7.2 THE FUNDAMENTALS IN METEOROLOGY INVENTORY: MOTIVATION AND DEVELOPMENT OF A NEW METEOROLOGY EDUCATION TOOL

Casey E. Davenport*
University of North Carolina at Charlotte

A.J. French
South Dakota School of Mines and Technology

T.L. Koehler and D.R. Vollmer
United States Air Force Academy

1. INTRODUCTION

Instructors often find that the material they teach does not coincide with the material that students actually learn and understand (e.g., Driver 1985; Schneps 1997; Fisher and Moody 2000). There are many stumbling blocks to learning, one of which is the extent of prior knowledge and conceptual understanding. The presence of any persistent *mis*understandings, otherwise known as misconceptions, provides a poor base for additional learning, and can consequently result in poor performance on formal assessments (e.g., Hestenes et al. 1992).

The field of meteorology can be particularly susceptible to students bringing in misconceptions as a result of years of personal experience with the weather (Rappaport 2009). To eradicate misconceptions and improve learning, they must be identified and dealt with head-on (e.g., Posner et al. 1982). Several science disciplines have had great success toward this end by developing standardized assessment exams that are designed to identify common misconceptions for their student populations, including physics (Halloun and Hestenes 1985; Hestenes et al. 1992), astronomy (Zeilik et al. 1997; Hufnagel 2002), biology (Anderson et al. 2002), statistics (Allen et al. 2004), and the geosciences (Libarkin and Anderson 2005). To date, no such broad-reaching exam exists for the discipline of meteorology, putting instructors who want to ensure that their students attain a scientific and accurate understanding of meteorology at a significant disadvantage.

The Fundamentals in Meteorology Inventory (FMI) is a new assessment tool under development whose long-range goal is to improve student understand of basic meteorological concepts. The FMI is a multiple-choice exam that covers all of the broad topics typically covered in an introductory meteorology course. Results from the FMI would be pinpoint consistent areas of struggle for students learning the fundamentals of meteorology, allowing instructors to develop focused and effective teaching techniques that improve student understanding.

* *Corresponding author address*: Casey E. Davenport, University of North Carolina at Charlotte, Department of Geography and Earth Sciences, Charlotte, NC 28223; email: Casey.Davenport@uncc.edu

2. BACKGROUND

2.1 MISCONCEPTIONS IN LEARNING

There are numerous roadblocks that can impede student learning, such as motivation, affect, aptitude, learning style, and exceeding cognitive load (e.g., Sweller 1988; Pass et al. 2003; Roebber 2005; McConnell and van Der Hoeven Kraft 2011; Sweller et al. 2011). Perhaps the most significant impediment to learning involves previous conceptions about the subject held by a student. Incorrect prior knowledge has been shown to significantly impair the accuracy of students' conceptual understanding of a particular subject, thus impacting performance on course assessments (e.g., Hestenes et al. 1992). Indeed, these *mis*conceptions impede students from attaining a more scientific viewpoint and are extremely resistant to traditional classroom instruction, particularly when instructors are ignorant of their existence and do nothing to directly address or correct them (e.g., Halloun and Hestenes 1985; Hestenes et al. 1992; Wandersee et al. 1994).

Meteorology in particular can be susceptible to misconceptions, as students often have years of personal experience with the day-to-day weather before starting formal instruction. Thus, students may have difficulty reconciling the presented material with their observations, allowing poor learning to take hold (Rappaport 2009). Previous research has often focused on the misconceptions of children, who often simply lack meteorological knowledge (e.g., Henriques 2002). At the undergraduate level, previous research has examined misconceptions associated with specific phenomena such as tornadoes or fog (Lewis 2006; Rappaport 2009; Polito 2010). These misconceptions existed at numerous cognitive levels (i.e., freshmen through seniors), and for both majors and non-majors (Polito 2010).

Consistent with the findings of Polito (2010), the meteorology program at the United States Air Force Academy (USAFA) identified misconceptions that

carried throughout the three years of required coursework utilizing a home-grown longitudinal assessment exam known as the Meteorology Program Assessment Test (MPAT). The MPAT is a 39 item test that covers topics across the entire USAFA meteorology program and evaluates the academic evolution of meteorology majors as they progress through the curriculum. In the initial version of the exam, 15 of the 39 questions related to fundamental concepts taught in the first-semester course. The graduating Class of 2010 was the first set of meteorology majors to take the MPAT before, during, and after instruction of their 11 required meteorology courses (hereafter referred to as the pre-curriculum, mid-curriculum, and post-curriculum periods, respectively).

