

Faculty Grading of Quantitative Problems: A Mismatch between Values and Practice

Heather L. Petcovic · Herb Fynewever ·
Charles Henderson · Jacinta M. Mutambuki ·
Jeffrey A. Barney

© Springer Science+Business Media B.V. 2012

Abstract Grading practices can send a powerful message to students about course expectations. A study by Henderson et al. (*American Journal of Physics* 72:164–169, 2004) in physics education has identified a misalignment between what college instructors say they value and their actual scoring of quantitative student solutions. This work identified three values that guide grading decisions: (1) a desire to see students' reasoning, (2) a readiness to deduct points from solutions with obvious errors and a reluctance to deduct points from solutions that might be correct, and (3) a tendency to assume correct reasoning when solutions are ambiguous. These authors propose that when values are in conflict, the conflict is resolved by placing the burden of proof on either the instructor or the student. Here, we extend the results of the physics study to earth science ($n=7$) and chemistry ($n=10$) instructors in a think-aloud interview study. Our results suggest that both the previously

H. L. Petcovic (✉)

Department of Geosciences and the Mallinson Institute for Science Education, Western Michigan University, 1187 Rood Hall, Kalamazoo, MI 49008, USA
e-mail: heather.petcovic@wmich.edu

H. Fynewever

Department of Chemistry and Biochemistry, Calvin College, 1726 Knollcrest Circle, SE, Grand Rapids, MI 49546, USA
e-mail: herb.fynewever@calvin.edu

C. Henderson

Department of Physics and the Mallinson Institute for Science Education, Western Michigan University, 1120 Everett Tower, Kalamazoo, MI 49008, USA
e-mail: charles.henderson@wmich.edu

J. M. Mutambuki · J. A. Barney

The Mallinson Institute for Science Education, Western Michigan University, 3225 Wood Hall, Kalamazoo, MI 49008, USA

J. M. Mutambuki

e-mail: jacinta.m.mutambuki@wmich.edu

J. A. Barney

e-mail: jeffrey.a.barney@wmich.edu

identified three values and the misalignment between values and grading practices exist among science faculty more generally. Furthermore, we identified a fourth value not previously recognized. Although all of the faculty across both studies stated that they valued seeing student reasoning, the combined effect suggests that only 49% of faculty across the three disciplines graded work in such a way that would actually encourage students to show their reasoning, and 34% of instructors could be viewed as penalizing students for showing their work. This research may contribute toward a better alignment between values and practice in faculty development.

Keywords Grading · Quantitative problem solving · Faculty development · Physics · Chemistry · Earth science

Introduction

Grading practices, as a form of assessment, can have a tremendous impact on what students do in a college course. Feedback from the instructor to the student, typically in the form of a grade with or without explanatory comments, is known to have a powerful effect on student learning (e.g., Black and Wiliam 1998; Elby 1999; Schoenfeld 1988). Research in physics education has documented a tension between what instructors say they value in grading quantitative, free-response student problem solutions, and their actual grading practices (Elby 1999; Henderson et al. 2004). For example, many instructors say they want to see reasoning in a student solution to make sure that the student really understands, but then assign a grade or numerical score (often without explanatory comments) in a way that penalizes students for showing their reasoning, or rewards the omission of clear reasoning (Henderson et al. 2004).

Henderson et al. (2004) propose that this tension exists because hidden internal values conflict with expressed values. Based on analysis of interviews with physics faculty, these authors identified three values that guide decisions for assigning a numerical score to student work: (1) a desire to see student reasoning to tell if a student really understands, (2) a reluctance to deduct points from solutions that might be correct and a readiness to deduct points from solutions that are clearly incorrect, and (3) a tendency to project correct thought processes on ambiguous student solutions. These values can conflict when assigning a numerical grade to a given student solution (Fig. 1). For example, an instructor may assign a high grade without conflict to a solution that shows clear and correct reasoning, and has a correct answer. Conversely, if a student shows little to no reasoning in his/her answer and has an incorrect solution, the instructor may experience no conflict between values and assign a low grade. A conflict might arise, however, if a solution shows little reasoning but has a correct answer. In this case, Value 1 suggests a low grade but Values 2 and 3 suggest a high grade. Similarly, if a solution clearly demonstrates reasoning that is incorrect, a conflict arises between Value 1, which suggests a high grade, and Value 2, which suggests a low grade.

These authors develop the “burden of proof” construct to explain how faculty resolved these conflicts (Henderson et al. 2004, p. 167). A burden of proof on the instructor means that the instructor must have explicit evidence that the student applied knowledge or procedures incorrectly in order to deduct points. A burden of proof on the student means that the instructor must have explicit evidence that the student applied knowledge or procedures correctly in order to earn points. The majority of physics faculty in the study placed the burden of proof on themselves, which may result in students’ perceiving that they are penalized for showing their reasoning. This, in turn, can create an undesirable

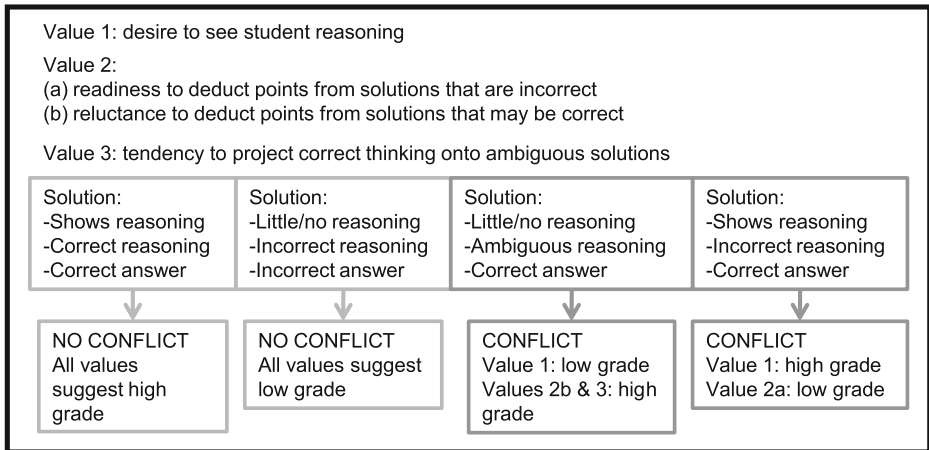


Fig. 1 Potential conflicts among grading values identified by Henderson et al. (2004) for physics faculty

situation that may discourage students from showing their reasoning (Table 1). The authors suggest that the burden of proof construct is a useful tool to create cognitive conflict that can better align faculty grading practices with expressed values.

