scientist who also spoke about learning criteria for making and judging knowledge claims through interacting within a community of scientists described how processes of scientific argumentation also becomes internalized in part through the process of writing:

Scientist: My guess is that, and I’m in the profession of science, so, yes, your training, but it’s your training through doing. And so, by making some of these bone-headed arguments and having them picked apart by your fellow students and professors, you start to hone your ability to develop an argument. . . . The other thing I would say that worked for me I suppose came from writing in itself. From writing scientific papers and arguments.

Interviewer: From writing and the review process coupled?

Scientist: Yeah, but especially from writing my own scientific papers. Where you make some of these same bone-headed arguments to yourself, and you say, wait a minute, that’s not right.

Interviewer: So you’re sort of internalizing your audience in a way, do you think?

Scientist: Yeah, partly internalizing and then it’s like when you think through a problem and you start thinking of plausible explanations that you criticize them yourself and then you, maybe you don’t reject them completely, but you start to dimensionalize it and realize it could be B or C, and I thought it was A at first. And a lot of that comes for me from writing about my project.

Overall, then, the process these scientists describe is one of initially learning criteria for evaluating arguments and conclusions through interaction with their peers. The criteria they referred to can be stated as: A conclusion or explanation must be plausible, consistent with the
evidence, and superior to alternative interpretations. Through proposing conclusions to other scientists who then apply these criteria to judge what has been proposed, scientists learn to apply the same criteria to their own and others’ work. Although the criteria are socially constructed and implemented, they also serve as important epistemological criteria for scientists to use as they work alone to craft their arguments, such as when writing scientific papers. Thus, criteria for judging scientific conclusions can be described as moving from the interpersonal plane to the intrapersonal plane through a process of internalization (Vygotsky, 1978). Once internalized, the epistemological criteria become central elements of scientists’ critical thinking for generating their own robust knowledge claims.

Discussion

In this study we analyzed how four different groups—students, nonscientist adults, science technicians, and scientists—evaluated a set of hypothetical conclusions drawn from a set of evidence. Judging conclusions is a key process in scientific knowledge production for both professionals and nonscientists. For professionals, judging conclusions and having one’s own conclusions judged by others is a gate-keeping process for allowing new knowledge claims into the scientific canon. For nonscientists, judging conclusions is one process by which they decide what claims to accept as a building block for expanding their own understanding of a phenomenon. In addition to being an important reasoning process in its own right, however, judging conclusions also evokes people’s epistemological criteria, or the standards they use for judging the limits and validity of knowledge claims. Given our interest in exploring the role that metaconceptual frameworks such as epistemological criteria play in scientific reasoning, we focused our analyses on this dimension of participants’ judgments of conclusions.

The scientists’ and technicians’ predominant epistemological criterion for judging conclusions can be summarized as: Conclusions should cohere with the range of evidence. Students and adult nonscientists also sometimes applied that criterion, but they also applied three alternative criteria more often than did technicians and scientists: Conclusions should be plausible based on the inferences I draw from Mr. Tesser’s evidence; Conclusions should be plausible based on their adherence with the evidence plus my own ideas about how the system works; or simply, Conclusions should be plausible relative to my own ideas and values. Thus, the major difference in the judgments of the science practitioners and the nonscientist groups was in their emphasis on empirical consistency versus plausibility of the conclusions.

Relative emphases on the epistemic criteria of plausibility and empirical consistency in science students’ metaconceptual frameworks could have important implications for their science learning, such as engaging in inquiry and undergoing conceptual change. Epistemic criteria for what constitutes sufficient backings for a knowledge claim could affect students’ engagement in science investigations because if they believe that they already have sufficient answers to scientific questions based on intuitive beliefs or experience, they will not be motivated to test their ideas empirically (Varelas, 1996). Also, students’ epistemic criteria are components of their conceptual ecology (Strike & Posner, 1992) that can influence if they change their prior ideas about a phenomenon: for instance, by determining whether they value the need to examine their personal beliefs, goals, and biases critically while evaluating new information. Thus, by presenting detailed contrasts between students’ and scientists’ epistemic criteria for judging conclusions we highlighted some underpinnings of the differences in novice and expert scientific reasoning that have an important bearing on the central aims of science education.
Interpreting the Results in Light of Standards for Scientific Rationality

Concluding that the students lacked expertise as scientific reasoners when they judged conclusions using epistemic criteria of coherence with their personal theories or inferences risks holding them to an outmoded, positivist standard of rationality that emphasizes objectivity through the application of theory-independent rules of inference. In contrast, postpositivist perspectives in the philosophy of science (e.g., Boyd et al., 1990) acknowledge the intricate interplay of theory and methodology in science. Cognitive scientists portray this interplay as a problem solving process that involves the recursive search of two problem spaces: the experiment space of data and methods, and the hypothesis space of conjectures and theories (Klahr, Fay, & Dunbar, 1993). Domain-specific information about the plausibility of hypotheses influences scientists’ search of both hypothesis and experiment spaces, whereas domain-general heuristics constrain and guide their explorations. Thus, conceptual knowledge and logical validity interact in scientific theory building (Thagard, 1994).