The post-curriculum assessment taken by the Class of 2010 revealed that 4 of the 15 questions related to fundamental concepts had an average score below 60%, demonstrating poor understanding and suggesting the persistence of misconceptions even *after* extensive instruction. Out of the four low-scoring questions, three of them had a considerably *lower* average than on the mid-curriculum assessment. One additional question related to introductory material (with a score above 60% on the post-curriculum assessment) *also* had a lower average than the previous year, resulting in a total of 4 questions with a decreasing average. In other words, performance on nearly one-quarter of the questions related to fundamental concepts *decreased* from the mid-curriculum to the post-curriculum MPAT assessment. Similar results exist for the other graduating classes that took the MPAT their junior and senior year (Classes of 2009-2014; not shown).

In terms of misconceptions focusing on introductory topics, to the best of the authors' knowledge, only a single study has been conducted, Kahl (2008). A survey was devised (and administered at a single institution) with three types of questions asked within each topical subject area: content, application, and deeper application. Questions related to content learning scored quite well (more than 75% saw improvement over the semester in each topical area), while questions related to applications and deep applications of topics saw much less improvement (8-43% of students). The overall portion of students demonstrating a correct understanding of introductory meteorology topics varied significantly in the application questions, between 9% and 78%, with the deeper application questions on the lower end of that range. These results indicate that students tend to excel at *memorizing* content, but struggle to truly *understand* concepts and apply them correctly to given situations.

2.2 CONCEPT INVENTORIES

Conceptual inventories, often given as a multiple-choice exam at the beginning and end of a semester, have been a demonstrated source of vital information for instructors on the depth of student learning. For example, the Force Concept Inventory (FCI; Hestenes et al. 1992) revealed the superficial nature of conceptual understanding of introductory physics topics by a significant proportion of college students. The application of the results from this inventory dramatically shifted perceptions of the teaching and learning of physics, and subsequently radically transformed conventional college-level physics instruction (Gonzales-Espada 2003). Recognizing the successes of the physics community, numerous other disciplines have also developed similar assessment exams, including astronomy (Zeilik et al. 1997; Hufnagel 2002), biology (Anderson et al. 2002), statistics (Allen et al. 2004), and the geosciences (Libarkin and Anderson 2005). Consequently, the authors are confident that the FMI can be a similarly successful source of advancement in teaching meteorology.

3. DEVELOPMENT OF THE FMI

Strong evidence exists that meteorology has misconceptions, but additional work is needed to determine those that are universal at the fundamental level. To date, no study has systematically identified misconceptions of meteorology common to undergraduates at a variety of institutions. A standardized assessment exam known as the Fundamentals in Meteorology Inventory (FMI) has been developed to fill this gap by assessing student understanding of basic concepts addressed in introductory meteorology courses (Davenport et al. 2014). The main goal of the exam is to assist instructors in identifying concepts that may cause the most difficulty for their students. Additionally, the exam provides a means with which learning and teaching effectiveness can be evaluated.

Previous standardized assessments in the science community like the FCI have typically been developed as a multiple-choice exam due to a variety of reasons: objective grading, ease of testing large numbers of students, minimal time commitment for instructors to grade, as well as the ease of applying standard statistical analysis (Engelhardt 2009). As noted in Anderson, et al. (2002), student interviews are the most effective means of identifying misconceptions, and have been used along with open-ended questions in developing many other concept inventories, especially

in identifying meaningful item distractors. However, Anderson et al. (2002) also note that interviews and open-ended questions are logistically impossible to include in such assessments due to the large number of students being measured. Thus, the FMI uses a multiple choice format to maximize objectivity while also allowing for efficient implementation and analysis of results.

In developing the current version of the FMI, assessing higher-order student understanding (as opposed to rote memorization) of meteorological concepts was a central goal. The authors followed the guidelines of Haladyna et al. (2002) in formatting each question. Following an analysis of numerous studies on multiple-choice item-writing guidelines, Haladyna et al. (2002) provides a synthesis of 31 recommendations for writing test items, which the authors sought to follow as much as possible, as appropriate for the intended goals of the FMI.

