5.5 THE FUNDAMENTALS IN METEOROLOGY INVENTORY: RESULTS FROM THE DEVELOPMENT OF A NEW METEOROLOGY EDUCATION TOOL

Casey E. Davenport*
University of North Carolina at Charlotte

Adam J. French
South Dakota School of Mines and Technology

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 (e.g., 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, 2006).

The Fundamentals in Meteorology Inventory (FMI) is an assessment tool currently under development that is designed to measure the presence of student misconceptions of basic meteorological concepts. The FMI is a 35 question multiple-choice exam that covers many broad topics typically discussed in an introductory meteorology course (see Davenport et al. 2015 for more details on how the FMI was developed). Results from the FMI would be able to 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.

Over the past few years, extensive small-scale testing of the FMI has been conducted, iterating on the most effective wording of questions, answer choices, and various visualizations of concepts to test *higher-order* student understanding, as opposed to rote memorization. As we narrow in on a final version that is ready to be distributed to meteorology instructors for wide-spread use, it is vital that the exam undergo reliability and validation testing so that instructors can be confident in the results. Reliability refers to the degree to which the exam produces stable and consistent results; validity indicates the extent to which the exam measures what it is intended to (in this case, meteorological understanding; Engelhardt 2009). There are a variety of ways to measure reliability and validity; not all will be addressed here, but some preliminary statistical analyses of these measures will be described.

2. LARGE-SCALE TESTING

During the Fall 2017 semester, 8 institutions across the country administered the FMI version 1.6 to students enrolled in introductory meteorology courses. The institutions ranged from community colleges up through research-intensive schools, included both private and public institutions, and spanned a variety of geographic locations. The FMI was given as a pre-test on the first day of class, before any instruction began, and on the last day of class as a post-test.

To achieve statistically reliable measures of validity and reliability, it is recommended that validation testing uses a sample size of 5-10 times the number of test items (Englehardt 2009); for the FMI, that means 175-350 students need to be tested. In Fall 2017, a total of 252 students took both the pre- and post-test, well within the desired sample size.

Summary statistics of student performance on the FMI is provided in Table 1. 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 17.64 (out of 35 questions, giving a ~50% correct response rate) suggests that there is a sizeable fraction of material that students struggle to fully understand, some of which could be due to misconceptions. Additionally, the statistics point to a larger range of scores for the post-

* *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

test, indicating that some students improved their scores much more than others. Figure 1 illustrates this shift, with a clear stretching of the distribution of post-test scores compared to the pre-test distribution.

Statistic	Pre-test	Post-test
Mean	12.90	17.64
Median	13	17
Mode	11	13
Standard deviation	3.19	5.77
Standard error of the mean	0.24	0.36
Range of scores	21	26

Table 1: Summary statistics of FMI scores (out of 35) for testing during Fall 2017.

3. VALIDATION AND RELIABILITY RESULTS

To assess the FMI's validity and reliability, a number of statistical analyses will be shown, examining individual test items, as well as the exam as a whole. We will begin with the straightforward measure of item difficulty, which is simply the percentage of students answering each question correctly. When it comes to creating a high-quality assessment instrument, it is desirable for questions to have a range of difficulty, between 0.3 and 0.9, with an average around 0.5 (Engelhardt 2009). This produces a range of scores, and suggests that the test is able to discriminate between students who truly do and do not know the material. For the FMI post-test, 33 out of 35 questions were within the ideal range, with only 2 items slightly below 0.3 (Fig. 2). Additionally, the average difficulty was 0.5, an ideal value for producing maximum discrimination (Engelhardt 2009).

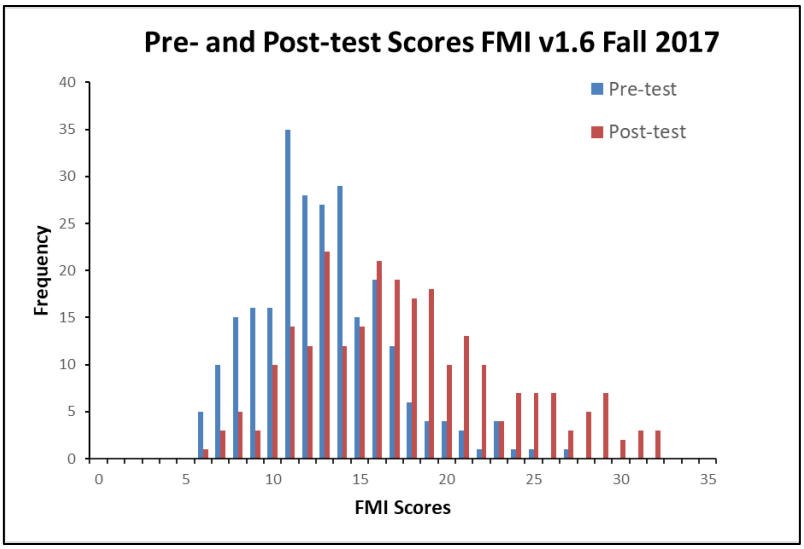

Figure 1: Histogram of FMI pre- and post-test scores .