Background

Quantitative Problem Solving in Introductory Science Courses

Most science courses require students to solve problems on homework, quizzes, and exams. These problems may involve conceptual analysis, numerical manipulation, or both. Here, we adopt the following definition of problem solving:

A cognitive process that involves (1) forming an initial representation of the problem, that is, an external presentation of the problem is encoded into an internal

Table 1 How an instructor’s view of grading might affect student behavior

	Instructor’s conflict resolution in scoring student solutions	
Conflict resolution	In favor of Value 1	In favor of Values 2 and/or 3
Instructor View of Grading	Burden of Proof on Student - Explicit evidence of correct knowledge & procedures needed to earn points	Burden of Proof on Instructor - Explicit evidence of incorrect knowledge & procedures needed to deduct points
Resulting Grading Practice	A solution will only receive points for the portions of a solution that are correct and are explained by the correct reasoning	For a solution that is not completely correct, as more reasoning is shown there is more opportunity to deduct points
Message Sent to Students	Students cannot get points without showing their reasoning	With the same level of understanding, students may receive more points for not showing reasoning
Effect on Student Behavior	This encourages students to show their reasoning	This discourages students from showing their reasoning

representation, (2) planning potential sequences of actions such as strategies and procedures to solve the problem, and (3) executing the plan and checking the results (Qin et al. 1995, p. 130).

Furthermore, we define quantitative problem solving as involving numerical manipulation in order to achieve a numerical solution to a problem. In this study we will focus only on free-response problems, by which we mean that the student writes out the solution as opposed to selecting from among several constructed responses.

Quantitative problem solving is a central feature of many college level physics and chemistry courses. Thus, significant research within these disciplines has identified student deficiencies in quantitative problem solving and has investigated instructional strategies that address these deficiencies (e.g., Gabel and Bunce 1994; Maloney 1994). However in earth science, quantitative problem solving has received significantly less attention in the education-research literature, as it is less common for instructors to require numerical analysis in introductory earth science courses (Manduca et al. 2008). Problem solving in introductory earth science is generally taken to mean conceptual reasoning about aspects of the earth system (e.g., Ault 1994). However, when quantitative problem solving is used in earth science it is commonly in the form of physics or chemistry problems applied to earth systems (Manduca et al. 2008). Thus, we expect the findings from these other disciplines to be relevant to problem solving in the earth sciences as well.

From the body of research on problem solving, several conclusions can be established. First, students leave introductory physics and chemistry courses with poor problem solving skills (e.g., Chi et al. 1981; Cooper et al. 2008; Gabel and Bunce 1994; Larkin 1980; Leonard et al. 1996; Maloney 1994; Reif and Scott 1999; Van Heuvelen 1991). Students often take an algorithmic view of problem solving, and view it as an activity in which the goal is to find the appropriate formula for a given problem. In contrast, experts reason about problems from fundamental principles, and then only apply the mathematical solution once they have a conceptual understanding of the problem (e.g., Reif and Scott 1999; Simon and Simon 1978; Stains and Talanquer 2008).

Second, research on physics problem solving suggests that students benefit from a requirement to make their reasoning explicit when solving a problem (e.g., Larkin 1980; Leonard et al. 1996; Maloney 1994; Mestre et al. 1993; Reif and Scott 1999). One benefit of requiring students to show their reasoning in problem solutions is that it forces the students to think about their reasoning. Being forced to explain their reasoning helps students to reflect on their reasoning process and develop more appropriate reasoning processes. This also helps students develop a better understanding of the concepts that are used in the reasoning process. Requiring students to explain their reasoning on problem solutions is an important part of many instructional strategies that have been shown to be successful in improving both student problem solving skills as well as their understanding of physics concepts (e.g., Heller and Hollabaugh 1992; Leonard et al. 1996; Mestre et al. 1993; Van Heuvelen 1991). Additionally, the instructor can see what difficulties students are having with the material and provide appropriate formative feedback (Black and Wiliam 1998).

Grading Practices Shape Student Behavior

Assessment practices play a critical role in determining what and how students learn (Black and Wiliam 1998; Brown et al. 1997; Ramsden 2003). Grading, perhaps the most common form of assessment, serves multiple although sometimes competing purposes such as:

evaluating the quality of student work, facilitating communication between faculty and student about the quality of work, providing the motivational framework that affects how students learn in the course, focusing organizational efforts by the instructor and student, and yielding information that both students and faculty can use to reflect upon and improve learning and teaching (Walvoord and Anderson 2010). In particular, grading practices influence how and what students study, as well as how much time and effort they spend on a class (Walvoord and Anderson 2010). Grading practices define what students regard as important, how they study, and how they spend their time in a course (Brown et al. 1997), often superseding the explicit curriculum (Ramsden 2003).

Unfortunately, many students believe that grading practices used by faculty reward rote learning strategies over deep understanding (Elby 1999; Entwistle and Entwistle 1992; Ramsden 2003). For example, Elby (1999) demonstrated that grading practices can have a profound effect on how students learn. He asked students to compare their own study habits in an introductory physics course to those study habits that they would recommend to a friend who wanted to obtain a deep understanding of physics and did not care about getting a good grade. Students recommended deep learning strategies for their friend (thus indicating that they were aware of these strategies), but conversely indicated that they actually used more rote learning strategies since these were viewed as leading to a better grade. In most introductory science courses, especially in physics and chemistry, a significant portion of an undergraduate students' grade comes from their performance on solving quantitative problems for homework, quizzes, and exams. Thus, the grading of these quantitative problems is an important aspect of the course and can be expected to significantly impact students study habits and learning outcomes.

Some researchers argue that faculty should reduce the amount of student work that is graded. When students receive graded work, they focus on the grade, or on getting a better grade than their peers, more so than on the instructor feedback or on reflecting on the quality of their work (Black et al. 2003; Rust 2002). Although we are sympathetic to this argument, we also believe it is unlikely that grading will cease to be an important part of teaching college science courses. Thus we think that it is important to align grading practices as closely as possible with desired study habits and learning outcomes.

