Using Data Mining to Predict K–12 Students’ Performance on Large-Scale Assessment Items Related to Energy

Xiufeng Liu, Miguel E. Ruiz

State University of New York at Buffalo, Buffalo, New York

Received 13 July 2006; Accepted 15 July 2007

Abstract: This article reports a study on using data mining to predict K–12 students’ competence levels on test items related to energy. Data sources are the 1995 Third International Mathematics and Science Study (TIMSS), 1999 TIMSS-Repeat, 2003 Trend in International Mathematics and Science Study (TIMSS), and the National Assessment of Educational Progress (NAEP). Student population performances, that is, percentages correct, are the object of prediction. Two data mining algorithms, C4.5 and M5, are used to construct a decision tree and a linear function to predict students’ performance levels. A combination of factors related to content, context, and cognitive demand of items and to students’ grade levels are found to predict student population performances on test items. Cognitive demands have the most significant contribution to the prediction. The decision tree and linear function agree with each other on predictions. We end the article by discussing implications of findings for future science content standard development and energy concept teaching.

Keywords: physics; general science; statistics/multivariate

For better or worse, science education standards are currently driving U.S. science education reforms. The most influential national standards in the United States are the Benchmarks for Science Literacy (American Association for the Advancement of Science [AAAS], 1993) and National Science Education Standards (National Research Council [NRC], 1996). After the release of the above national standards in science, and particularly because of the No Child Left Behind Act (NCLB, 2001), all states must have science learning standards and the accompanying state standardized tests by 2007. The movement toward standards has also been taking place in other countries such as Canada (Council of Ministers of Education of Canada [CMEC], 1997) and UK (Department for Education and Employment [DfEE], 1999). Countries like China and Singapore that have traditionally maintained national science content standards have also initiated major revisions to their national standards (Curriculum Planning & Development Division [CPDD], 2000; Wei & Thomas, 2005). The role of standards in today’s science education worldwide is so great that the validity of standards cannot simply be assumed.

Correspondence to: X. Liu; E-mail: xliu5@buffalo.edu
DOI 10.1002/tea.20232
Published online 13 December 2007 in Wiley InterScience (www.interscience.wiley.com).

© 2007 Wiley Periodicals, Inc.
There can be many types of science education standards. For example, the *National Science Education Standards* (NRC, 1996) contain science content standards, science teaching standards, and science assessment standards, to name a few. This study is concerned about science content standards. Two issues are particularly important in content standards: (a) the sequence of content, and (b) the expectation of students mastering the content. The sequence defines the learning progression of students from K to 12, and the mastery expectation defines students’ attainment levels on different types of learning content. Central to the above two issues is students’ competence. Competence is concerned with what students are capable of performing. Competence in this study is defined as the performance level of a student population (e.g., fourth grade) as a whole on a national assessment. The reason to focus on student populations on national assessments is because a national science content standard is intended for all students in the nation, although different local and subpopulations may perform differently—higher or lower than the national norm. Without knowing the national norm of students’ performances, any decision on the sequence of science content and expected student mastery of science content would be arbitrary.

The research base for developing the current national science content standards is largely derived from qualitative studies that involve small and convenient samples (please refer to the research base discussed in NRC [1996] and AAAS [1993]). Ongoing validation of science content standards using nationally representative samples is desirable. Consistent with the recommendation in the *National Science Education Standards* (NRC, 1996), a national research council committee recently also recommends that science standards (content and achievement) should be reviewed and updated at least every 10 years to maintain their validity (NRC, 2006).

In this study, we focus on one unified concept—energy that is commonly included in national science content standards. The critical role of unified concepts for organizing content standards has been well documented (AAAS, 2000; Bybee, 1998, 2003). This critical role is to ensure that curriculums to be derived from the content standards are coherent. A coherent curriculum is one that is “articulated over time as a sequence of topics and performances consistent with the logical and, if appropriate, hierarchical nature of the disciplinary content from which the subject-matter derives” (Schmidt, Wang, & McKnight, 2005, p. 528). According to Schmidt et al. (2005), a curriculum is coherent when some foundational topics are introduced in earlier grades and then continue to be developed in later grades, a feature that is evident in the curriculums of top-performing countries on the Third International Math and Science Studies (TIMSS) but lacking in that of the United States. Ornstein and Hunkins (1998) summarize that a coherent curriculum contains the following features: (a) continuity: major ideas and skills are continuously practiced and developed; (b) integration: all types of knowledge and experiences are unified at a higher level; (c) articulation: various aspects of the curriculum are interrelated; and (d) balance: appropriate weights are given to each aspect of the curriculum. Using unified concepts, such as energy, is an effective way to organize topics into a coherent curriculum, because they provide not only a possible developmental progression for students to develop individual concepts but also a thematic network to help students make connections among the concepts (AAAS, 2000).