The specific content of the FMI was largely driven by the broad topics covered in many introductory meteorology courses, split into seven categories divided among 35 total questions: clouds and precipitation, wind, fronts and air masses, temperature, stability, severe weather, and climate (e.g., Fig. 1). In the Fall 2013 semester, the questions were administered in the introductory course at USAFA as an in-class check for understanding. Feedback was solicited from students on question language and answer choices. Additional feedback and edits were given by meteorology faculty at USAFA. Responses were collected and taken into careful consideration with revisions as needed to ensure appropriateness of content, language, and answer choices. Further review and iterations of question content will be ongoing as additional feedback is solicited from additional meteorology faculty across the country, which will be described later as future work.

10. Given the following forecast for Oklahoma City, Oklahoma in late spring, what type of weather boundary is expected to pass through later?

"Warm and humid today, with southerly winds and increasing cloudiness with a chance of thunderstorms in the afternoon. Towards evening, continued warm, drier, with gusty westerly winds."

a. Cold front
b. Warm front
c. Occluded front
d. Dryline

Figure 1. Sample FMI question from the fronts and air masses category.

4. PILOT STUDY RESULTS

To assess the viability of the FMI in identifying misconceptions in introductory meteorology, a pilot study was conducted by faculty at USAFA. Students enrolled in the Spring 2014 Introduction to Meteorology and Aviation Weather (Meteor 320) course at USAFA had Version 1.1* of the FMI administered as a pre-test and post-test. The pre-test, given at the beginning of the semester, measured the extent of prior knowledge that students brought to the course, while the post-test, given on the last day of the class, measured the extent to which student understanding improved beyond prior knowledge. By comparing the pre-test and post-test results, learning gains were determined, in addition to indicating any misconceptions. Meteor 320 represents the first course in the meteorology major course sequence at USAFA, though the majority of students enrolled in the course over the past three years were non-majors (93%), with three majors typically taking the course each semester*.

The FMI pre-test was given on January 8-9, 2014, to 49 of 56 USAFA cadets enrolled in the Spring 2014 offering of the Meteor 320 course. The post-test was then administered on May 8-9, 2014 to 48 of the 56 registered students. Due to students adding the course late or dropping the course altogether, there was an overlap of 41 students who took both the pre-test and the post-test. Thus, our analysis will focus on the scores of the 41 students who completed both tests. The data collected for each student consisted of the responses to each of the 35 test items, whether the response was correct (1) or not (0), and the total number of correct responses.

Table 1 provides a set of summary statistics of the pre- and post-test scores. It is encouraging that the mean, median, and mode score improved from the pre-test to the post-test, indicating a gain of meteorological knowledge. Even so, the post-test average of 19.56 (out of 35 questions, giving a 56% correct response rate) suggests that there is a sizeable fraction of material that students struggle to fully

* The FMI has since been updated to Version 1.3, which modified 5 of the 35 questions. This version was administered to students enrolled in the Fall 2014 Meteor 320 course. The differences between Version 1.1 and 1.3, as well as the comparatively small number of students taking both the pre- and post-test (12), precludes the presentation of the Fall 2014 results here.

* Typical enrollment in Meteor 320 is 70 students per academic year, with more enrolled in the spring.

Statistic	Pre-test	Post-test
Mean	14.66	19.56
Median	15	20
Mode	16	20
Standard deviation	2.82	4.00
Standard error of the mean	0.44	0.62
Range of scores	10	18

Table 1. Summary statistics on FMI scores for the spring 2014 Meteor 320 offering at USAFA, based on formulas from Engelhardt (2009).

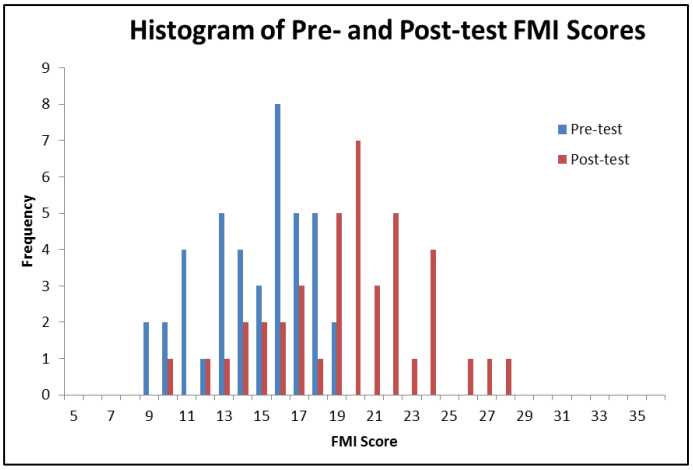

Figure 2. Histogram of the FMI pre- and post-test scores from the Spring 2014 Meteor 320 course at USAFA.

understand, some of which could be due to misconceptions. Additionally, the statistics point to a nearly doubled range of scores for the post-test, indicating that some students improved their scores much more than others. Figure 2 illustrates this shift, with a clear stretching of the distribution of post-test scores compared to the pre-test distribution.