Tools for Teacher Reflection

It is common in all parts of life for people to experience conflicts between their values and their actions. Identifying and reflecting on these discrepancies has been identified as a key component leading to improvement in performance (e.g., Schon 1983). This process of identifying and reflecting on mismatches between values and practice also improves teaching performance (Hubball et al. 2005; Kane et al. 2004). Teachers, especially those not skilled in critical reflection, may benefit from external ideas to guide the reflection process (Hubball et al. 2005; Richardson 1990). Rowe's (1974a, b) concept of wait time is an example of an external idea that can help guide reflection (Richardson 1990; Blase and Blase 1999). Wait time works as an excellent tool for reflection because it is simple to understand and evaluate, teacher practices are often in conflict with their value of allowing adequate wait time, there are actionable ways to reduce the conflict, and increases in wait time can have a large impact on student learning. Thus, we believe that it is fruitful for researchers to identify additional tools that can help teachers reflect on and improve their practice.

Research Focus

From the literature review we conclude that quantitative problem solving is an important feature of most introductory, college-level science courses. Despite requirements that students “show their work” on most quantitative problems, research suggests that students typically leave these courses with poor problem-solving skills. This problem may be due in part to faculty grading practices, especially the practice of assigning a grade or numerical score without explanatory comments. Whereas instructors often say they want to see reasoning in a student solution, they may actually penalize students for showing their reasoning, or reward unclear reasoning. The “burden of proof” construct explained by Henderson et al. (2004) may explain how faculty negotiate between competing values when grading student work.

In the current study, we extend the prior work by Henderson et al. (2004) with faculty in chemistry ($n=10$) and earth science ($n=7$), in order to document whether the misalignment between explicit values and grading practices exists across science faculty more generally. In our analysis of interviews with faculty as they grade and discuss example student solutions, we address the following key research questions:

- Which, if any, of the three previously identified values are expressed by chemistry and earth science faculty as they grade quantitative problems?
- How do faculty from chemistry and earth science weigh expressed and implicit values in their grading decisions?
- Are chemistry and earth science faculty more likely to place the burden of proof on themselves or on their students?

Methods

Overview

Data were collected via semi-structured interviews with chemistry and earth science faculty as they graded typical examples of student work. The research design is modelled after a think-aloud protocol (e.g., Ericsson and Simon 1980) in that faculty participants were prompted to verbally report their reasons for grading choices as they were engaged in grading example student work. Chemistry faculty graded and discussed only example student solutions to the chemistry problem, and earth science faculty graded and discussed only example student solutions to the earth science problem. Numerical scores assigned to student problems were recorded and analyzed. Interviews were transcribed and coded, and themes were compared to values reported by Henderson et al. (2004).

Participants and Recruitment

Ten chemistry faculty and seven earth science faculty participated in the study. In order to compare our results to previous work, only tenure-track or tenured faculty (rank of Assistant, Associate, or Full Professor) who had taught introductory science courses within the last 3 years at research-intensive universities were included in the study. The chemistry faculty participants, all male and from a single department at one doctoral-granting university in the midwestern United States, had a range of 4–15 years of teaching experience. The earth science faculty participants were from geography or earth science

departments at three different institutions, as these departments tend to be smaller and have only a few faculty who teach physical geography courses. One earth science faculty participant was from the same university as the chemistry participants, five were from a nearby doctoral-granting research university in the midwestern United States, and one was from a masters-granting university in the western United States. The earth science faculty included 3 females and 4 males, and a range of 5–40 years of teaching experience.

Potential participants were initially contacted by phone and invited to participate in the study. Once informed consent was obtained, a time for the interview was scheduled. To preserve anonymity, each participant was assigned an instructor number. All of the research activities were conducted under a protocol approved by our Human Subjects Institutional Review Board.

Quantitative Problems and Student Solutions

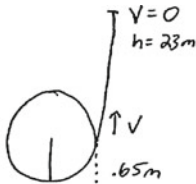
We created one chemistry problem and one earth science problem that carefully mirrored the original physics problem used by Henderson et al. (2004). Important features of the problem design were that each problem: (1) is appropriate for an introductory course in the discipline, (2) requires a quantitative solution, (3) involves combining more than one important concept from the discipline, and (4) involves multiple steps that have more than one possible path to a correct solution. The physics problem and two student solutions from the original study are provided in Fig. 2. The chemistry problem (Fig. 3) and the earth science problem (Fig. 4) were chosen as typical quantitative, free-response problems encountered in an introductory, college-level course. For chemistry, we chose a stoichiometry problem that could be assigned in a general chemistry course. For earth science, we used an adiabatic rise problem that could be assigned in a physical geography course.

We then created five student solutions to each problem based on actual student work; solutions again mirrored the original five solutions from Henderson et al. (2004; Fig. 2). Two of the five student solutions, which were deliberately constructed to elicit conflicts among values, are discussed in this paper. Student Solution D (SSD; Figs. 2, 3, and 4) is a detailed solution that shows student thinking but has significant conceptual and mathematical errors, yet the error combination results in a correct answer. Student Solution E (SSE; Figs. 2, 3, and 4) is brief and does not clearly show student thinking, but has a correct answer. SSE could have made the same combination of errors as SSD, or could have done the problem correctly; the reasoning expressed in the solution is ambiguous.

Interviews

Several days prior to the interview, each faculty participant was emailed the problem text and asked to review the problem (chemistry faculty received the problem text from Fig. 3, and earth science faculty received the text from Fig. 4). During the interview, faculty were presented with the five student solutions and asked to rank them from best to worst. Following the procedure used in the original Henderson et al. (2004) study, obvious errors in the student solutions were identified by written boxed comments as shown in Figs. 2, 3, and 4; these were pointed out to faculty during the interview. Faculty were told to assume that this problem had been assigned on a quiz or exam, and that it was late enough in the semester that students were familiar with the expectations for completing this type of work. Participants were then asked to give each solution a numeric grade from 0 to 10, and to explain their grading of each solution. Finally, they were asked to go through each solution and explain, as best as possible, the student's thinking as she/he solved the problem. Interviews were audio and video recorded.

PHYSICS: You are whirling a stone tied to the end of a string around in a vertical circle having a radius of 65 cm. You wish to whirl the stone fast enough so that when it is released at the point where the stone is moving directly upward it will rise to a maximum height of 23 meters above the lowest point in the circle. In order to do this, what force will you have to exert on the string when the stone passes through its lowest point one-quarter turn before release? Assume that by the time that you have gotten the stone going and it makes its final turn around the circle, you are holding the end of the string at a fixed position. Assume also that air resistance can be neglected. The stone weighs 18 N.