Developmentally appropriate progression is central to a coherent curriculum. Early studies on students’ conceptions of energy focused on a limited range of ages or grade levels and employed primarily qualitative research methods (e.g., Boyes & Stanisstreet, 1990; Solomon, 1983a, 1983b; Watts, 1983; Watts & Gilbert, 1983). More recent studies investigated quantitatively the progression of energy concept development from K–12 (e.g., Liu & McKeough, 2005; Liu & Tang, 2004). Driver, Squires, Rushworth, and Wood-Robinson (1994) suggested that students’ conceptions of energy progress through the following sequence: (a) awareness of personal energeticness, (b) extending energeticness to other living things, (c) awareness of nonliving things.
spontaneously being able to do things, (d) extending energeticness to some nonliving things that possess energy, (e) awareness of stored energy in elastic materials, (f) awareness of gravitational potential energy, (g) being able to tell the energy story, (i.e., describing events in energy terms), (h) awareness of energy conservation (i.e., describing events in quantitative terms), and (i) awareness of energy degradation (i.e., recognition that things run down) and efficiency. Based on Driver et al.’s (1994) suggestion and other findings in the literature, Liu and his colleagues (Liu & Collard, 2005; Liu & McKeough, 2005) developed and tested an hypothesis of progression on students’ energy concept development from elementary to high school that is characterized by the following distinct stages: (a) perceiving energy as activities or abilities to do work (abbreviated as Activity/Work), (b) identifying different energy sources and forms (abbreviated as Form/Source), (c) understanding the nature and processes of energy transfer (abbreviated as Transfer), (d) recognizing energy degradation (abbreviated as Degradation), and (e) realizing energy conservation (abbreviated as Conservation). The above stages are hierarchical, that is, a higher stage subsumes all lower stages. In terms of students’ competence levels along the above sequence, Liu and his colleagues (Liu & Collard, 2005; Liu & McKeough, 2005) further hypothesized and tested that students between ages 7 and 9 (around grade 3) could reach stage Activity/Work, students between ages 9 and 11 (around grade 5) could reach stage Form/Source, students between ages 11 and 13 (around grade 7) could reach stage Transfer, and students between ages 13 and 15 (around grade 9) could reach stage Degradation, and students between ages 15 and 19 (high school) could reach stage Conservation. The distinct stages in energy concept development were also found in a most recent study of student performance assessment data involving energy transfer by Dawson-Tunik (2006).

A limitation in the above studies (e.g., Liu & Collard, 2005; Liu & McKeough, 2005; Liu & Tang, 2004) is that energy is treated as a unidimensional concept (i.e., content type only) that consists of different topics independent of learning context and cognitive reasoning processes. Teaching and learning the energy concept involves many other factors, both psychological and sociocultural. The traditional view is that a content-free psychological structure, for example, logicomathematical reasoning, determines how and what kind conceptual understanding students can develop at different ages (Piaget, 1960; Piaget & Inhelder, 1969). Recent research in science education and developmental psychology has shown that students’ science concept development is also content specific (Carey, 1985, 1999; Driver & Easley, 1978; Metz, 1995). For example, Carey (1999) identified the following sources for conceptual development: (a) domain-general supports such as students’ fundamental type-of/kind-of and causal reasoning—a subset of logico-mathematical reasoning, (b) domain-specific supports such as students’ preconceptions and fundamental understanding in core domains (e.g., number, contact causality, intentional causality, personal psychology, intuitive mechanics, etc.), and (c) other cognitive processes such as analogical mapping and coherence networking.

Given the complexity of content and learning processes described above, student learning competence depends on many factors. Novice and expert research in the 1970s and 1980s clearly show that experts’ knowledge is organized around important ideas such as unified concepts (Chi, Feltovich, & Glaser, 1981; Chi, Glaser, & Rees, 1982). The process of developing expertise including higher levels of cognitive structures is more than an individual effort; it requires active participation and acceptance of learners into the learning organization and environment (Lave & Wenger, 1991). The sociocultural aspects associated with learning processes and outcomes also play an important role in cognitive development (Vygotsky, 1978, 1986). Therefore, student performances on test items may be affected by many factors including content, context, and cognitive demand of items as well as student developmental stages.

Journal of Research in Science Teaching
Implications of the above nature of the energy concept development for a coherent curriculum on energy are many. First, the energy concept may appear again and again in the science curriculum at different grades, but the teaching of the energy concept may not take place at the same cognitive demand. Following the revised Bloom’s taxonomy can be one approach to teaching the energy concept progressively toward more complex and integrative understanding (Anderson & Krathwohl, 2001). For example, students may simply memorize that light and electricity are forms of energy (remembering) by fourth grade, but be able to explain how light energy is transferred to electric energy in a solar device (understanding) by eighth grade, and finally be able to construct a model house powered by sunlight (creating) by 12th grade. Second, any content related to energy may be taught using everyday examples or imaginary or hypothetical scenarios. Helping students make the transition from reasoning in an everyday context to reasoning in an abstract context is an important objective of a coherent curriculum on energy. Third, energy is a type of conceptual knowledge; it can be conceptualized to consist of five hierarchical subtopics of knowledge reviewed earlier. Because of its comprehensiveness, different subtopics of the energy concept, that is, energy activity, energy sources, energy transfer, energy degradation, and energy conservation, may be introduced at different grade levels. Fourth, energy teaching following a developmentally coherent curriculum must take students’ prior knowledge and experiences into consideration so that learning becomes meaningful to students (Mintzes & Wandersee, 1998). The above requirements for a coherent curriculum, that is, cognitive demands, subtopics, and learner background, plus grade levels, interact in science teaching to effect different student competences in understanding of the energy concept. To date, there has been no report in the literature on how the above four factors are related to student performances on assessment items related to energy at the national population level—a gap that the present study intends to fill.