A common measure of statistical discrimination for assessments like the FMI is known as the discrimination index (DI). 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 (typically, the top 27% students) and low test performers on each item (the bottom 27% students). The DI is calculated for each test item in the following manner:

$$DI = \frac{H-L}{N}$$

where H (L) represents the number of correct answers from the top (bottom) test performers and N represents the number of students that represent 27% of the overall sample size. 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

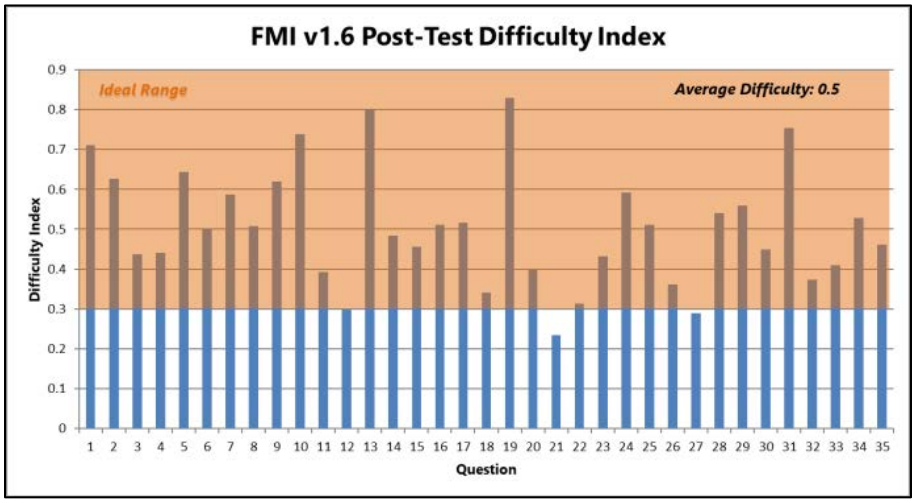

Figure 2: The FMI post-test difficulty index for each question, with the ideal range highlighted in orange.

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 2 shows the DI values calculated for each item on the post-test. Encouragingly, over ¾ of the questions are considered "excellent" or "good" discriminators, the majority of which are excellent. Five questions are within the "acceptable" range, while only 2 are "poor" and need revision.

Post-test v 1.6

Item	DI
1	0.44
2	0.38
3	0.35
4	0.32
5	0.38
6	0.72
7	0.41
8	0.46
9	0.56
10	0.46
11	0.54
12	0.16
13	0.21
14	0.51
15	0.44
16	0.44
17	0.25
18	0.46
19	0.31
20	0.25
21	0.24
22	0.38
23	0.63
24	0.50
25	0.19
26	0.43
27	0.26
28	0.63
29	0.53
30	0.37
31	0.31
32	0.44
33	0.54
34	0.44
35	0.47

Table 2: Discrimination index of each FMI test item, based on post-test scores.

DI < 0	Discard
0.0 ≤ DI ≤ 0.19	Poor
0.2 ≤ DI ≤ 0.29	Acceptable
0.3 ≤ DI ≤ 0.39	Good
DI ≥ 0.4	Excellent

A related item statistic is the point biserial correlation, which measures the correlation between item correctness and the whole exam score. It is calculated using the following equation:

$$r_{pbs} = \left(\frac{\bar{x}_{correctly} - \bar{x}_{whole\ test}}{\sigma_{whole\ test}}\right)\sqrt{\frac{p_i}{1-p_i}}$$

where $\bar{x}_{correctly}$ is the average total score for those students who answered item i correctly, $\bar{x}_{whole\ test}$ is the average total score for the whole sample, $\sigma_{whole\ test}$ is the standard deviation of the total score for the whole sample, and p_i is the difficulty index for item i. Clearly, it is desirable to have a good, positive correlation between answering a test item correctly or incorrectly and the overall test score. The necessary threshold to meet for each test item is a correlation ≥ 0.2 (Engelhardt 2009). As shown in Fig. 3, every question has a correlation ≥ 0.2 except for question 25, which is just under that value. Given that question 25 also has a low DI, this measure provides further evidence that this item should be edited or simply removed from the final version.

