Methodological Concerns About AAAS’s Project 2061 Study of Science Textbooks
William G. Holliday
Science Teaching Center, University of Maryland, College Park, MD 20742

The question under consideration in this commentary is not whether textbooks help students learn good science, but whether Kesidou and Roseman’s (2002) article raises important methodological concerns.

The Training of the Rating Teams
Kesidou and Roseman (2002) in their work with the American Association for the Advancement of Science (AAAS) reported statistical summaries of textbook ratings arrived at by two-person teams of science teachers and “specialists in research on teaching and learning” -- teams assessing “nine widely used programs” (Kesidou and Roseman, 2002, p. 522). These teams rated students’ and teachers’ textbook editions after receiving seven days of training.

But, information contained in the Kesidou and Roseman (2002) article and located in AAAS’s websites on August 31, 2002 (the dated month and year of this published article) suggests that these rating teams were unknowingly biased or prejudiced against commercial publishers’ science textbooks. This apparent bias or prejudice presumably occurred during the seven days of training, prior to the rating of textbooks.

The nature of the training sessions used to prepare the teams was addressed in the method section but only outlined in two long sentences printed on page 526, as follows:

“Before the analysis, reviewers in each topic area received 7 days’ training that had the following components: (a) clarification of the set of ideas that served as the basis of the analysis; (b) modeling the application of criteria, indicators, and scoring schemes using a range of examples from different curriculum materials; (c) practice using the Project 2061 curriculum analysis procedure with feedback on reviewers’ understanding of the criteria, indicators, and scoring schemes both from peers and Project 2061 staff. A notebook with clarifications, scoring schemes, and a range of examples rated from poor to excellent on how well they addressed each criterion was available to reviewers during their training and during the subsequent analyses.”

Readers need a much more detailed description of this training because of its influence on the way the teams of teachers and specialists may have assessed the textbooks.

In this regard, Roseman, Kesidou, and Stern (n.d.) in a similar study describe how they worked with reviewers looking at different science materials, resulting in, perhaps, unknowingly prejudicing the reviewers to engage in “more rigorous analysis,” as illustrated from this excerpt:
“Results of using the (training) procedure are often surprising. Analysis teams find that quick judgments about alignment to benchmarks or content standards are frequently contradicted by a more rigorous analysis. This held for single units or across several units in a program. A superficial examination most often overestimates what a material can be expected to accomplish. For example, when educators were asked how well River Cutters (a grade 6-9 curriculum module) addressed benchmarks, their initial judgments were far more optimistic than their judgments after completing the 2061 analysis. After initially listing 22 benchmarks that a cursory read led them to suspect were addressed in River Cutters, they found actual sightings for only 12 of them. After studying the meaning of the benchmarks carefully and revisiting the sightings with this more sophisticated understanding, they found that only 6 had a respectable content match. And on considering the instructional strategy of the material, only 1 was found to be instructionally well-supported, as shown below” (Roseman, Kesidou, & Stern, n.d., http://www.project2061.org/newsinfo/research/roseman/roseman2.htm).

My fear is Kesidou and Roseman in their study unknowingly biased the teams until these reviewers were more inclined to judge rigorously (i.e., negatively) the textbook-based programs assessed in the present study. Additional evidence supporting this fear of biasing the teams, especially the teacher-half, is that teachers generally are positive in their ratings of textbook-based programs (Weiss, 2001; National Research Council, 1990).

Weiss (2001) noticed this parallel contradiction in her highly publicized research findings. She concluded, “It is interesting to note that while national experts in science and mathematics education are often critical of textbook quality (American Association for the Advancement of Science, 1999; 2000), most teachers consider their textbooks to be of relatively high quality” (Weiss, 2001, p. 86).

Kesidou and Roseman’s article contradicts the judgment of practicing middle school science teachers by suggesting that a narrow sampling of information from commercial textbooks rated good, very good or excellent by these teachers (Weiss, 2001) are, in fact, poor or unsatisfactory. Science teachers generally like their textbooks (Weiss, 2001), ancillary materials and publishers’ websites, both of which increasingly play a major role in much of their teaching but were not evaluated in this study. Just because teachers like publishers’ products does not validate any products’ instructional effectiveness. But, making such a contrary negative suggestion to so many classroom practitioners about their textbook requires a clear and unambiguous argument.

Thus, the onus is on Kesidou and Roseman to convince us that the reviewing teams did not receive biased training resulting in a tendency to unjustifiably downgrade the quality of the examined textbooks.

Readers also need other information about these teams. Perhaps the experienced teachers making up half the teams did not represent typical science teachers. Perhaps the specialists dominated the rating discussions surrounding of the textbooks. This point begs the question: How did these specialists on the teams rate the programs compared to the
teachers on the same teams? How were the research-specialists’ part of the two-person teams kept from potentially dominating the rating process? Kesidou and Roseman did not describe how the scores differed between these two professional groups or whether they fairly rated materials independently of each other within each team. Did the specialists have the opportunity to persuade unknowingly the teacher to vote, rate or award a negative score on each criterion?