SSD

Energy conservation between top and release

$$\frac{1}{2}mv^2 = mgh$$

$$v^2 = 2gh$$

$$v = \sqrt{2(-9.8)23}$$

$$v = 21.2$$

uses h instead of h-R

makes sign error

changes sign

between release and bottom $T \perp v$ so no work done

\therefore Energy is conserved and velocity is the same

$$\Sigma \vec{F} = m\vec{a}$$

$$T - mg = \frac{mv^2}{R}$$

$$T = 18 + \frac{18}{9.8} \cdot \frac{21.2^2}{0.65}$$

$$= 1292 \text{ N}$$

uses v_{release} instead of v_{bottom}

$$v^2 = 2gh$$

SSE

$$F - mg = \frac{m2gh}{R}$$

$$F = 18 + \frac{2 \cdot 18 \cdot 23}{0.65} = 1292 \text{ N}$$

◀ **Fig. 2** Original text of the physics problem, student solution D (SSD), and student solution E (SSE) as appearing in Henderson et al. (2004). Obtained with permission of C. Henderson. This problem and student solutions were used as a model to develop the chemistry and earth science problems for our study (Figs. 3 and 4)

Data Analysis

Numerical scores were examined to generate simple descriptive statistics for each solution for each discipline. In particular, we noted whether faculty graded SSD or SSE more highly. Interviews were thematically coded using an initially emergent scheme (e.g., Strauss and Corbin 1990) as described below. However, the themes we generated were so similar to values identified by Henderson et al. (2004) that we later adopted those authors' categorizations. HyperRESEARCH, a qualitative data analysis software package, was used in assigning and organizing the codes.

All interviews were transcribed and any typographical errors or misspellings were corrected. Emergent coding of the ten chemistry faculty interviews was performed first, and proceeded in five stages as follows. First, two researchers with a background in chemical education (HF and JM) reviewed all transcripts to obtain a general sense of the data (Strauss and Corbin 1990). Second, these researchers independently recorded meaningful ideas, merged together similar ideas into codes, and independently applied the relevant codes to the four chemistry transcripts that were randomly selected to develop the preliminary coding scheme. Third, after the independent analysis, the two researchers compared their codes, resolved any disagreements via discussion, and finalized the coding scheme. Fourth, the finalized coding scheme was applied by each researcher independently to all of the transcripts (through constant comparison, e.g., Strauss and Corbin 1990) with new emerging codes also included. Fifth and finally, salient coding categories were then developed into common themes. The saliency of themes was judged by frequency of occurrence, uniqueness, and apparent connectedness to other categories. Themes generated in this manner closely matched the three values identified by Henderson et al. (2004), except for one theme we determined to be new. Pre-discussion agreement for identifying themes in the data and assigning burden of proof was 93%, and coding was compared and discussed until complete agreement was reached for all interviews.

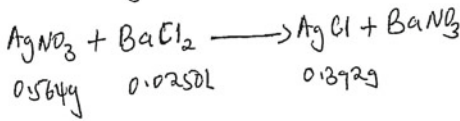
The earth science faculty interviews were coded after analysis of the chemistry interview data was complete. Two researchers with backgrounds in earth science education (HP and JB) together reviewed two of the earth science interviews and developed a preliminary coding scheme based on the themes identified by the chemistry research team and values identified by Henderson et al. (2004). One earth science education researcher (JB) then applied the coding scheme to all of the interviews, allowing for additional codes and themes to emerge from the data as needed. The second earth science education researcher independently coded two interviews. Pre-discussion agreement for identifying themes in the data and assigning burden of proof was 80%, and coding was compared and discussed until complete agreement was reached for all interviews.

Results

In this and following sections, we report our data and compare findings to the prior work by Henderson et al. (2004). The lead author of the previous study (CH) participated in the current work to ensure that our methods faithfully replicated the earlier work. Thus we can make direct comparisons between the two studies, and discuss the combined effects.

CHEMISTRY: 0.564 grams of AgNO_3 is dissolved in 25.00 mL of 0.250 molar BaCl_2 . A precipitate forms and is isolated and weighed. Its mass is 0.392 grams. What is the percent yield of the reaction?

Limiting reactant problem



writes BaNO_3 instead of $\text{Ba}(\text{NO}_3)_2$ and does not balance the equation.

Find limiting reactant by finding moles of product that could be made.

$$0.564\text{g AgNO}_3 \left(\frac{\text{mol AgNO}_3}{169.88\text{g/mol}} \right) = 0.00332\text{ mol} \leftarrow \text{smaller}$$

$$0.0250\text{L BaCl}_2 \left(\frac{0.250\text{mol BaCl}_2}{\text{L BaCl}_2} \right) = 0.00625\text{ mol} \leftarrow \text{larger}$$

$\therefore \text{AgNO}_3$ is limiting

Find theoretical yield by $\frac{\text{theoret.} \times 100\%}{\text{actual}}$

uses theoretical yield instead of percent yield.

$$\frac{0.392\text{g AgCl}}{(0.00332\text{ mol}) \left(\frac{143.32\text{g}}{\text{mol}} \right)} \times 100\% = 82.4\%$$

SSD

$$\text{MW} = \text{g/mol}$$

$$M = \frac{\text{mol}}{\text{L}}$$

$$\frac{0.564}{169.9} = 0.00332\text{ mol AgNO}_3 = \text{AgCl}$$

$$\frac{(25.00)(0.250)}{1000} = 0.00625\text{ mol BaCl}_2$$

$$\frac{0.392\text{ AgCl}}{(0.00332)(143.32)} = 82.4\%$$

SSE

◀ **Fig. 3** Text of the chemistry problem, student solution D (SSD), and student solution E (SSE) used in this study

Numerical Scores

Faculty scores of each solution varied greatly among individuals, with scores ranging from 0 to 10 (Table 2). The mean scores of 7.8 for the chemistry and 8.1 for the physics SSD were similar, however, the earth science SSD mean score of 5.4 was lower (Table 2). For SSE, both the earth science mean score of 6.1 and the chemistry mean score of 5.1 were lower than the mean physics score of 8.2 (Table 2). Although some of the differences in the average scores between the disciplines are statistically significant, we do not think that this is a particularly meaningful result because the actual scores assigned would likely be dependent on faculty perceptions of the difficulty of the particular problem used. We did not seek to control for the difficulty level of the problems across disciplines.

