Bayesian Networks for Skill Diagnosis and Model Validation

Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA

Hua Wei

April 2014

ALWAYS LEARNING

PEARSON

Abstract

Domain models depict relationships among the important knowledge and skills that students are expected to learn in a subject domain. Even in a highly focused and well-defined content area, there may be different domain models capturing different assumptions, perspectives, and hypotheses regarding what students should learn and how they learn it. Domain models are based on expert opinions or theories, and rarely validated empirically. The framework of Bayesian networks is appropriate to the task of modeling relationships between subskills and making inferences about the modeled subskills with empirical data. This study illustrated how Bayesian networks can be used to diagnose specific subskills in the domain of fractions, and how different domain models compare with each other in terms of the extent to which they accurately predict students' mastery of different subskills

Keywords: Bayesian network, domain model, cognitive diagnosis

Bayesian Networks for Skill Diagnosis and Model Validation

Domain models depict relationships among the important knowledge and skills that students are expected to learn in a subject domain (Mislevy & Riconscente, 2006). The relationships represented in a domain model synthesize teachers, researchers, and domain experts' beliefs and theories about how learning evolves in the domain. Domain models have important implications for curriculum, instruction, and assessment in the domain.

Fraction is one of the most often discussed topics in mathematics education. It is also one of the most difficult topics to learn and teach in primary mathematics curriculum (Suydam, 1978; Wu, 1999). Many students are found not proficient at problem solving involving fractions, and the lack of proficiency has been reported among students across a wide range of grade levels and ages, including middle school students (Brown & Quinn, 2006), high school students (National Center for Educational Statistics, 2000), college students (Hanson & Hogan, 2000; Larson & Choroszy, 1985), and adults (Steen, 2007).

Research studies have been done to investigate what difficulties students experience when they respond to questions that require application of fraction concepts and operations on fractions, what may have caused the difficulties, and what types of errors students typically make in their responses. For example, Hanson and Hogan (2000) reported that students struggling with problems involving fractions were not familiar with the process of finding common denominators. Brown and Quinn (2006) examined student responses on a fraction test, and identified a number of common and unique errors students made that resulted from misconceptions or lack of computational fluency related to the subject of fractions.

Clearly, appropriate instructional practices need to be in place to develop and enhance students' proficiency in the domain of fractions. However, even in a highly focused and well-

researched content area such as fraction arithmetic, there are different domain models reflecting different assumptions, perspectives, and hypotheses regarding what students should learn and how they should learn it. In this study, we analyzed two mathematics curricula designed for college students, and found that each embodied a domain model with a different representation of interconnections among the subskills in the domain of fractions.

Domain models are based on expert opinions or theories, and rarely validated empirically (Chi & Koeske, 1983; Wang, 2005). The framework of Bayesian networks seems appropriate to the task of modeling relationships between subskills and making inferences about the modeled subskills with empirical data. The efficiency of Bayesian networks in making probability-based inferences for the purpose of cognitive diagnosis has been documented in the literature (Mislevy, 1995; Mislevy, Almond, Yan, & Steinberg, 1999). Bayesian networks have been successfully applied in various domains, such as physics problem solving (VanLehn & Martin, 1998), mixed-number subtraction (Mislevy, 1995), multicolumn subtraction (Lee & Corter, 2011), and proportional reasoning (Beland & Mislevy, 1996), to assess subskills and competencies, which are otherwise challenging to model using traditional measurement methodologies. The goal of this paper is to evaluate how Bayesian networks can be used to diagnose specific subskills in the domain of fractions, and how different domain models compare with each other in terms of the extent to which they accurately predict students' mastery of different subskills.

Method

Data

Data used in this study originated from an online homework, tutorial, and assessment system used by millions of students from thousands of colleges and universities in the United States. The system works by allowing teachers to create online assignments using

algorithmically generated exercises and maintain records of all student work. Students receive assignments and respond to exercises in their own pace and with extra help from a variety of multimedia resources. The online system accompanies a large number of textbooks that are commonly used in college mathematics classes of all levels, ranging from remedial and developmental mathematics to advanced calculus. The exercises generated from the system are either taken from the textbooks or adapted versions of exercises in the textbooks.

In this study, we selected 38 multiple-choice exercises administered in the online system. These exercises were aligned to a developmental mathematics curriculum, and measured students' abilities to perform basic arithmetic operations with fractions, with each exercise assessing one specific subskill. We selected student responses to these exercises obtained from an "assessment" environment, in which certain constraints were placed with regard to how and when students must respond to the exercises. For example, multimedia learning aids were not available in a test context whereas they were accessible in the homework environment. Because the students came from different classrooms and were taught by different instructors, each student only responded to a subset of the 38 exercises, and differed in terms of which and how many exercises they responded to. We also selected a certain time period during which students interacted with the exercises. As a result, in the data matrix that we worked with in the study, there were a total of 406 students and 38 exercises, with missing values in many cells.