Using large-scale assessment data of nationally representative student samples, we attempt to answer the following research question: Can K–12 students’ performance on items related to energy be predicted? If yes, what are the significant predictors?

Method

Data Sources

Data for the present study came from the following national and international assessments: the National Assessment of Educational Progress (NAEP), the 1995 Third International Mathematics and Science Study (TIMSS)—the U.S. national sample—the 1999 Third International Mathematics and Science Study—Repeat (TIMSS-R)—the U.S. National sample, and the 2003 Trend in International Mathematics and Science Study (TIMSS)—the U.S. national sample. The above assessment programs involved U.S. nationally representative samples of students in both private and public schools; they provided accurate estimations of students’ science achievements for three populations—population 1 (grades 3 and 4), population 2 (grades 7 and 8), and population 3 (grades 10, 11, 12, and advanced science course takers).

Public items released on the Web sites of the TIMSS, TIMSS-R (http://timss.bc.edu), and NAEP (http://nces.ed.gov/nationsreportcard/itmrls/) were reviewed and items related to energy were identified. Although the United States also participated in the Program for International Student Assessment (PISA) in which science achievement was also assessed, the science assessment items for PISA were not released for public use; thus, we were not able to make use of PISA assessment data. In total, 39 items were identified, among which 23 were from 1995 TIMSS, 5 from 1999 TIMSS, 3 from 2003 TIMSS, 6 from 2005 NAEP, and 2 from 2000 NAEP. Although different students answered the above questions on different assessments, because we used...
weighted percentages correct, student performance levels on the above questions were comparable at the population level.

The identified items were then classified by types of content, context, and cognitive demand. The types of content were in the following hierarchical categories following Liu and McKeough (2005): energy’s ability to do work or being responsible for various activities (Activity/Work), energy sources and forms (Form/Source), energy transfer (Transfer), energy degradation (Degradation), and energy conservation (Conservation). The identified items were also classified as involving everyday context if an item deals with a phenomenon related to typical American students’ everyday experiences, or noneveryday context if an item deals with a phenomenon beyond typical American students’ everyday experiences. The types of cognitive demands were conceptual understanding, and reasoning/investigating.

Table 1 shows the correspondence between the above two cognitive demands used in this study and the cognitive demands TIMSS and NAEP used to classify items. Description of the TIMSS cognitive demands is available in Martin and Kelly (1996), Mullis, Martin, Smith, Garden, Gregory, Gonzalez, Chrostowski, and O’Connor (2003), and description of NAEP cognitive demands is available from the National Assessment Governing Board (NAGB) (http://nces.ed.gov/nationsreportcard/about/nagb/). Table 2 presents the distribution of items by content, context, cognitive demand, and grade.

As an example of classification, the following question is from 1995 TIMSS:

G6. By what process do most stars release energy?
A. Electromagnetic induction resulting from strong magnetic fields
B. Rapid rotation of the star
C. Radioactivity in the interior of the star
D. Nuclear fusion in the interior of the star
E. Heat that was stored when the star was “born”

This question was classified as conceptual understanding by TIMSS. We further classified it as noneveryday context because none of the students would have any experience with this phenomenon. The item was also considered as being related to energy transfer, because it asks about a process of energy conversion.

In this study, we were interested in population norms of student performances on individual items. We defined a population norm as the weighted percentage correct at a population level. Because students who participated in the TIMSS and NAEP studies had different chances of being selected due to stratified random sampling and oversampling of some minority groups, sampling weights must be used in computing the average percentage correct of an item. A weight was the total number of students in the population represented by the student in the sample; a weighted

Table 1
Correspondence among types of cognitive demands

<table>
<thead>
<tr>
<th>Category in Present Study</th>
<th>TIMSS 1995 and 1999</th>
<th>TIMSS 2003</th>
<th>NAEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding</td>
<td>Understanding</td>
<td>Factual knowledge</td>
<td>Conceptual understanding</td>
</tr>
<tr>
<td></td>
<td>Using tools, routine procedures and science processes</td>
<td>Conceptual understanding</td>
<td></td>
</tr>
<tr>
<td>Reasoning and investigating</td>
<td>Theorizing, analyzing, and solving problems</td>
<td>Reasoning and analysis</td>
<td>Scientific investigation</td>
</tr>
<tr>
<td></td>
<td>Investigating the natural world</td>
<td>Investigating the natural world</td>
<td>Practical reasoning</td>
</tr>
</tbody>
</table>
percentage was the percentage of all students in the population. We used house weights to calculate the weighted percentage correct for each item. Weighted percentages produced an unbiased estimate of population performances. The total weights, which corresponded to population sizes, varied from item to item. Because the total weights were so large (from a minimum of 100,000 to over 1 million), and the assignment of students to different items was random, we considered that the weighted percentages correct represented the competence levels of the targeted US student populations (e.g., all fourth-grade students, eighth-grade students, etc.). For the NAEP items, weighted percentages correct were reported along with the released items on the NAEP website. For constructed response questions, percentages correct were percentages of completely correct responses, that is, partial correct was considered incorrect. Because it was typical in large-scale assessment that a same item was given to more than one targeted population (e.g., fourth and eighth graders), the above-identified 39 items generated 76 weighted percentages correct for different populations.