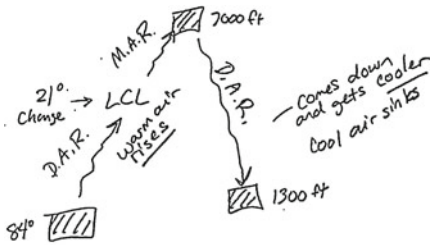
We were particularly interested in how the scores assigned to SSD compared to those assigned to SSE. Similar to the physics study, earth science faculty were evenly split between those who scored SSD higher than SSE (3) and those who scored SSE higher than SSD (3; Fig. 5 and Table 2). The remaining earth science instructor gave the two solutions identical scores. Conversely, 8 of the 10 chemistry faculty scored SSD higher than SSE, with 2 remaining faculty giving the two solutions equivalent scores (Fig. 5 and Table 2). Results of a two-tailed Fisher's exact test suggest that there was a statistically significant difference between the three discipline groups in terms of how the scores assigned to SSD compare to SSE ($p=.022$).

Manifestation of Values in Interview Data

During the interview analysis procedure, we concluded that the emergent codes for the chemistry and earth science faculty data were not different enough from the original physics study to warrant inclusion in our discussion of the results. Thus we report our result in the context of the previously identified values of Henderson et al. (2004), except that we also include one new value identified in the current work. Interview quotes illustrating these values are given in Table 3. Similar to the results reported by Henderson et al. (2004) for physics faculty, all of the chemistry and earth science faculty expressed Value 1; all participants stated that they wanted to see evidence of student reasoning in the solution to a problem, as the reasoning could help faculty diagnose whether the student understood the concepts, as well as determine how to correct the student's mistakes (Table 3). This value was dominant when comparing SSD to SSE. For example, many faculty commented that they preferred SSD over SSE because (1) it clearly showed student reasoning, and (2) the faculty could determine from the solution where the student had made mistakes and therefore could offer feedback for how to correct the mistakes.

For Value 2, all of the chemistry and earth science faculty indicated a desire to find clear evidence of student mistakes in order to deduct points (Table 3). Although SSD had explicit reasoning and had the correct answer, the mistakes in this solution served as evidence for 9 of 10 chemistry faculty and all of the earth science faculty to deduct points. On the other hand, most faculty acknowledged that SSE had little reasoning shown; however, 5 chemistry and 3 earth science faculty explicitly expressed a reluctance to deduct points from what they viewed as a potentially correct solution (Table 3). Of these, one chemistry and two earth science faculty gave solution E full credit.

EARTH SCIENCE: An air parcel is forced to rise over a mountain to a height of 7000 feet. The air parcel's starting temperature is 84°F at sea level on the windward side of the mountain. It reaches its dew point at approximately 63°F. What is the approximate temperature of this air parcel when it descends back to 1300 feet on the leeward side of the mountain? Assume that the air parcel is not saturated during its descent.



Adiabatic rise
Problem

STEP 1 - rising air has 21° Temp Change

$$\frac{21^\circ}{5.5} \times 1000 \text{ ft} = 3818 \text{ ft}$$

$$\begin{array}{r} \text{Still has } 7000 \\ - 3818 \\ \hline 3182 \text{ ft to go} \end{array}$$

$$\frac{3182 \text{ ft}}{1000 \text{ ft}} \times 3.3^\circ = 10^\circ$$

$$\text{rising total} = 84 + 19 + 10 = 115^\circ \text{ at top}$$

STEP 2 - air sinks and cools

$$\frac{7000 \text{ ft}}{1300 \text{ ft}} \times 5.5^\circ = 31^\circ$$

$$115 - 31 = \boxed{84^\circ}$$

Incorrect reasoning

- rising air cools
- descending air warms

SSD

$$21 \times \frac{1000}{5.5} = 3818$$

$$7000 - 3818 = \frac{3182}{1000} \times 3.3 = 10$$

$$\frac{7000}{1300} \times 5.5 = 31$$

$$\boxed{84^\circ}$$

SSE

◀ **Fig. 4** Text of the earth science problem, student solution D (SSD), and student solution E (SSE) used in this study

The majority of faculty expressed Value 3 when evaluating SSE (Table 3). Seven chemistry faculty and 4 earth science faculty felt that SSE had the correct thought processes, only that the student did not display his or her reasoning. Since there was no evidence in the solution that the student had made an obvious mistake, these faculty assumed that the student had done the problem correctly.

Table 2 Physics (P), chemistry (C), and earth science (E) faculty numeric grades of SSD and SSE, and instructor grading orientation. Mean and standard deviation, median, and mode of scores for each discipline are also reported. Physics data from Henderson et al. (2004) are included in the table for comparison to the earth science and chemistry faculty scores

Instructor	SSD grade	SSE grade	Burden of proof
P1	5.5	4	On student
P2	9.5	10	On instructor
P3	6.8	9.2	On instructor
P4	10	7	On instructor
P5	7.5	10	On instructor
P6	9	9	On instructor
<i>P mean</i>	<i>8.1</i>	<i>8.2</i>	
<i>P stdev</i>	<i>1.7</i>	<i>2.3</i>	
<i>P median</i>	<i>8.3</i>	<i>9.1</i>	
<i>P mode</i>	<i>NA</i>	<i>10</i>	
C1	7	5.5	On student
C2	6	6	On instructor
C3	4	2	On student
C4	10	10	On instructor
C5	9.5	8	On instructor
C6	9	1	On student
C7	8	4	On instructor
C8	8	7	On student
C9	8	4	On student
C10	8.5	3.8	On student
<i>C mean</i>	<i>7.8</i>	<i>5.1</i>	
<i>C stdev</i>	<i>1.8</i>	<i>2.7</i>	
<i>C median</i>	<i>8</i>	<i>4.75</i>	
<i>C mode</i>	<i>8</i>	<i>4</i>	
E1	4	8	On instructor
E2	1	0	On student
E3	6	10	On instructor
E4 ^a	6.5	6.5	On instructor
E5 ^a	9	4	On student
E6	5	10	On instructor
E7	6	4	On student
<i>E mean</i>	<i>5.4</i>	<i>6.1</i>	
<i>E stdev</i>	<i>2.5</i>	<i>3.7</i>	
<i>E median</i>	<i>6</i>	<i>6.5</i>	
<i>E mode</i>	<i>6</i>	<i>4, 10</i>	

^a Instructor gave the solution a range of possible scores; mean of the range is reported in the table.