Domain Models

The two domain models being compared in this study are informed by the analysis results of two prerequisite structures that each accompanies a mathematics curriculum for college students. Both models indicate that there are five important subskills to learn in the domain of fraction arithmetic. The five subskills are the elementary skill working with fractions, including

converting integers, mixed numbers or decimals to fractions, identifying numerators and denominators, writing a fraction in simplest form, and finding the prime factorization of a fraction, and the procedural skills of adding, subtracting, multiplying, and dividing fractions. Compared with the four procedural skills which are about operations with fractions, the elementary skill is more about understanding fraction notions and terminology (such as denominator, numerator, and least common multiple), and relationships among fractions and other forms of numbers. Another commonality of the two domain models is that they both specify the dependence relationship of procedural skills on the learning of the elementary skill. This is consistent with Bezuk's (1989) recommendation that operations on fractions should be taught after students have had a firm grasp of the concepts and ideas of the order and equivalence of fractions.

However, the two domain models differ in terms of the interconnections among the five subskills. Model 1 suggests that the four procedural skills are dependent on the elementary skill but are conditionally independent of one another, given the elementary skill. Model 2, on the other hand, suggests that addition and multiplication with fractions are dependent on the elementary skill, while subtraction and division with fractions are dependent on addition and multiplication, respectively. A graphical representation of the two domain models is shown in Figure 1.

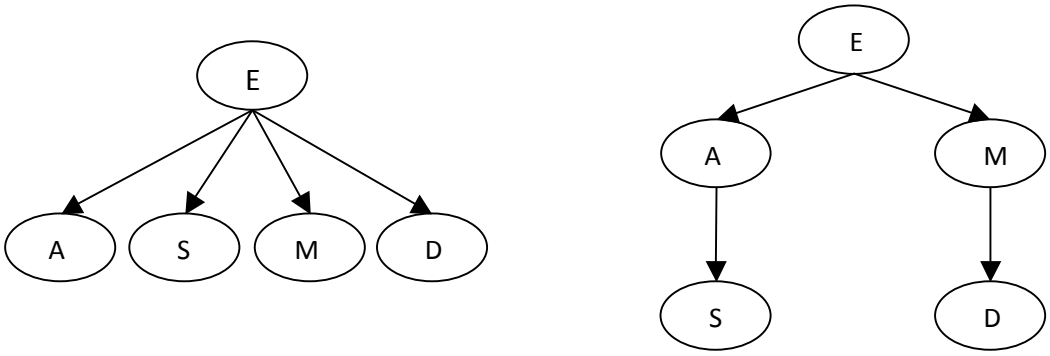

Figure 1. Two Domain Models

Notes. 1. Nodes in each model represent the different subskills in the domain of fractions. 2. E represents the elementary skill. A, S, M, and D represent the procedural skills of adding, subtracting, multiplying, and dividing fractions, respectively.

Overview of Bayesian Networks

Bayesian networks (BNs), also known as belief networks are used to represent knowledge about an uncertain domain. Nodes represent random (often discrete) variables, while edges between nodes represent probabilistic dependencies among them. The structure of a Bayesian network represents conditional independence relationships. Specifically, each variable is independent of its non-descendents given the state of its parents. This property is used to reduce the number of parameters required to define the joint probability distribution of the variables, and to compute the posterior probabilities in an efficient way, given data. The quantitative parameters of a Bayesian network are expressed as the conditional probability distribution of each node depending only on its parents. For discrete random variables, this conditional probability is described with conditional probability tables (CPTs) where the local probability that a variable takes on each of a finite number of values is dependent on each

combination of values of its parents. The joint distribution of a collection of variables can be determined uniquely by these local CPTs.

Once the structure and the parameter values of the conditional probabilities are known, a Bayesian network can be used to make inferences about the modeled variables. New evidence is entered into the network in the form of specific values for certain variables (e.g., observable variables like test items), and the output consists of updated posterior probability distributions for the set of nodes of interest (e.g., unobservable variables like knowledge and skills in a content domain). Updated posterior probabilities can be obtained through Bayes' theorem as:

Posterior Probability = α Likelihood × Prior Probability, or

$$P(H_i|e) = \frac{P(e|H_i)}{P(e)} = \alpha L(H_i|e_1, e_2, e_3, \ldots, e_n)P(H_i),$$

where e consists of the set of evidence $e_1, e_2, e_3, \ldots, e_n$, H_i is the hypothesis of interest, and α is the normalizing constant.