Although different students answered different items as the result of the matrix design in TIMSS and NAEP, because we were not interested in individual students’ total test scores and their variances, we did not use specialized procedures such as Jackknife variance estimation to compute variances; that is, no variances for students or items were computed in the present study.

Data Mining

We use data mining to analyze data. Data mining is also called knowledge discovery in databases (KDD) (Han & Kamber, 2001; Witten & Frank, 2005). Developed from database management systems technology and conventional statistics, data mining goes beyond retrieving, analyzing and representing information in databases; it focuses particularly on uncovering hidden patterns in large data sets. Today, data mining involves not only database and statistics, but also machine learning, information science, and visualization. It is being applied in sciences (e.g., bioinformatics), business, internet security, and many other fields.

Data mining performs two functions: one is to identify regularities among data records (e.g., concept cluster, concept comparison, and discrimination), another to find relations among variables in the data that will predict unknown or future values of the variables. Unlike descriptive and inferential statistical analyses that rely on means and standard deviations, data mining uses both logical and mathematical (deterministic, and parametric and nonparametric statistical)
reasoning to analyze data records. Data mining usually involves creating a data warehouse in a particular structure so that special data mining query language or algorithms can be applied to answer researchers’ questions. Data mining is both a bottom-up and top-down approach to discovering patterns. An example of a bottom-up approach is market basket analysis, which generates association rules to identify frequent patterns or correlations in sales data to help merchants display certain products in clusters. An example of top-down approach is using a predefined hierarchy to generate frequent patterns (i.e., we could group students according to grade levels such as elementary, middle, and high school). Top-down data mining is useful in problems involving a large number of variables (or dimensions) but a relatively small number of cases. In this case, top-down data mining can group variables that correlate to reduce the dimensionality so that the new variables will have a larger set of cases to support the discovery of patterns (Liu, Han, Xin, & Shao, 2006).

Specifically, data mining includes the following steps:

1. **Creating a data warehouse**: this involves creating a common metadata schema, data cleaning and transformation, and mapping of missing and unknown values. To create this data warehouse we would need to define a common integrated metadata schema and then map each of the data resources to this new integrated schema to create the data warehouse. For this study we combined the data from two sources: that is, TIMSS and NAEP. Because TIMSS and NAEP used different survey instruments, it was necessary to develop a new common scheme so that data collected in both TIMSS and NAEP could be integrated. We used the weighted percentages correct at different grade levels (i.e., grades 3, 4, 7, 8, 10, 11, 12, and advanced physics course takers) on individual items to integrate TIMSS and NAEP data. Using weighted percentages correct at the population level, we avoided dealing with different raw data structures of TIMSS and NAEP, and instead we dealt with TIMSS and NAEP data at the population level. Due to the small number of items, in the present study we did not employ a usually complex data warehouse structure; instead we used a 76 \times 8 table to perform the data warehouse function, with the rows for the 76 weighted percentages correct, and eight columns for item, content, context, cognitive demand, grade, percentage-correct, competence (i.e., satisfactory and unsatisfactory, to be described later), and data source.

2. **Data reduction and projection**: this involves building data cubes to reduce the total number of variables and to find invariant representations for the data that would facilitate the application of data mining algorithms. Data cubes are not necessarily three dimensional; in fact, they can be n-dimensional depending on the number of variables involved. In this study, we used item characteristics, that is, content type, application context, cognitive demand, and grade levels to create data cubes. For example, the sample item on stars releasing energy discussed earlier was described as conceptual understanding, noneveryday context, and being related to energy transfer. Because each item was also associated with a weighted percentage correct at a population level (e.g., fourth grade), each data cube was described by the following five dimensions: performance (weighted percentage-correct), content, context, cognitive demand, and grade level. All the 76 weighted percentages in the present study were described by these same five dimensions, resulting in 76 data cubes.

3. **Data mining**: this involves the application of pattern matching algorithms that will create a representation of the data to understand the relations among variables, or generate models to predict future values. In this study we used a decision tree and a linear function to find possible patterns in data cubes. The specific data mining algorithms will be described in the next section.

4. **Interpretation and visualization**: this involves representing the discovered patterns and creating the best visualization of them. For this study, a tree format was used in the
present study to present the decision tree, and a linear function was used to present relations among the variables.

Figure 1 presents the overall data mining process.

**Classification Algorithms**

Two classification algorithms, C4.5 and M5, were used in this study, one for predicting nominal classes (i.e., satisfactory and unsatisfactory), and another for predicting numerical values (i.e., percentages correct). An algorithm is a sequence of commands a computer executed to perform a task—classification. We used the C4.5 classification algorithm (Quinlan, 1993) to build a decision tree for predicting student performances. Due to the small number of data cubes, various schemes for classifying student performances based on percentages correct, such as no-understanding (≤25%), beginning (25–50%), developing (50–75%), and mastery (≥75%), were tried; the best prediction was made when student performance levels were classified into two categories—satisfactory and unsatisfactory—with 55% as a cutoff value. Student performances were considered satisfactory if the weighted percentage correct was > 55%; student performances were considered unsatisfactory if the weighted percentage correct was equal to or <55%.