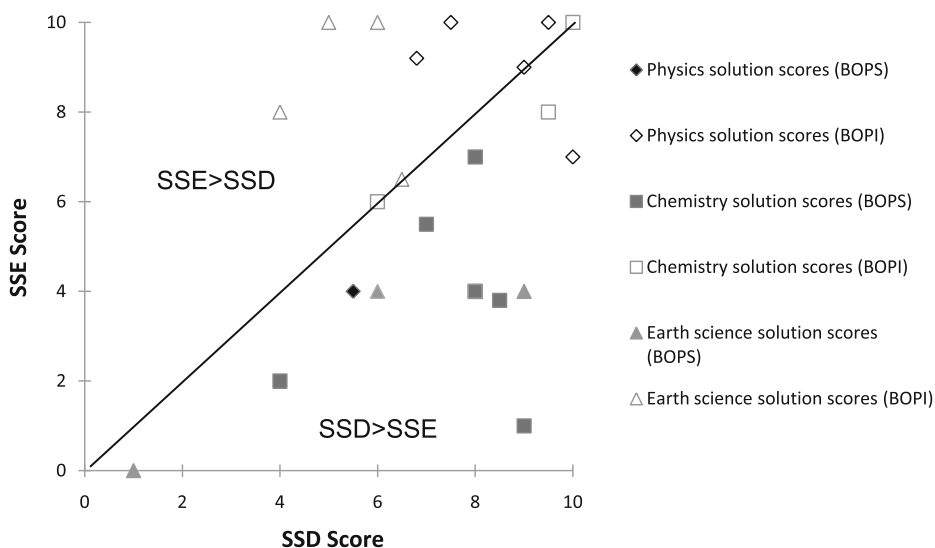


Fig. 5 Faculty numeric scores of SSD and SSE for three disciplines. BOPS indicates burden of proof on the student; BOPI indicates burden of proof on the instructor. Note that all of the faculty who scored SSE > SSD placed the burden of proof on themselves. Physics scores from Henderson et al. (2004) are included in the plot for comparison to chemistry and earth science faculty scores

In addition to these, a fourth value arose predominantly in the chemistry interview data but was also apparent in the earth science data (Table 3). The chemistry instructors felt that discipline and organization in indicating the units, labels, conversions, correct use of significant figures, and writing and balancing reaction equations is vital in chemistry problem solving. Similarly, all of the earth science instructors expressed a desire to see units clearly labelled in the student solutions, and 3 of the instructors also explicitly mentioned that they preferred solutions that included a diagram. Our analysis suggests that the majority of instructors value solutions that do not just show student thinking, but show it in a methodical and organized fashion.

Discussion

Value Conflicts and the Burden of Proof

All chemistry and earth science faculty expressed more than one value while grading student solutions. When values conflicted, the final grading decision was due to the perceived importance given to each value. For example, in grading Student Solution E (SSE), Value 1 (wanting to see reasoning) and Value 4 (indicating units, equations and/or diagrams, etc.) would suggest a low grade, whereas Value 2 (not wanting to deduct points from a student who might be correct) and Value 3 (projecting correct thought processes onto a student solution) would suggest a high grade (Fig. 1). The resultant grade would therefore depend on whether the instructor valued Value 1 and/or 4, or Value 2 and/or 3 more highly. Thus, it appears that the relative strength of these four values and how these values interacted with a particular solution determined what grade was assigned.

Table 3 Examples of values identified in the interview data for chemistry and earth science faculty

Value	Chemistry example	Earth science example
One: desire to see student reasoning to know if the student really understands	Instructor C7: "I appreciate student solution D because it does give me a chance to better understand what the student was thinking as they did the problem... at least my ability to interpret whether they are in need of some guidance, I think, is much easier. For student E... [I would not] be able to say 'this I believe is where you made a mistake...'"	Instructor E3: "I always say show your work...and diagrams would be helpful. ...diagrams would be helpful for the people who would have gotten partial credit - at least I see where they messed up."
Two: desire to deduct points from solutions that are clearly incorrect	Instructor C8: "This one [SSD] on my scale, that's minus two for not balancing the reaction; they did these [compared moles] both correctly; that's based on that [the limiting reactant has smaller moles], so they got that. So they get 8 out of 10."	Instructor E2: "This student [SSD] complied with expectations but did not think it through correctly... wrong numbers and wrong physical processes...severe problems, I'd give a one [point] because there is work shown... [but] reasoning is wrong."
Two: reluctance to deduct points from solutions that might be correct	Instructor C4: "student solution E has got the correct answer and he used a very simple way to write the solution, but all the stages are right; all the conversions are correct, so I give him 10... I try to give [students] more credit as long as they write something which seems right."	Instructor E3: "Well, this person [SSE] didn't show their work, but they got the right number and it looks like they did everything right. I guess we've got no choice but to give them a 10."
Three: tendency to project correct thinking on to ambiguous solutions	Instructor C7: "This student [SSE], I think this student knew what they were doing; they actually had the ability to do all of the detail work... they clearly indicate what they know about stoichiometry and solutions at the top, but I just think that they felt like they didn't have to write down any details."	Instructor E1: "[SSE has] no organization, no units, and it's impossible to follow the logic. I always debate on this how much to penalize because I always say to show all work. There is enough chicken scratching for me to know they knew what they were doing, so it's a minor penalty."
Four*: desire to see an organized, methodical solution with units clearly labeled	Instructor C5: "When I give a problem and I say I want these elements in the problem; I want the correct reaction balanced or charges; mass; I want quantities labeled; I want the units in there and if you do that even if you get the problem wrong you gonna get a half credit."	Instructor E7: "And, you [the student] can't just throw some numbers together in your head and get an answer - you have to check your units. You have to draw a picture. You have to identify what's known, and most importantly, identify what's unknown."

*Value not previously identified by Henderson et al. (2004).

As found by Henderson et al. (2004) for physics faculty, burden of proof is a hidden construct apparent in both the chemistry and earth science interview data that may explain

faculty grading decisions. Based on our analysis of the interviews, 6 of 10 chemistry faculty and 3 of 7 earth science faculty were judged as consistently placing the burden of proof on the student (Table 2), requiring students to show evidence of understanding in their solutions. The remaining 8 faculty placed the burden of proof on the instructor (Table 2), requiring the instructor to find explicit errors in student reasoning in order to deduct points. These differences in the placement of the burden of proof between the chemistry and earth science faculty are not statistically significant (two-tailed Fisher's exact test, $p=0.11$).