One of the major issues in developing a conceptual inventory such as the FMI is evaluating the suitability of individual items in identifying misconceptions. Utilizing item response theory is thus of use, as it does not assume that each test item is equally difficult (Hambleton et al. 1991). For example, Hestenes et al. (1992) used a straightforward item analysis of the FCI that suggested that items with a high correct response rate only provide weak discrimination in identifying misconceptions, and thus should probably be dropped from the inventory. From the Spring 2014 FMI offering at USAFA, the pre-test indicated that Items 5, 8, and 31 had correct response rates greater than 75% (not shown), while the post-test showed that Items 5 and 31 maintained this high correct response rate (Table 2). Based on a straightforward item analysis, these will be considered for removal in future FMI versions since a large fraction of students appear to have a good grasp on the tested concept.

Perhaps the analysis of most interest for the proposed research is comparing items that students performed equally poorly on or worse in both the pre-test and the post-test. Such an analysis would be able to highlight any sustaining student misconceptions. A straightforward and common metric for concept inventories (e.g., Halloun and Hestenes 1985) to identify problem areas is learning gain:

$$Gain = \frac{(posttest\ score - pretest\ score)}{(100 - pretest\ score)}$$

Calculating gains is extremely useful in identifying misconceptions because they are normalized using the maximum possible improvement from pre-test to post-test (100—pretest score). Thus, small or even negative learning gains provide clear evidence that a misconception is present. The overall learning gain for the Spring 2014 offering of the FMI at USAFA was 24%, which is consistent with Table 1 demonstrating a rise in scores from the pre- to post-test. However, learning gains calculated for individual questions tells us a bit more nuanced story. Table 2 indicates 12 questions (approximately a third of all FMI test items) where *negative* learning gains occurred, meaning that students performed *worse* on those questions in the post-test than they did in the pre-test. While these questions spanned a range of topics, over half of them were related to the categories of temperature or climate. Whether or not this is representative of all students in introductory meteorology remains to be seen, and will be evaluated as FMI testing is expanded.

In order to provide confidence that these overall and item-specific learning gains are meaningful, it can be useful to apply a discrimination index (DI to the results (e.g., used by Allen et al. 2004). This index essentially evaluates the effectiveness of test items to discriminate between students who know the answer and those who do not. Specifically, comparisons are made between high test performers and low test performers on each item (statistics textbooks generally suggest the highest and lowest 25% or 27% of scores). Thus, a large and positive correlation suggests that students who get any one question correct also have a relatively high score on the overall exam. Strong negative correlations indicate the opposite effect and could suggest that low-performing students are using test taking techniques to guess the correct answer, or that high-performing students are justifying a wrong answer in some way (e.g., Hufnagel 2002). Thus, analyzing the

discriminatory power of each item gives more confidence that any measured learning gains would be meaningful, providing context for identifying the topics and questions students struggled with the most.

According to Ebel (1972), a negative DI would indicate an item to be discarded (i.e., not a good discriminator); a DI value between 0.0 and 0.19, poor discriminator (needing revision); a DI value between 0.2 and 0.29, acceptable discriminator; a DI value between 0.3 and 0.39, good discriminator; and a DI value greater than or equal to 0.4, an excellent discriminator. Table 3 shows the DI values calculated for the post-test FMI, based on the top and bottom 27% performers. It is clear than the large majority of FMI questions are well within the necessary values for us to be confident in their effectiveness. However, item 22 has a negative DI, indicating that it needs to be discarded from future FMI versions. Additionally, 12 more items are found to be in the "poor" range, signifying a need for revision. This result is perhaps not too surprising, given that the FMI is still in its infancy of development and will undergo several more revisions before achieving a final product. Nevertheless, the DI will be a valuable tool in illustrating the questions needing the most refinement.