*Journal of Research in Science Teaching*
C4.5 constructed a decision tree where the internal nodes were the set of attributes, and the leaves the classes to be learned or predicted. The set of attributes were content (activity/work, form/sources, transfer, degradation, and conservation), context (everyday and noneveryday), cognitive demand (understanding, reasoning/investigating), and grade level (elementary, middle school, high school). The classes to be learned were satisfactory and unsatisfactory. The C4.5 built the decision tree by selecting a best attribute to place at the root and placing every other value into a branch. This process continued until all the attributes were placed in the tree. The criteria used for selection of the best attributes were based on measuring the information gain of the set of attributes for yielding the largest information gain to maximize the prediction rate. Finally, the C4.5 algorithm used a method to prune the tree into a more efficient tree in order to minimize the prediction error.

The M5 algorithm (Quinlan, 1992) was used to build a linear function to predict student percentages correct. The process was similar to building a decision tree by C4.5, except that a function was constructed to predict the numerical value for each of the branches. Although C4.5 built branches of a decision tree by splitting and grouping attributes to maximize the prediction rates, M5 built a function based on numerical values from various branches of a decision tree to minimize the variation of class values within each branch.

**Evaluation of Accuracy**

There are a number of measures that can be used for estimating the accuracy of a classification algorithm. The most commonly used measures are Recall, Precision, F-measure, and Kappa. To compute these measures we first define a $2 \times 2$ contingency table representing the possible outcomes of the classification: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications. TP is when an instance belonging to a class is classified into that class; TN is when an instance not belonging to a class is classified into other classes. TP and TN are correct classifications; the higher the rates, the better. FP is when an instance not belonging to a class is classified into that class; FN is when an instance belonging to a class is classified into other classes. FP and FN are errors; the lower the rates, the better. Recall, also known as sensitivity, is defined as the proportion of instances of the class that are correctly identified by the algorithm. Using the contingency table, Recall is computed as follows: $\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$.

Precision, also known as a positive predictive value, is the proportion of correct classifications achieved by the algorithm: $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$.

A derived measure, called F-measure, is used to measure the overall accuracy of classification. The F-measure corresponds to the harmonic mean of recall and precision. The F-measure is defined as: $\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})} = \frac{2 \times \text{TP}}{(2 \times \text{TP} + \text{FP} + \text{FN})}$.

The F-measure ranges from 0 to 1; the higher the value of the F-measure, the better. Another statistic, called Kappa, is also used to measure the agreement between predicted and observed classification by correcting for agreement by chance. Kappa is defined as: $\text{Kappa} = \frac{\text{Pr}(\text{agreement}) - \text{Pr}(\text{chance})}{1 - \text{Pr}(\text{chance})}$, where Pr(\text{agreement}) represent the probability of agreement, and Pr(\text{chance}) is the probability that the agreement occurs by chance. The values of Kappa range from $-1$ to $1$. A value of 1 represents perfect agreement between the method and the ground truth.

In addition to the above statistics for describing the classification accuracy, a confusion matrix may also be used to indicate the number of instances that are correctly classified and the number of instances that are incorrectly classified. Although the F-measure gives overall accuracy rate, the
confusion matrix indicates exactly how many and what instances are correctly classified, and how many and what instances are incorrectly classified.

For numerical prediction using algorithm M5, the commonly used accuracy measures are multiple correlation coefficient $R$, the squared multiple correlation coefficients ($R^2$), and the absolute error of prediction. $R$ and $R^2$ are interpreted in the same way as in regression. The higher the $R$ and $R^2$ and the lower the absolute error of prediction, the better.

**Computer Program**

We used Weka 3.4.6, a free Java-based tool that offers a collection of machine learning algorithms for data mining (Weka is available at http://www.cs.waikato.ac.nz/ml/weka/). Weka is currently one of most popular open source data mining systems.

The standard 10-fold crossvalidation was used (Witten & Frank, 2005). The 10-fold crossvalidation method divided the whole set of data cubes randomly into roughly 10 equal subsets of cubes, and the computer program selected a combination of nine subsets of data cubes to learn the rules that could be used to build a decision tree. This decision tree was tested on the remaining one subset of data cubes to evaluate the accuracy of the predictions. The computer program continued the process using all possible combinations of nine subsets of data cubes. The results obtained were averaged over the 10-fold and contained estimates of the classification effectiveness of the model in terms of F1, recall, precision, accuracy, and Cohen’s Kappa statistics.

**Results**

**Decision Tree for Predicting Competence**

The average rate of correct classification was 76.3%, or 58 out of 76 instances, and the Kappa statistics was 0.52. Table 3 presents the breakdown of Recall and Precision, as well as the F-measure. It can be seen that both satisfactory and unsatisfactory classes had overall good classification accuracy (Recall 0.79, Precision 0.66, and F-measure 0.72 for satisfactory, and Recall 0.75, Precision 0.85, and F-Measure 0.80 for unsatisfactory).

A confusion matrix indicates exact numbers of correctly and incorrectly classified instances. Results showed that, among the 29 satisfactory instances, 23 were correctly classified as satisfactory, 6 were misclassified as unsatisfactory. Among the 47 unsatisfactory instances, 35 were correctly classified as unsatisfactory, and 12 were misclassified as satisfactory.