In addition to assessing the validity and reliability of individual test items, statistical measures of the exam as a whole are also calculated. The Kuder-Richardson 20 (KR-20) metric is a measure of exam consistency for dichotomously scored test items (i.e., right or wrong) by quantifying the extent to which different groups of questions would produce similar or different results. Specifically, it represents a correlation that is calculated in the following manner:

$$KR-20 = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} p_i(1-p_i)}{\sigma_t^2}\right)$$

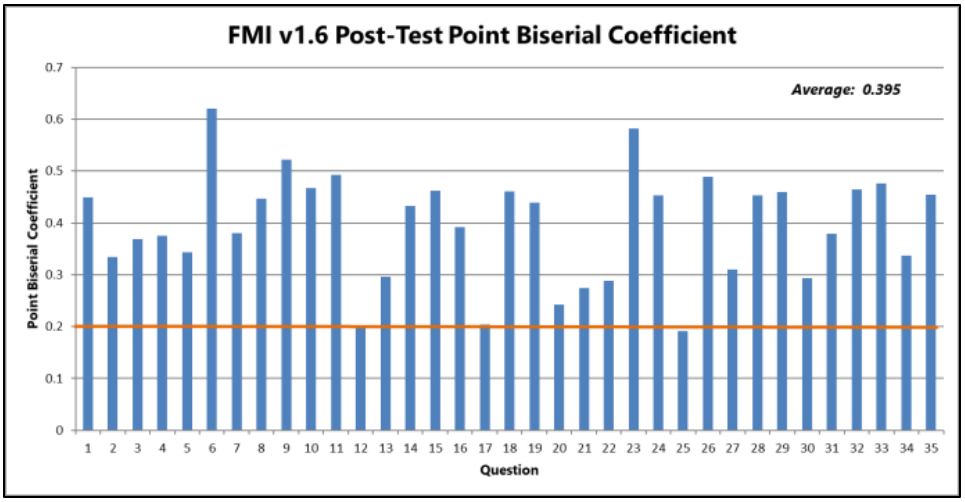

Figure 3: The FMI post-test point biserial coefficient for each question, with the ideal threshold highlighted in orange.

where k is the number of test items, σ_t^2 is the total test variance, and p_i is the difficulty index for item i. To ensure that the exam is sufficiently consistent to be a reliable measurement of understanding, the KR-20 metric should be in excess of 0.7. For the FMI post-test, the KR-20 value was calculated to be 0.78, indicating that the results are statistically consistent and reliable.

Ferguson's Delta is another common statistical measure that quantifies the discrimination of the test as a whole; different pairs of student scores are compared and the ratio between the number of unequal pairs of scores and the maximum number of pairs that the test can produce is calculated. The formula is given as

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - \frac{N^2}{K+1}}$$

where N is the number of students in the sample, K is the number of test items, and f_i is the frequency (number of occurrence) of cases at each score. An acceptable value for δ is at least 0.9; for the FMI post-test, $\delta = 0.97$, indicating a test that discriminates very well between students.

4. 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. In Fall 2017, the FMI was administered to 252 introductory meteorology students across the country to help assess the reliability and validity of the exam. A number of statistical measures of reliability and validity were calculated for individual test items and for the exam as a whole. Nearly every test item met the desired reliability and discriminatory thresholds, and the FMI as a whole was found to discriminate well between students.

These preliminary validity and reliability assessments are encouraging, though a more thorough statistical analysis is needed to further confirm the results. For example, not every introductory meteorology course will necessarily cover the content associated with each individual question; based on instructor surveys, additional item analyses will be conducted that remove students who were not exposed to material related to specific questions. We would also like to compare FMI performance for different student demographics, institution types, and geographic locations to assess the extent of potential biases. Similar statistical assessments are needed for the FMI pre-test scores as well, along with correlations with overall course performance to determine the predictive capability of the FMI. Finally, once the exam is fully validated, it will be utilized to identify student misconceptions so that the meteorological community can work to address these issues.

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, 2015: Motivation for and Development of a Standardized Introductory Meteorology Assessment Exam. *Bulletin of the American Meteorological Society*, **96**, 305—312.

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.

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

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.

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*.

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.

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.