5. SUMMARY AND FUTURE WORK

The development of the FMI was motivated by a clear need in the meteorological community to identify the persistent and common stumbling blocks of students. Initiated by discussions among a pool of advanced-degreed meteorologists at USAFA, questions were established through an iterative editing process. These collaborative discussions worked to ensure the validity of both the concepts being tested and the specific responses to the questions (i.e., following Engelhardt 2009). To ensure content validity of the FMI, input will be sought from instructors at institutions offering an undergraduate degree (major or minor) in meteorology or atmospheric science, targeting those with experience in teaching introductory meteorology courses. Instructor feedback will be aggregated to identify common suggestion themes for each question, resulting in edits to each question to produce a more broadly-testable exam. Further iterations will certainly be necessary as more is learned about the reliability and validity of the exam, but this effort to include insight from the meteorological community will provide a more unified vision of what should be tested at the introductory level.

An initial pilot study at USAFA revealed that the FMI has the ability to identify common meteorology misconceptions. While significant work remains to refine the questions, the exam shows promise in assisting the academic meteorology community in promoting enhanced learning outcomes. To ensure that the FMI can be universally applicable to all introductory meteorology courses, testing will soon expand to two additional schools, including the University of North Carolina at Charlotte, and the South Dakota School of Mines and Technology, offering considerably different student populations with which to test.

Post-test

Item	A	B	C	D	% Correct	LG %
1	4	2	34	1	82.9	78.9
2	1	27	10	3	65.9	-10.4
3	19	14	5	3	46.3	44.8
4	8	0	33	0	80.5	54.5
5	1	40	0	0	97.6	82.9
6	15	0	26	0	63.4	30.1
7	14	10	1	16	39.0	32.4
8	5	25	6	5	61.0	-134.3
9	9	7	4	21	51.2	21.2
10	23	5	2	11	26.8	16.9
11	9	32	0	0	78.0	74.4
12	5	7	15	14	36.6	-16.5
13	37	0	3	1	90.2	78.4
14	4	17	1	19	41.5	-11.8
15	4	7	9	21	22.0	-9.6
16	4	10	23	4	56.1	-2.5
17	2	23	11	5	56.1	35.5
18	12	7	21	1	29.3	-2.5
19	6	6	8	21	51.2	45.9
20	4	7	27	3	65.9	11.2
21	21	14	6	0	34.1	16.7
22	6	21	2	12	51.2	41.0
23	7	26	1	7	17.1	3.4
24	10	2	15	14	34.1	-15.9
25	2	29	3	7	70.7	50.8
26	8	31	2	0	75.6	55.4
27	7	31	2	1	75.6	35.9
28	27	7	7	0	65.9	-19.6
29	4	1	5	31	12.2	-8.7
30	14	0	5	22	53.7	38.5
31	39	0	2	0	95.1	70.7
32	5	3	0	33	80.5	67.2
33	4	20	16	1	48.8	-95.7
34	5	15	3	18	43.9	-2.5
35	24	14	1	3	58.5	7.8
Correct Answer			0-25%	25.1-50%	50.1-75%	75.1-100%

Table 2. Response distributions and learning gains (LG) for the post-test FMI. The correct response is shaded blue, and the correct response rate is shaded by the scale shown. Negative learning gains are indicated by red font.

Post-test

Item	DI
1	0.24
2	0.40
3	0.16
4	0.16
5	0.16
6	0.00
7	0.40
8	0.48
9	0.48
10	0.48
11	0.16
12	0.24
13	0.32
14	0.64
15	0.08
16	0.32
17	0.56
18	0.08
19	0.16
20	0.32
21	0.32
22	-0.24
23	0.08
24	0.48
25	0.48
26	0.56
27	0.56
28	0.56
29	0.16
30	0.08
31	0.24
32	0.48
33	0.64
34	0.08
35	0.40

$DI < 0$	Discard
$0.0 \leq DI \leq 0.19$	Poor
$0.2 \leq DI \leq 0.29$	Acceptable
$0.3 \leq DI \leq 0.39$	Good
$DI \geq 0.4$	Excellent

Table 3. Discrimination Index (DI) for each item computed for the FMI post-test in spring 2014. DI values are color-coded according to their utility in discriminating between high- performers and low-performers.

6. REFERENCES

Allen, K., A. Stone, T.R. Rhoads, and T.J. Murphy, 2004: The statistics concept inventory: Developing a valid and reliable instrument. Preprints, *Proceedings of the 2004 American Society for Engineering Education Annual Conference and Exposition*, Amer. Soc. for Eng. Edu., Salt Lake City, UT, 1—15.