Figure 2 presents the decision tree and classification rules. From Figure 2 we see that the cognitive demand of assessment items was the first deciding factor. If cognitive demand of items was reasoning and investigation, student competence would be unsatisfactory no matter what content and context of the items and what grade of students. If the cognitive demand of items was understanding, then context of the items would decide students’ competence. If the context of items was noneveryday, then student performance was unsatisfactory no matter what content of test items and what grade of students. If the context of items was everyday, then the student grade
would decide the competence. If students were at middle or high school (above seventh grade), then student competence would always be satisfactory no matter what content of test items. For elementary grades (third and fourth), different content types would decide student competence. If the content was about energy’s ability to do work, then student competence was satisfactory; if the content was about energy forms or sources, energy transfer, energy degradation, and energy conservation, then student competence would be unsatisfactory.

Figure 2 also presents actual numbers of correctly and incorrectly classified instances in the parentheses on the tree leaves. Note that the branches in the decision tree were built based on a full set of data cubes; the total number of 76 instances is obtained by adding the first number in each of the leaf nodes. We can see that the first leaf closest to the root was decided by the cognitive demand. There were 13 instances on the first leaf, and all of them were classified correctly (the first number refers to correct classification instances and the second number refers to incorrect classification instances). Similarly, for the second leaf branching from context, 14 instances were classified as unsatisfactory, and one of the instances was incorrectly classified. Examining the numbers for all leaves, we see that overall, the correctly classified instances significantly outnumbered the incorrectly classified instances.

**Linear Function**

The following linear function was obtained through M5:

\[
\text{Percentage Correct} = 120.5 - (11.1 \times \text{content}) - (27.8 \times \text{context}) - (39.0 \times \text{cognitive}) \\
+ (17.5 \times \text{population})
\]
where

\[
\text{Content} = \{1 - \text{Activity/Work}, 2 - \text{Form/Source}, 3 - \text{Transfer}, 4 - \text{Degradation}, 5 - \text{Conservation}\}
\]

\[
\text{Context} = \{1 - \text{Everyday}, 2 - \text{Noneveryday}\}
\]

\[
\text{Cognitive} = \{1 - \text{Understanding}, 2 - \text{Reasoning/Investigation}\}
\]

\[
\text{Population} = \{1 - \text{Elementary}, 2 - \text{MiddleSchool}, 3 - \text{HighSchool}\}
\]

From the above function, we can see the following relationships:

1. There is a negative relationship \((r = -0.530, n = 76, p < .000)\) between percentage correct and content level; as content increases from 1 (activity/work) to 5 (conservation), the percentage correct decreases.

2. There is a negative relationship \((r = -0.283, n = 76, p = .013)\) between percentage correct and context; as context increases from 1 (everyday context) to 2 (noneveryday context), the percentage correct decreases.

3. There is a negative relationship between percentage correct and cognitive demand \((r = -0.587, n = 76, p < .000)\). As the cognitive demand increases from 1 (understanding) to 2 (reasoning/investigation), the percentage correct decreases.

4. There is a positive relationship between percentage correct and population level, despite the fact that the bivariate correlation between the two is not statistically significant \((r = -0.228, n = 76, p = .51)\). As the population level increases from 1 (elementary) to 3 (high school), the percentage correct increases.

From the above function, we also see that the cognitive demand of items had the most impact on the percentage correct; an increase in the cognitive demand from understanding to reasoning/investigation would reduce the percentage correct by 39%. The next most significant factor determining the percentage correct was the context of test items; changing item context from everyday to noneveryday would reduce the percentage correct by 27.8%. Content had least significant impact on the percentage correct; an increase in content level (e.g., from energy form to energy transfer) would reduce percentage correct by 11.1%. On the other hand, increase in the population level (e.g., from elementary school to middle school) would increase percentage correct by 17.5%. Therefore, it was the combined effect of the above factors, cognitive demand, context, content, and grade level, that would decide student performance on test items.

The above function had an average multiple correlation coefficient of 0.78 \((R = 0.78)\) over 10-fold computations. That is, the function could account for 60.8% of variance in the data cubes \((R^2 = 0.608)\). The mean absolute error of prediction was 11.6%.

Using the above linear function, meaningful prediction of students’ competence can be made. For example, in elementary school, energy is usually taught as activity or work (e.g., attributing change to energy, energy does work); in middle school, energy is usually taught as transfer (e.g., energy can be transferred from one form to another); and in high school, energy is usually taught as conservation (e.g., energy is conserved during transfers). If the above energy content is taught in everyday context, and with the understanding cognitive demand, the predicted percentages correct would be 60.1% (elementary), 55.9% (middle school), 53.7% (high school). Two of the percentages are at the satisfactory level and one is unsatisfactory, indicating that expectations of
elementary and middle school students are reasonable, but those of high school students are not. As another example, if energy is taught to middle school students with an expectation that they develop conceptual understanding of energy degradation in an everyday context, then the predicted percentage correct is only 44.3%, indicating that the expectation is beyond students’ competence.