As an example of burden of proof on the instructor, Instructor E3 expressed all four values, but ended up resolving the conflict and assigning SSE full credit in favour of Values 2 and 3. In scoring SSE, Instructor E3 says:

Well, this person got it right and it looks like their logic was right. That's the best paper so far, even though they didn't draw a nice mountain. Guess they just knew it cold and didn't need to put it together like I do.

Similarly, Instructor C4 also resolved the conflict in favour of Values 2 and 3:

I don't like [SS]E - although he or she may be smart to get the correct answer and everything right, but from a simple writing you cannot check his thinking, you know. I don't want to take any credit off but I will just tell him directly that he should give people a little more writing to enhance understanding just in case the final result is wrong.

These faculty felt that although work was not shown, the student was thinking correctly because there was no evidence to the contrary, so they were reluctant to deduct points.

A burden of proof on the student is exemplified by Instructor C6, who gave SSE the lowest score of all chemistry faculty. This instructor expressed more than one value, but resolved the conflict in favour of Value 1:

...there's no explanation how it [the problem] was done, I cannot see.... if the student knew this or if it was just copied from somewhere. So this student [SSE] might actually be better than this one [SSD] but since the method of solving the problem is not exposed correctly, I cannot grade that work.

A similar sentiment was expressed by Instructor E2, who scored SSE a zero:

I don't really know what student E was thinking... I fault student E because nothing is labelled, crudely the work is shown... it's not clear what the work refers to. Personally I'm irritated by this kind of scant answer.

Since the solution showed no evidence of student understanding, these faculty felt unable to give the student credit despite the correct final answer.

Although there were no statistically significant differences in where faculty from the three disciplines placed the burden of proof, there were statistically significant differences between the disciplines in terms of which solution was rated the highest. The chemistry faculty were significantly more likely to grade SSD higher than SSE (two-tailed Fisher's exact test $p=.022$). Eighty percent of chemistry faculty did this, compared to 29% of earth science faculty and 17% of the physics faculty. A tentative explanation for this finding is that chemistry problems nearly always involve sub-microscopic systems. Although graphic representations of the systems involved are certainly helpful in physics and earth science; in chemistry they are crucial to bring concreteness to what is otherwise invisible to the human eye. Facility in using the various common symbolic representations for chemicals (names, formulas, graphical models) has been shown to correlate with problem solving ability (Kozma and Russell 1997). Successful problem solvers are more likely to write out the

chemical equation and the symbols for the chemical species involved (Camacho and Good 1989). Given the critical nature of the use of graphical representations in chemistry, it is perhaps not surprising that chemists in our study were more skeptical of SSE's solution, which appeared to be able to find the correct answer without ever explicitly symbolizing the chemical system. Additional work, using other problems and more targeted questioning, will need to be done to test this tentative explanation.

Grading and Messages Sent to Students

Table 1 shows schematically how an instructor's orientation towards grading may affect a student's inclination to show their reasoning on a problem solution. Students typically begin introductory science courses without showing their reasoning on problem solutions, either because they have not yet learned how to show reasoning, or because they have learned from prior experience that showing reasoning tends to hurt their course grade (Elby 1999). If students are graded in a way that places the burden of proof on the instructor (as 47% of the earth science and chemistry faculty did), they will likely receive more points if they do not expose much of their reasoning and allow the instructor to instead project his/her understanding onto the solution. On the other hand, if they are graded in a way that places the burden of proof on the student to either demonstrate his/her understanding or produce a scientific argument, they will receive very few points unless they show their reasoning. Most instructors tell students that they want to see reasoning in problem solutions, however students quickly learn about an instructor's *real* orientation towards grading by comparing their graded assignments with those of their classmates, or by comparing their own grades from one assignment to the next.

The grades assigned to SSD and SSE by all faculty, which includes the chemistry and earth science instructors who participated in this study combined with results from the 30 physics instructors who participated in the Henderson et al. (2004) study, suggest that about half (49%) of all science instructors gave students an incentive for showing their reasoning (i.e., graded $SSD > SSE$). For these instructors, their values were aligned with their actual grading practices. However, a significant minority of instructors (34%), in fact, displayed a misalignment between values and grading practices in that they penalized students for showing their reasoning (i.e., graded $SSD < SSE$). Our analysis is consistent with the earlier findings of Henderson et al. (2004), which suggests that many instructors have internal conflicts related to grading student solutions. These conflicts may result in instructors placing the burden of proof on themselves when grading and, thus, sending the unintended message to students that reasoning is not valued. The results of this study imply that science faculty could benefit from reflecting on their grading practices to make sure that they are not sending these unintended messages to students. Identifying and reflecting on mismatches between values and practice has been shown to improve teaching performance (Hubball et al. 2005; Kane et al. 2004). Professional development providers might use the concept of burden of proof to encourage this reflection.

Strengths, Limitations, and Future Work

The nature of qualitative research often precludes broad generalizations, so we are unable at present to determine how widespread this misalignment between faculty grading practices and values may be. In particular, we suspect that faculty responses to the grading task may be significantly impacted by the nature and perceived difficulty of the specific problem that was used in the study. Nonetheless, the presence of similar values and decision processes among faculty across three disciplines suggests that the values and the concept of burden of

proof may be widely applicable. Future work could substantially extend our initial findings by examining grading practices and values among a much larger faculty population using more than one problem in each discipline. This work also raises intriguing questions as to whether there are subtle differences between the values of different disciplines (e.g., why the chemistry faculty graded SSD higher than SSE), which could be explored using a larger faculty population and targeted questions. Finally, future work in faculty development could pursue our suggestion that the burden of proof concept may serve as a useful tool to prompt faculty to closely examine the alignment between their grading practices and values.

Conclusions and Implications

This study documents a potentially widespread misalignment between the expressed desire to encourage students to show their work on problem solutions and actual grading practices. The 34% of participants who scored SSE higher than SSD across three disciplines (combined data from the current study and Henderson et al.'s 2004 study), could be viewed as not providing an incentive for students to show their reasoning by scoring the solution without clear reasoning higher than the solution with reasoning. These faculty might also be viewed as actually penalizing reasoning by taking off points for the obvious errors in SSD but not for the lack of reasoning in SSE.