Anderson, D.L., K.M. Fisher, and G.J. Norman, 2002: Development and validation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, **39**, 952—978.

Davenport, C.E., C.S. Wohlwend, and T.L. Koehler, 2014: Motivation for and Development of a Standardized Introductory Meteorology Assessment Exam. *Bulletin of the American Meteorological Society*, in press.

Driver, R., 1985: Children's Ideals in Science. Milton Keynes, UK: Open University Press, 208 pp.

Ebel, R.L., 1972: Essentials of Educational Measurement. Oxford, England: Prentice-Hall, 622 pp.

Engelhardt, P.V., 2009: An introduction to classical test theory as applied to conceptual multiple-choice tests. *Getting Started in PER*, **2**, 1.

Fisher, K.M. and D. E. Moody, 2000: Students' misconceptions in biology. Mapping Biology Knowledge, Fisher, K.M., Wandersee, J.M., and Moody, D.E. Dordrecht, The Netherlands: Bluwer Academic, 55—76.

Gonzales-Espada, W.J., 2003: Physics education research in the United States: A summary of its rationale and main findings. *Revista de Educacion en Ciencias*, **4**, 5—7.

Haladyna, T.M., S.M. Downing, and M.C. Rodriguez, 2002: A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, **15**, 309—334.

Halloun, I. and D. Hestenes, 1985: The initial knowledge state of the college physics students. *Amer. Journal of Physics*, **53**, 1043—1055.

Hambleton, R.K., H. Swaminathan, and H.J. Rogers, 1991: Fundamentals of item response theory. Sage, 179 pp.

Henriques, L., 2002: Children's ideas about weather: A review of the literature. *School Science and Mathematics*, **102**, 202—215.

Hestenes, D., M. Wells, and G. Swackhamer, 1992: Force Concept Inventory. *The Physics Teacher*, **30**, 141—158.

Hufnagel, B., 2002: Development of the Astronomy Diagnostic Test. *Astronomy Education Review*, **1**, 47—51.

Kahl, J.D.W., 2008: Reflections on a Large Lecture, Introductory Meteorology Course: Goals, Assessment, and Opportunities for Improvement. *Bulletin of the American Meteorological Society*, **89**, 1029—1034.

Lewis, T.R., 2006: The tornado hazard in southern New England: History, characteristics, student and teacher perceptions. *Journal of Geography*, **105**, 258—266.

Libarkin, J. C. and S.W. Anderson, 2005: Assessment of learning in entry-level geoscience courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*, **53**, 394—401.

Libarkin, J.C., and S.W. Anderson, 2006: Development of the geoscience concept inventory. *Proceedings of the National STEM Assessment Conference, Washington DC*.

McConnell, D.A., and K.J. van Der Hoeven Kraft, 2011: Affective domain and student learning in the geosciences. *Journal of Geoscience Education,* **59**, 106—110.

Paas, F., A. Renkly, and J. Sweller, 2003: Cognitive load theory and instructional design: Recent developments. *Journal of Educational Psychology*, **38**, 1—4.

Polito, E.J., 2010: Student conceptions of weather phenomenon across multiple cognitive levels. Dissertation, San Francisco State University, 97.

Posner, G.J., K.A. Strike, P.W. Hewson, and W.A. Gertzog, 1982: Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, **66**, 211-227.

Rappaport, E.D., 2009: What undergraduates think about clouds and fog. *Journal of Geoscience Education*, **57**, 145—151.

Schneps, M., 1997. Minds of our own: Lessons from thin air [Video]. Cambridge, MA: Harvard University. Science Media Group.

Sweller, J., 1988: Cognitive load during problem solving: Effects on learning. *Cognitive Science*, **12**, 257—285.

Sweller, J., P. Ayres, and S. Kalyuga, 2011: *Cognitive load theory*. Springer, 274 pp.

Wandersee, J.H., J.J. Mintzes, and J. D. Novak, 1994: Research on alternative conceptions in science. *Handbook of Research on Science Teaching and Learning*, D. Gabel, Ed., Simon & Schuster Macmillan, 177—210.

Zeilik, M., C. Schau, N. Mattern, S. Hall, K.W. Tague, and W. Bisard, 1997: Conceptual astronomy: A novel model for teaching postsecondary science courses. *American Journal of Physics*, **65**, 987—996.