Table 4 presents different scenarios in which energy may be taught in K–12. We can see that by the end of elementary school (i.e., grade 4), students’ performances can be expected to be satisfactory if teaching is at conceptual understanding, in everyday context, and on energy doing work. By the end of middle school (i.e., grade 8), students’ performances can be expected to be satisfactory if teaching is at conceptual understanding, in everyday context, and on energy doing work, or energy form/source, or energy transfer. By the end of high school (i.e., grade 12), students can be expected to demonstrate satisfactory performances in the following types of learning: (a) conceptual understanding, everyday context, and all types of energy content; (b) conceptual understanding, noneveryday context, and energy doing work, energy form/sources; and (c) reasoning/investigation, everyday context, and energy doing work.

Comparing the predicted competence levels in Table 4 with the predicted competence levels in Figure 2, we see that most predictions are consistent with each other with a few exceptions (in italic in Table 4). The following predictions are different based on the function and on the decision tree:

1. The decision tree predicted that when the cognitive demand was beyond conceptual understanding, that is, reasoning and investigation, then student performance would be unsatisfactory no matter what the context, content and grade level were. However, the function predicted the same as the above with one exception, that is high school students’ performances could be satisfactory (56%) if the cognitive demand was reasoning/investigating, in everyday context, and the energy content was about energy doing work.
2. The decision tree predicted that when the cognitive demand was at conceptual understanding but the context was noneveryday context, then student performance would be unsatisfactory no matter what content and grade level were. However, the function predicted the same as the above with the exception that high school students’ performances would be satisfactory if energy content was about energy doing work or energy form/source.
3. The decision tree predicted that middle school student performance would be satisfactory if energy teaching was at conceptual understanding, in everyday context, and no matter what energy content was. However, the function predicted that in the same conditions as the above, if energy content was about energy degradation and conservation, then middle school student performance would be unsatisfactory.

The above discrepancy accounts for 5 out of 60 predictions, thus the prediction agreement between the decision tree and function was 92% (55 of 60).

Discussion

This study aimed at answering the following questions: can K–12 students’ performance on national assessment items related to energy be predicted? If yes, what are the significant predictors?

Results from this study show that the answer to the above research question is positive. In the application of two data mining algorithms (decision tree and linear function), all four factors, that is, energy content, context, cognitive demand, and grade level, contribute to predicting students’
Table 4
*Predicted performance levels based on characteristics of test items and student grade levels*

<table>
<thead>
<tr>
<th>Cognitive Demand</th>
<th>Context</th>
<th>Population</th>
<th>Content</th>
<th>Predicted Percentage Correct</th>
<th>Predicted Competence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding</td>
<td>Everyday Elementary</td>
<td>A/W</td>
<td>60.1</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>49.0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>37.9</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>26.8</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>15.7</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>A/W</td>
<td>77.6</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>66.5</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>55.4</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>44.3</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>33.2</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>A/W</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>100</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>94.6</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>83.5</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>72.4</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noneveryday Elementary</td>
<td>A/W</td>
<td>32.3</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>21.2</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>10.1</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>A/W</td>
<td>49.8</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>38.7</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>27.6</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>16.5</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>5.4</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>A/W</td>
<td>67.3</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>56.2</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>45.1</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>34.0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>22.9</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>Reasoning/investigating</td>
<td>Elementary</td>
<td>A/W</td>
<td>21.1</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>10.0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>A/W</td>
<td>38.6</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>27.5</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>16.4</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>5.3</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>A/W</td>
<td>56.4</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>45.3</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>34.2</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>23.1</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>12.0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noneveryday Elementary</td>
<td>A/W</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>F/S</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0</td>
<td>U</td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
The predicted competence levels of different student populations on different types of test items (i.e., content, context, and cognitive demand) are consistent with previous research. For example, unsatisfactory understanding of energy conservation by high school and university engineering students, as well as preservice science teachers has been reported (Kruger, Palacio, & Summers, 1992; Liu, Ebenezer, & Fraser, 2002; Liu & Tang, 2004; Trumper, 1997). Findings from the present study show that even high school students may not be expected to reach a satisfactory performance level on energy conservation even if items are at the understanding cognitive demand and is introduced in everyday context. Findings from this study may also provide an explanation to the mixed results on students’ competence levels of energy concept development in previous research as summarized in Liu and McKeough (2005).

Why are students’ performance levels overall disappointing? Although findings from this study cannot suggest any causal relationship, they do suggest factors that relate to student performances on test items on energy. No single factor can determine student performance; it is the interaction among factors that accounts for performance differences among individuals. Mayer (2003) developed a model to combine various developmental theories to explain individual differences in cognitive performances. He suggests that student knowledge on a specific domain, such as understanding a concept, is attributable to both ability, which is strongly related to age before high school, and deliberate practice (i.e., science teaching and learning experiences). Student performance on test items is directly related to student knowledge, but is also dependent on the nature of performance tasks (such as cognitive demands, context, etc.). Consistent with the above theory, our findings from this study suggest that teaching energy needs to be developmentally appropriate, as suggested by Driver et al. (1994), and Liu and McKeough (2005). Developmental appropriateness is one key requirement of a coherent curriculum (AAAS, 2000; Schmidt et al., 2005).