If we truly value seeing student reasoning in quantitative problem solutions – something recommended by most educational research – then placing the burden of proof on the student, as was done by only 43% of the faculty across three disciplines (combined data from the current study and Henderson et al.'s 2004 study), is one way to resolve grading conflicts. Telling students to show their work is not enough; grading practices must reinforce and reward showing work. We hope that this research can serve as a tool that educational researchers and faculty developers can use to promote cognitive conflict in faculty. This cognitive conflict may in turn lead to reflection on and changes in practice.

Acknowledgements We would like to thank the faculty members who participated in this research project. Dr. Lisa DeChano-Cook and Dr. Robert Ruhf contributed to the development of the earth science problem and student solutions. Comments and feedback by faculty and graduate student colleagues Dr. David W. Rudge, Caitlin Callahan, Matthew Ludwig, and Kate Rowbotham have greatly improved this manuscript. We are also grateful to the two reviewers and editorial staff whose comments enhanced this paper. This project was funded in part by our university, and work by one graduate student researcher (JB) was funded in part by the National Science Foundation (USA) under grant #0733590.

References

- Ault, C. R. (1994). Research on problem solving: Earth science. In D. Gabel (Ed.), *Handbook of research on science teaching and learning*. New York: Macmillan.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, 7–74.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. New York: Open University Press.
- Blase, J., & Blase, J. (1999). Principals' instructional leadership and teacher development: teachers' perspectives. *Educational Administration Quarterly*, 35, 349–378.
- Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.
- Camacho, M., & Good, R. (1989). Problem-solving and chemical equilibrium: successful versus unsuccessful performance. *Journal of Research in Science Teaching*, 26, 251–272.

- Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorizations and representations of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Cooper, M. M., Cox, C. T., Jr., Nammouz, M., Case, E., & Stevens, R. (2008). An assessment of the effect of collaborative groups on students' problem-solving strategies and abilities. *Journal of Chemical Education*, 85(6), 866–872.
- Elby, A. (1999). Another reason physics students learn by rote. *American Journal of Physics*, 67, S52–S57.
- Entwistle, A., & Entwistle, N. (1992). Experiences of understanding in revising for degree examinations. *Learning and Instruction*, 2, 1–22.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215–251.
- Gabel, D. L., & Bunce, D. M. (1994). Research on problem solving: Chemistry. In D. Gabel (Ed.), *Handbook of research on science teaching and learning*. New York: Macmillan.
- Heller, P., & Hollabaugh, M. (1992). Teaching problem solving through cooperative grouping: 2. Designing problems and structuring groups. *American Journal of Physics*, 60, 637–645.
- Henderson, C., Yerushalmi, E., Kuo, V. K., Heller, P., & Heller, K. (2004). Grading student problem solutions: the challenge of sending a consistent message. *American Journal of Physics*, 72, 164–169.
- Hubball, H., Collins, J., & Pratt, D. (2005). Enhancing reflective teaching practices: implications for faculty development programs. *Canadian Journal of Higher Education*, 35(3), 57–81.
- Kane, R., Sandretto, S., & Heath, C. (2004). Excellence in tertiary teaching: emphasising on reflective practice. *Higher Education*, 47, 283–310.
- Kozma, R. B., & Russell, J. (1997). Multimedia and understanding: expert and novice responses to different representations of chemical phenomena. *Journal of Research in Science Teaching*, 34(9), 949–968.
- Larkin, J. H. (Ed.). (1980). *Teaching problem solving in physics: The psychological laboratory and the practical classroom*. New Jersey: Erlbaum Associates.
- Leonard, W. J., Dufrense, R. J., & Mestre, J. P. (1996). Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems. *American Journal of Physics*, 64, 1495–1503.
- Maloney, D. (1994). Research on problem solving: Physics. In D. Gabel (Ed.), *Handbook of research on science teaching and learning*. New York: Macmillan.
- Manduca, C. A., Baer, E., Hancock, G., Macdonald, R. H., Patterson, S., Savina, M., et al. (2008). Making undergraduate geoscience quantitative. *EOS, Transactions of the American Geophysical Union*, 89(16), 149–150.
- Mestre, J. P., Dufrense, R. J., Gerace, W. J., & Hardiman, P. T. (1993). Promoting skilled problem-solving behavior among beginning physics students. *Journal of Research in Science Teaching*, 30, 303–317.
- Qin, Z., Johnson, D., & Johnson, R. (1995). Cooperative versus competitive efforts and problem solving. *Review of Educational Research*, 65(2), 129.
- Ramsden, P. (2003). *Learning to teach in higher education*. London: Routledge.
- Reif, F., & Scott, L. (1999). Teaching scientific thinking skills: students and computers coaching each other. *American Journal of Physics*, 67, 819–831.
- Richardson, V. (1990). Significant and worthwhile change in teaching practice. *Educational Researcher*, 19(7), 10–18.
- Rowe, M. B. (1974a). Wait time and rewards as instructional variables, their influence in language, logic, and fate control: part 1. Wait time. *Journal of Research in Science Teaching*, 11(2), 81–94.
- Rowe, M. B. (1974b). Wait time and rewards as instructional variables, their influence in language, logic, and fate control: part 2. Rewards. *Journal of Research in Science Teaching*, 11(4), 291–308.
- Rust, C. (2002). *The impact of assessment on student learning. Active learning in higher education*. London, CA, and New Delhi: SAGE Publications. doi:10.1177/1469787402003002004.
- Schoenfeld, A. H. (1988). When good teaching leads to bad results: the disasters of “well-taught” mathematics courses. *Educational Psychologist*, 23, 145–166.
- Schon, D. A. (1983). *The reflective practitioner*. New York: Basic Books.
- Simon, D. P., & Simon, H. A. (Eds.). (1978). *Individual differences in solving physics problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Stains, M., & Talanquer, V. (2008). Classification of chemical reactions: stages of expertise. *Journal of Research in Science Teaching*, 45(7), 771–793.
- Strauss, A. L., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory practice and techniques*. Newbury Park: SAGE Publications.
- Van Heuvelen, A. (1991). Learning to think like a physicist: a review of research-based instructional strategies. *American Journal of Physics*, 59, 891–897.
- Walvoord, B. E., & Anderson, V. J. (Eds.). (2010). *Effective grading: A tool for learning and assessment* (2nd ed.). San Francisco: Jossey-Bass.