The most widely used algorithm for performing the belief updating in Bayesian networks is the message-passing algorithm (Pearl, 1988). However, even with the efficient formula for computing posterior probabilities when a set of evidence is introduced into the network, the message-passing algorithm still requires complex computations. Currently, a number of computer programs can be used to implement message propagation calculations, including First Bayes (www.tonyohagan.co.uk/1b/), HUGIN (www.hugin.com), MSBNx (research.microsoft.com/adapt/MSBNx/), WinBUGS (www.mrc-bsu.cam.ac.uk/bugs/), GeNIe (genie.sis.pitt.edu/about.html), and Netica (www.norsys.com). In this study, GeNIe was used to construct the Bayesian networks and perform skill diagnosis and model comparison by calculating posterior probabilities.

Network Representation

In this study, diagnosis of the subskills with fraction arithmetic was made by using student performance data on a number of exercises in the test context. Specifically, each subskill was measured by seven to nine exercises, and student responses to these exercises were scored as either correct or incorrect. Two networks, each based on one domain model as shown in Figure 1, were constructed in accordance with the following assumptions: (1) One and only one subskill is a direct cause of each exercise; (2) Given the state (mastery vs. non-mastery) of a subskill, all the exercises that measure that subskill are mutually independent. The resulting two networks are shown in Figures 2 and 3.