The task used in the present study, however, created a context of judging conclusions, not producing them. Because descriptions and interpretations of scientific work in cognitive, philosophical, historical, and sociological studies of science tend to focus on scientific discovery processes and on the dynamics that prevent or incite major paradigm shifts, it is not clear how the balance between domain-specific and domain-general knowledge play out in the everyday activity of judging others’ conclusions. There is some indication, however, that different epistemic criteria become more or less important in different stages of scientific inquiry (Twene & Chitwood, 1995). As one of the scientists we interviewed explained: “There are different phases of the research process. When deciding which preliminary data to ignore and which to pursue, you use intuition. You check them against your mental model of the system to see if they make sense. But at later stages, you use more logical criteria.”

From the perspective of timing, then, it seems that when they judged conclusions based on their theories and inferences, the students were doing the right thing at the wrong time. A driving purpose of science is to develop theoretical explanations that go beyond the data to explain how results might have occurred by proposing unseen mechanisms. Therefore, students’ theory-laden judgments of conclusions, and especially their suggestions or appreciation of mechanisms that supported the conclusions, are evidence of their ability to use scientific reasoning skills and criteria that are essential for explanatory model building. Indeed, their ideas about marsh ecology were sound, such as those regarding the mechanisms and effects of plant competition, the fragility of certain plant species, and the connections among plants, habitats, and fish survival.

However, students’ use of the criterion of plausibility to judge conclusions often superceded use of the criterion of empirical support for claims of causality. In contrast, although the scientists acknowledged when patterns in the data indicated that there might be causal links between variables, and when the conclusions proposed plausible mechanisms for the linkages, they consistently faulted the conclusions for lacking sufficient evidentiary backings to support strong claims of causation. Given that the latter criterion was less important in the overall reasoning of students as a group, they did not fully display the technical along with the conceptual sensitivities that are inherent in scientific rationality.

Linking Differences in Reasoning to Contexts of Practice

What accounts for the differences in the groups’ use of epistemological criteria for judging conclusions? The differences did not seem to be due primarily to age and associated
developmental factors because the performance of nonscientist adults was not statistically
distinguishable from the performance of students. Also, neither prior knowledge about
ecological relationships nor prior achievement level in school science made a difference in
students’ performances. One obvious factor that could account for the differences, then, is the
contextual frames that the groups brought to the task based on their life-worlds and professional
experiences.

The term conclusion evoked for the scientists and technicians a set of standards and formal
criteria for judging a summative scientific knowledge claim. The students and nonscientist
adults, however, seemed to interpret the term conclusion in a broader, everyday sense as a
reaction to a set of observations, and judged the conclusions according to whether they seemed
like reasonable things to say based not only on the evidence, but also on what they personally
knew and believed. We argue that this is not so much a matter of absolute reasoning skill as it is a
matter of knowing what to value when, which is a kind of knowledge that depends on one’s
sphere of sociointellectual practice. Explanations and conclusions need only to meet standards of
pragmatic precision in everyday life—that is, to be precise enough for the purposes of the
inquirer (Hawkins & Pea, 1987). In science, however, conclusions are developed to share with a
larger community and thus must meet normative, or socially constructed standards.

Given their daily work, the scientists were familiar with experimentation as an important
means of developing new scientific knowledge, especially for establishing causality. Experi-
mentation is not a typical way of developing knowledge in everyday life, however. We tend
instead to generalize from single experiences, or gain knowledge from reading or other means of
direct transmission (Klahr & Carver, 1995). Therefore, without referencing a mental model of
experimentation, it might not have occurred as readily to the students and nonscientists that the
causal claims in the conclusions could have been better supported.

In both school and everyday life, displaying one’s knowledge about a topic is a way to gain
respect and rewards. Thus, showing one’s knowledge about ecology and using it as a basis for
judging the validity of conclusions makes sense in response to everyday and school performance
demands. One student’s comment about the need to state what you know with confidence when
writing school essays as the basis for not liking a conclusion that hedged in making a claim
indicates that at least some students intermingled standards for forming conclusions with
standards for typical schoolwork tasks. Although scientists’ professional contexts also present
performance demands for displaying their knowledge, the context of judging conclusions is not
one in which one’s personal knowledge about natural phenomena is considered to be as relevant
as one’s knowledge about rules of inference, as was indicated in one of the scientist’s quotations
about fairness presented earlier.

In effect, then, the entire cultural institution of science is embodied within the reasoning
practices of scientists, whereas nonscientists lack the same level of exposure to that cultural
knowledge (Brewer & Samarapungavan, 1991). Thus, if students and scientists apply different
criteria when reasoning because they are members of different communities of practice,
educational strategies to enhance students’ scientific reasoning should focus on the influence
that classroom community practices can have on individual development. This is the topic we
turn to next.