Our findings suggest that cognitive demand is the most significant predictor to students’ performances on energy, thus, it is important to help students to reason with different cognitive demands for a same topic. Wiggins and McTighe (1998) call for an approach called uncovering,
that is, bringing to the surface misconceptions, the subtle, the nonobvious, the problematic, the controversial, the obscure, and the missing and lost. For example, there can be six facets of understanding of a same topic; they are facet 1: explanation—sophisticated and apt explanations and theories to provide knowledgeable and justified accounts of events, actions, and ideas; facet 2: interpretation—interpretations, narratives, and translations to provide meaning; facet 3: application—ability to use knowledge effectively in new situations and diverse contexts; facet 4: perspective—critical and insightful point views; facet 5: empathy—the ability to get inside another person’s feelings and worldview; and facet 6: self-knowledge—the wisdom to know one’s ignorance and how one’s patterns of thought and action inform as well as prejudice understanding.

Our findings also suggest that relevant student experiences matter because context of test items is the second significant predictor of student performances. Thus, teaching energy needs also expose students to different experiences, such as both the life—world experiences and abstract symbolic reasoning as suggested by Solomon (1983a). Following a world-wide movement toward Science-Technology-Society (STS) in the 1970s and 1980s, there has been recently renewed calls for context-based science teaching and learning (e.g., Gilbert, 2006; Pilot & Bulte, 2006).

The coherence of learning outcomes related to energy in the national content standards needs also to be evident, and the competence models developed in this study could help in this regard. For example, the competence models could inform future decisions on expectations of student achievement to increase the specificity of content standards. To increase the specificity of content standards in stating the expected student achievement levels, it is necessary to give consideration to the effects of cognitive demands and context of test questions or problems we expect students to solve. Many state content standards are still too general; they contain basically propositional statements about scientific theories and principles, a situation considered to be unacceptable by NCLB (NRC, 2006). Although it is unrealistic for any content standards to be as specific as an industrial standard, that is, engineering standards, for science content standards as standard, it is highly desirable for them to be specific to avoid open interpretation that may undermine its authority and enforceability. We believe there is much to be improved in the current U.S. national science content standards in this regard. The need for more specificity is demonstrated by recent curriculum evaluation studies conducted by Project 2061 (Kesidou & Roseman, 2002; Stern, 2003; Stern & Ahlgren, 2002). In those evaluation studies of curriculum materials and assessments, Project 2061 staff reinterpreted the learning outcomes in the national science education standards to develop operational sets of criteria. Disagreement with Project 2061’s reinterpretation of some learning outcomes (e.g., KMT) has been raised (e.g., Shiland, 2003).

The above issues of coherence and specificity of science content standards raise a question about the nature of science content standards. To increase the coherence of a content standard, there should be a companion standard to define expected student achievement on how good is good enough. Ravitch (1996) argues that there should be three types of standards in education: (1) content standards specifying what teachers should teach and what students should learn, (2) performance standards defining degrees of mastery or levels of attainment by students, and (3) opportunity-to-learn or school delivery standards defining the availability of programs, staff, and other resources schools, districts, and states must provide for teachers to teach and students to learn for achieving the performance standards. Ravitch (1996) argues that the three types of standards are closely interrelated, and if any is omitted, the remaining standards become meaningless. The NCLB also differentiates content standards from achievement or attainment/performance standards; it requires that every state develop both content standard and achievement standard. Although all states now have content standards, many still do not have achievement standards (NRC, 2006).
In the present study, we defined competence based on student population performance on test items, that is, > 55% correct as satisfactory and ≤ 55% as unsatisfactory. More levels of competence, such as meeting standards ≥ 65% and meeting standards with distinction (≥ 85%) as used in New York State regents exams, may also be studied in future research when a large number of data cubes become available. Defining satisfactory or setting attainment levels is an issue about standard setting, a topic beyond the present study. Interested readers may reference Cizek and Bunch (2007).

The above implications need to be taken with an awareness of the limitations of the methodology used in this study. Despite the reasonably satisfactory accuracy of data mining in predicting student performances, the classifications and predictions are not perfect. Still, 40% variance is unexplained by the linear function, and between 20% and 25% instances are misclassified by the decision tree, which is reflected in the recall values of 80% and 75%. Thus, although the four factors investigated in this study all contribute to the prediction of student competence levels, there still exist other factors contributing to students’ competence levels. Other potential factors may include students’ ability level, gender, social–economic status, teachers’ competence, school resource, etc. We believe the power of this new methodology is to conduct exploratory studies to sort through a potentially large number of variables that may contribute to differences in student performances, and identify a limited few that are most significant. Follow-up confirmatory studies such as experimental intervention studies or contextual studies such as ethnographies in the classroom are needed to further test or understand the identified patterns. We consider this study to be only one in a continuum aiming at better understanding and improving student understanding of the energy concept.

The data mining approach is, of course, applicable to studies of other science concepts. We have shown how data mining can be a promising methodology to analyze nationally representative student assessment data to develop a clear and valid research base to inform the revision of national science content standards and classroom instruction of unified science concepts. The more data sources, assessment items, and science concepts are analyzed simultaneously, the more powerful data mining can be in discovering the hidden patterns about student competence; there is no limit for data mining as to numbers of data sources, assessment items, or concepts to be analyzed at the same time.

References