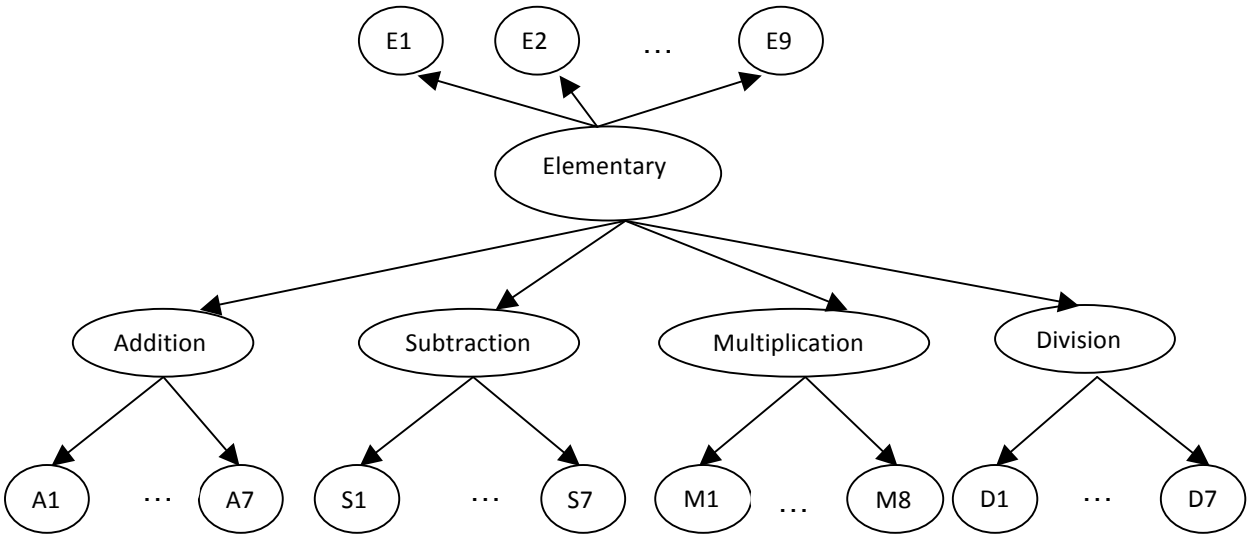

Figure 2. Network 1

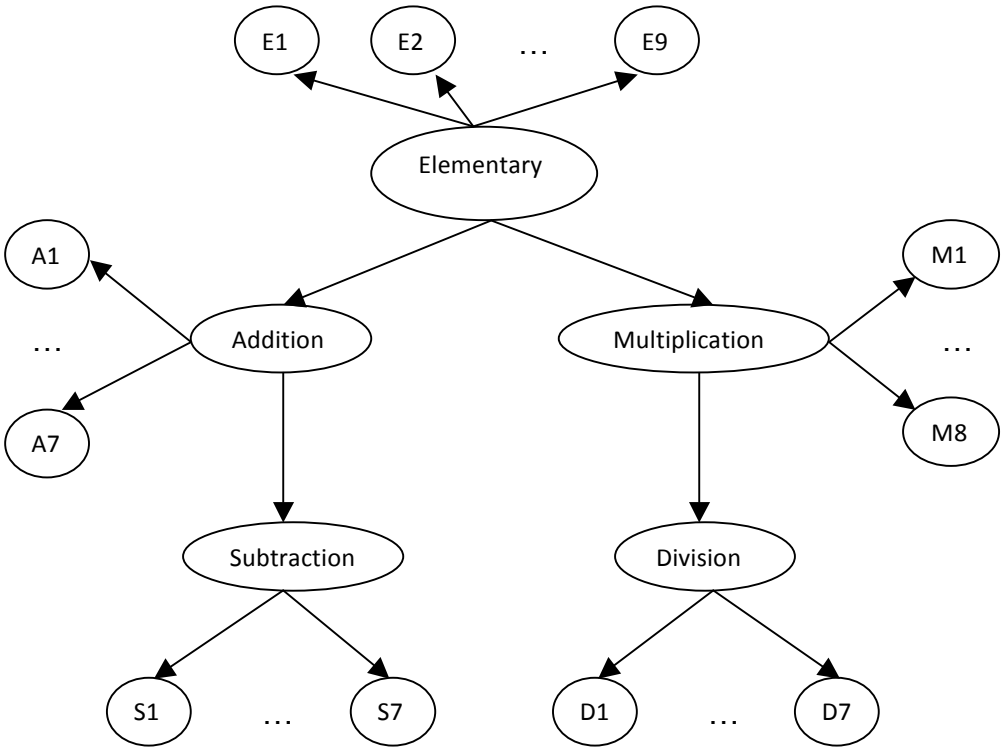

Figure 3. Network 2

In this study, the structure of each network was considered known, and the goal of learning was to determine the values of the network parameters with observed data through the maximum likelihood estimation method. Evaluation of each network was based on how accurately it diagnosed the five subskills among the students.

The entire student data were randomly split up into two sets: a calibration sample of 325 (80%) students, and a validation sample of 81 (20%) students. The calibration sample was used to estimate the parameters for each network, including marginal or conditional probability distributions for each of the five subskills. Empirical data for the calibration sample included each student's responses (correct, incorrect, or missing) to the 38 exercises plus his/her "observed" status on each subskill, which was obtained based on a pre-specified criterion. Specifically, a student would be determined as having mastered a subskill if he/she correctly

answered 50% or more of the exercises assessing that subskill; otherwise, his/her status on that subskill would be non-mastery. The network parameters were estimated in GeNIe.

Once the network parameters were estimated, the validation sample data were loaded into GeNIe to obtain posterior probabilities. The vector of item responses for each student in the validation sample was entered to produce the posterior probability of that student having mastered each of the five subskills. Different cut-off points were applied to the obtained posterior probabilities to predict the mastery or non-mastery status of each student on each subskill. The predicted mastery/non-mastery status of each subskill for each student in the validation sample was then compared to the "observed" mastery/non-mastery status to get the correct classification rate. Fixed cutoff points of 0.50, 0.60, and 0.70 were tried, along with a scheme using variable-specific cutoff points. For Network 1, these cut points were 0.87, 0.80, 0.79, 0.74, and 0.75 for the five subskills, respectively. For Network 2, the cut points were 0.87, 0.80, 0.82, 0.74, and 0.79. The two networks were compared in terms of the rate of correct classification and global model fit, indexed by Akaike's information criterion (AIC) and the Bayesian information criterion (BIC).

Results

By applying the criterion of correct response rate of 50% or greater, the "observed" mastery rates among students in the validation sample for the five subskills were 0.83, 0.78, 0.75, 0.73, and 0.80, respectively. The predicted mastery rates for the validation sample produced by the two networks are shown in Table 1.

Table 1

Observed and Predicted Subskill Mastery Rates in the Validation Sample

	Observed	Network 1	Network 2
Cut point = 0.50			
Elementary	0.83	0.88	0.88
Addition	0.78	0.85	0.81
Subtraction	0.75	0.83	0.79
Multiplication	0.73	0.77	0.75
Division	0.80	0.84	0.81
Cut point = 0.60			
Elementary	0.83	0.85	0.84
Addition	0.78	0.80	0.79
Subtraction	0.75	0.80	0.78
Multiplication	0.73	0.73	0.74
Division	0.80	0.79	0.79
Cut point = 0.70			
Elementary	0.83	0.80	0.80
Addition	0.78	0.75	0.78
Subtraction	0.75	0.78	0.75
Multiplication	0.73	0.72	0.70
Division	0.80	0.75	0.74
Variable-specific cut			
Elementary	0.83	0.75	0.75
Addition	0.78	0.70	0.72
Subtraction	0.75	0.65	0.67