Recognizing the Interplay of Epistemology, Culture, and Scientific Reasoning:
Implications for Classroom Practice

The scientists we interviewed credited their sociointellectual interactions within a scientific
community as the source of their personal epistemic criteria for judging conclusions. Thus,
learning the epistemological dimensions of scientific practice is at least in part a sociocultural process.

Students can experience sociocultural practices that are similar to those that scientists experience when classrooms are organized as knowledge-building communities whose members construct and apply standards of explanation as they work together to make sense of the phenomena they are studying (e.g., Anderson, Holland, & Palincsar, 1997; Beeth & Hewson, 1999; Bereiter, Scardamalia, Cassells, & Hewitt 1997). From the perspective of sociocultural theory that traces the origin of cognitive tools such as epistemological standards to immersion in cultural activity (Vygotsky, 1978), these pedagogical practices have more potential to foster students’ epistemological development than simply having students do activity-based science, which clearly does not help them gain insight into the epistemological foundations of scientific procedures (Schauble, Glaser, Duschl, Schulze, & John, 1995). Taking a social practices perspective on epistemological growth by recognizing how immersion in a full range of scholarly community practices can stimulate intellectual development provides an educationally promising alternative to developmental psychologists’ perspectives on stage-based constraints on epistemological and other cognitive development (Kitchener & Fischer, 1990; Kuhn, 1991), which have led educators to place unnecessary limitations on their science pedagogy (Metz, 1995).

One standard of the social practice of science that this study indicates should receive attention in science classroom communities concerns how and when to use the epistemological criterion of coherence with prior knowledge in scientific work. In this study, although the scientists acknowledged the importance of their prior conceptual frameworks in shaping their research, they placed priority on the criterion of coherence with empirical evidence when judging the validity of conclusions. Students on the other hand, more often judged conclusions based on their personal views and inferences. Therefore, it could be useful for teachers to introduce metaconceptual tools such as the idea of reciprocally searching an “experiment space” of data and a “hypothesis space” of ideas (Klahr et al., 1993) to build students’ conditional knowledge of norms for the timing and interplay of these two dimensions of scientific reasoning.

Our contention is that to build competency as scientific reasoners, students need to participate over time in explicit discussions of the norms and criteria that underlie scientific work. Viewing the practice of science as including certain cultural traditions and values that are manifest as epistemological standards is perhaps less familiar to teachers than recognizing that learning science entails mastering conceptual material and gaining procedural proficiency. We suggest that in addition to discovering what concepts students do and do not understand, or what procedures they can and cannot do as a basis for science instruction, it is also important for teachers to find out what students value as intellectual standards. Such knowledge can serve as a foundation on which science teachers can build efforts to help students use and critically assess the epistemological standards of scientific communities of practice within their classroom communities.

Summary and Conclusions

In this study, scientists, technicians, nonscientist adults, and middle school students did an identical task during interviews—they rated the validity of a set of conclusions drawn from a given body of evidence. We found that the participants without extensive science backgrounds and experience (students and nonscientist adults) and those with more extensive science backgrounds and professional experience (scientists and technicians) performed differently on
the task. Fine-grained analyses contrasting two subsets of these groups—the students and the scientists—depicted how scientists prioritized empirical consistency of evidence and conclusions, whereas students more often used their personal views and inferences as criteria for judging the plausibility of the conclusions. The scientists spoke about how they developed epistemological criteria such as empirical consistency through becoming active contributing members to their community of practice, and described how they translate these experiences into personal knowledge that they apply to evaluate their own and others’ knowledge claims. They thus internalized their sociocultural activity as cognitive tools that enabled them to act in professional contexts in ways that are rational given the objectives and norms of their discipline. We built on these insights to conclude that broadening cognitive views on the development of scientific reasoning to include sociocultural perspectives on the origin of epistemological criteria in cultural activity is a promising direction for classroom practice, and ultimately for fostering students’ growth as scientific reasoners.

Notes

1In our quest for volunteers, we obtained roughly equal numbers of student (24) and adult (21) participants. Given that we further subdivided the adult population into three groups, our plan was to seek more volunteers in each of those groups based on consistency or variability in subgroup responses, following protocol for the constant comparison method of data analysis (Strauss & Corbin, 1990), which is to sample until reaching saturation (low variability) of responses. Because there was enormous consistency in responses within each subgroup of our original adult sample, we did not think it necessary to recruit additional volunteers for this study. However, because our subsamples were small, future research with more participants is warranted to bolster or extend the findings we report.

2Later in this article, we examine this view of scientific rationality more closely by contrasting a core premise of positivist perspectives in the philosophy of science (i.e., that observations and conclusions can be entirely theory-free) with postpositivist, realist perspectives that acknowledge the roles of theories and social factors in scientific judgments, to show that the standard of prioritizing empirical support when judging conclusions is consistent with contemporary realist perspectives.

3There should have been 240 ratings—10 each for 24 students—but nine responses were lost owing to taping failures.

4Words in brackets within the quotations were not uttered by the participants, but were added to make incomplete or vague statements more clear to the reader.

References