Concerns about the Methods Used to Produce the Commentary

In the major part of Kesidou and Roseman’s results section, the authors’ added commentary condemning popular textbooks by citing selected research studies tied to isolated statements located in the examined textbooks, obviously focusing their attention on textbook “weaknesses.” Most of the 11 pages making up the article’s results section did not describe the statistical results of the empirical study described in the article’s method section. The authors, instead, described the verbal results of a related study not described in the method section of their article. This is unusual because readers expect an article’s results section to present the results of the research study detailed in the method section, not a major presentation of results derived from a related study not described in the article’s method section in the examined textbooks, obviously focusing their attention on textbook “weaknesses.”

Kesidou and Roseman’s article described mostly weaknesses (very seldom strengths) of textbook programs using selected quotes from textbooks linked to selected research studies. Kesidou and Roseman did state in the method section that “reports” produced by the teams were used to identify “patterns across all three topic areas [three sampled science concepts], illustrating them with findings mainly from the physical science topic area” (p. 527). But, they never described how the reports were solicited or produced by the teams, or used by the authors or others to produce the “pattern” information. The origin and development of these comments is important but remains undefined. These pages of commentary are also questionable in terms of their contextual relevance and importance. Perhaps Kesidou and Roseman or their colleagues constructed statements describing these alleged textbook weaknesses. The reader cannot tell.

Quality of Science Content Sampling Procedure

Kesidou and Roseman trained teams of reviewers to use22 criteria to evaluate the quality of student textbooks, teacher’s guides and other materials in terms of “…key scientific ideas specified in national science standards, and to identify typical strengths and weaknesses of these programs using research-based criteria” (Kesidou, & Roseman, 2002, page 522). In reality, only three science ideas out of many possible choices were selected for evaluation by the authors. The chosen concepts were, “(a) the kinetic molecular theory (in physical science), (b) flow of matter and energy in ecosystems (in life science), and (c) processes that shape the earth (in earth science)” (p. 525). A convincing argument for selecting the three concepts was never made, beyond the authors’ professional judgments, and the fact that these concepts were among the many concepts included in AAAS’s (1993) Benchmarks and NRC (1996) Standards documents.
The Quality of the 22 Criteria

The 22 criteria were clearly presented in the article’s appendix. Each criterion was reportedly applied using five indicators, thus requiring each team to understand and apply 110 (22 criteria times 5 indicators) indicators and their associated criteria. Unfortunately, only one set of five indicators used to judge each of the 22 criteria in the Kesidou and Roseman study was printed in the article. More importantly, these indicators were not made available to readers nor were references given to published papers or websites. Whether the 5 indicators truly measured 5 different factors is questionable because the researchers did not report the results of a factorial analysis.

Kesidou and Roseman provided no reasonable rationale why they chose the 22 criteria rather than an alternative set of equally valid criteria established by another group of science educators. Just because a set of science materials fails to meet one set of constructed criteria drawn from selected research doesn’t suggest that these science materials are mediocre, much less unsatisfactory. Readers need to examine the 22 criteria closely, perhaps admiring many and questioning others, as I did.

Textbook publishers rightly cannot just depend on one such set of research-based criteria. They must also apply other research-based criteria and common-sense criteria -- factors and questions that scientific research has not yet or cannot answer. Publishers have long engaged potential buyers in focus group sessions and drawn from their professional experience as publishers when making decisions about how to construct a science textbook that should come closest to meeting the demands of experienced science teachers likely to serve on state and local textbook adoption committees.

Kesidou and Roseman may have used AAAS’s earlier attempts to link research to their development of the 22 criteria, according to another study highly similar to this present study, as suggested by the following: “The 2061 procedure uses highly specific analysis criteria. For example, materials are examined for how well they alert teachers to commonly held student ideas (both troublesome and helpful) such as those described in Benchmarks Chapter 15: The Research Base and then for how well the material explicitly addresses those commonly held student ideas.”

This research chapter, presented in AAAS’s Benchmarks document, has often provided AAAS studying science textbooks with a research-base foundation. The Research Chapter also has scholarship-based problems, as highlighted by simply comparing the Benchmark section “Historical Perspectives” (AAAS, 1993, p. 354) with another well-known book chapter summarizing “Teaching and Learning History” (Downey & Levstik, 1991, p. 400) in social studies education.

AAAS’s Foundational studies

This article, in one sense, is part of a long-running assault by AAAS on the quality of commercial publishers’ science and mathematics textbooks, as illustrated on AAAS’s websites. Science teachers and researchers have read about AAAS’ downbeat studies of science textbooks in a front-page article, published in NSTA Report! (1999, December), an article in NSTA’s The Science Teacher (Stern, & Roseman, 2001), an article published
AAAS reportedly has shown interests in texts (Hilts, 1989) for many years. Yet their general efforts have not always been well received. During this period, an editorial (“Still No Beef”) in The Washington Post, characterized AAAS’ Project 2061 evolving work as adding, not subtracting, to the problems of science and math teaching in America, (Editorial, 1989). The message from AAAS during the past five to eight years regarding the quality of commercial publications has been clear and consistent: these science and math textbook-based programs, especially ones not sponsored by the funding unit of the National Science Foundation, are unsatisfactory with rare exceptions. Yet, these same textbook programs are popular with practicing science teachers and the school systems which purchase them. Of course, just because teachers like textbooks and schools purchase them doesn’t mean that they are in the best interest of students.

Concerns about AAAS’ Students’ Assessment Study
Stern and Ahlgren (2002) from AAAS just (November 2002) published a JRST article, “Analysis of students’ assessments in middle school curriculum materials: Aiming precisely at Benchmarks and Standards.” The authors describe a study assessing the same nine middle school science textbooks using presumably the same procedure but, of course, different criteria. The published results suggest that the nine textbooks’ assessment components were generally rated by educators as “poor,” compared to “fair,” “satisfactory,” and “very good” (see Stern and Ahlgren, 2002, Table 1, page 892).

But, what procedure was used to rate these textbooks? The authors describe their procedure in even more general terms compared to Kesidou and Roseman’s (2002) description, followed by the statement: “[for more details about the procedure and the specific criteria, see Roseman et al. (1996), Kulm (1999), Kesidou and Roseman (2002), Resources for Science Literacy: Curriculum Materials Evaluation (in preparation), and the Project 2061 web site at www.project2061.org]” (Stern and Ahlgren, 2002, p. 891). Examination of these five cited references in Stern and Ahlgren’s article is warranted because of the high similarity between their study and the central study authored by Kesidou and Roseman.

First, the Roseman et al. (1996) referenced document was “distributed at the National Research Council colloquium…in Washington, D.C.” (Stern and Ahlgren, 2002, p. 910) and is an unavailable (practically speaking) to readers. So, this reference is of little help. Second, the Kulm article is a three-page piece published in a pamphlet produced by the Council for Basic Education (an excellent organization but not a research association). Kuhm’s description of the procedure states, “After judging the textbooks…analysts trained in Project 2061’s Procedure…determined how well the texts addressed instructional criteria (Kuhm, 1999, p. 7). Third, I have discussed the lack of specifics located in the Kesidou and Roseman (2002) article, so this document was not helpful to
me. Fourth, the Resources document is still “in preparation,” thus unavailable. Fifth, the
web site reference in the Stern and Ahlgren article lacks reasonable details about the
methodology. The best statement referencing the procedure in my web site search
(November 17, 2002) was at the end of the document, Heavy Books Light on Learning:
Not One Middle Grades Science Text Rated Satisfactory By AAAS's Project 2061
(http://www.project2061.org/newsinfo/press/rl092899.htm), which read: “A summary of
the middle grades science textbook evaluation will be posted on the Project 2061 web site
at http://www.project2061.org. Full reports on each textbook will be available early next
year.”

In Conclusion
All of the authors’ pages of commentaries criticizing textbooks, moreover, need to be
reassessed in terms of textbook publishers’ goals, state standards, contextual
considerations, other research, and common sense. The real goal of the Kesidou &
Roseman seemingly is to argue a case why popular science programs, preferred by
practicing science teachers, are just awful for kids. In addition, the article’s results read
like two separate studies, an empirical study described in the article’s methodology
section where raters assess textbooks based on 22 criteria, and a related study resembling
a downbeat book review of science textbooks linking selected quoted textbook sentences
to Kesidou and Roseman or AAAS’s view of what is important teaching and learning
research.

References
American Association for the Advancement of Science (AAAS). (1993). 
American Association for the Advancement of Science (AAAS). (1999, Fall).
Heavy texts: Light on Learning. 2061 Today, 9, 1, 3.
American Association for the Advancement of Science (AAAS). (2001,
Spring/Summer). Collaborating to create better texts: Project 2061 host conference on
science textbooks. 2061 Today, 11, 3, 5.
Shaver (Ed.), Handbook of research on social studies teaching and learning: A project of
the National council for the social Studies. (pp. 400-410). New York: Macmillan
Publishing Company.
Hilts, P. J. (1989, February 24). Science-study reform sought; Aim is to broaden
all students’ skills. The Washington Post, A1, A6
programs measure up? Findings from Project 2061’s curriculum review. Journal of
Research in Science Teaching, 39, 522-549.
43(9), 6-8.
National Research Council. (1990). Fulfilling the promise: Biology education in