Multiplication	0.73	0.68	0.69
Division	0.80	0.73	0.73

As shown in Table 1, the differences between observed and predicted subskill mastery rates were small for each network, no matter which cutoff point was used. This is saying that each network predicted the observed subskill mastery status reasonably well. A comparison between the two networks shows that Network 2 seemed to be able to recover the mastery rates for all five subskills in the validation sample slightly better than Network 1. In addition, a comparison across the various cutoff points indicates that the cutoff point of 0.60 resulted in the closest match between observed and predicted mastery rates for both networks.

Correct classification rates are the percentage agreement between predicted and observed subskill mastery status. They are calculated as the percentage of students who have actually mastered a subskill and are predicted having mastered the subskill plus the percentage of students who have not mastered a subskill and are predicted not having mastered the subskill. The correct classification rates for both networks using the various cutoff points are presented in Table 2.

Table 2

Correct Classification Rates in the Validation Sample

Cut points	0.50	0.60	0.70	Variable-specific
Network 1				
Elementary	73 (0.90)	73 (0.90)	75 (0.93)	73 (0.90)
Addition	69 (0.85)	65 (0.80)	63 (0.78)	63 (0.78)
Subtraction	73 (0.90)	71 (0.88)	69 (0.85)	73 (0.90)
Multiplication	76 (0.94)	77 (0.95)	76 (0.94)	73 (0.90)
Division	78 (0.96)	80 (0.99)	77 (0.95)	75 (0.93)
Network 2				
Elementary	73 (0.90)	74 (0.91)	75 (0.93)	73 (0.90)
Addition	62 (0.77)	64 (0.79)	65 (0.80)	64 (0.79)
Subtraction	68 (0.84)	69 (0.85)	67 (0.83)	72 (0.89)
Multiplication	77 (0.95)	78 (0.96)	75 (0.93)	74 (0.91)
Division	78 (0.96)	78 (0.96)	76 (0.94)	75 (0.93)

Results in Table 2 show that the correct classification rates were high across different subskills, ranging from 0.77 to 0.99. In other words, no matter which network or cutoff point was used, the five subskills were correctly diagnosed most of the times. A comparison across different cutoff points did not lead to a single cut point with consistently better classification rates than the others. In addition, no consistent evidence was reported that favors one network over the other. Both networks were able to diagnose the multiplication, division, and elementary subskills better than they did the addition and subtraction skills.

In addition to correct classification rates, the two networks can be compared in terms of global model fit. Model fit indices, such as AIC and BIC, take into account the extent of model fit as well as degree of model complexity. The fit statistics calculated for the two networks are shown in Table 3.

Table 3

Model Fit Statistics

	Network 1	Network 2
Log-likelihood	-3550.33	-3496.82
AIC	-3635.33	-3581.82
BIC	-3796.14	-3742.63

The fit statistics for the two networks, including log-likelihood, AIC, and BIC are reported in the form without multiplying the constant of –2 used by Akaike. In other words, a lower absolute value indicates better fit. All three statistics point to Network 2 as a preferred model.

Discussion and Conclusions

This study provided an example of how Bayesian networks could be used for diagnosis of multiple subskills and validation of different domain models. The two domain models being compared were abstractions from analysis of the domain of fraction arithmetic, and captured different conceptualizations about how the important subskills in the domain were dependent on each other. Based on the two domain models, two Bayesian networks were built and compared with each other in terms of how well the subskill mastery rates were recovered in the student sample, how frequently each subskill was correctly diagnosed among individual students, and how well the model fit the data.

Results of the study showed that the two networks were able to recover the subskill mastery rates and predict the mastery status of each subskill among individual students moderately well. Although the two networks were nearly as effective as each other in making correct diagnosis of individual subskills, the second network showed a minimal advantage over the first one judged by recovery of subkskill mastery rates and fit statistics. In the second network, the subskills of addition and multiplication are each dependent on the elementary subskill, while the subskills of subtraction and division are dependent on addition and multiplication, respectively. This is consistent with the hypothesis behind many elementary mathematics curricula in which the notions and terminologies of fractions are first introduced, and then sections on addition/subtraction and multiplication/division follow.

Bayesian networks can be considered a member in the family of cognitive diagnostic models that make probability-based inferences about students' specific knowledge structures and processing skills. Appropriate use of these cognitive models in educational assessments has the potential of diagnosing students' strengths and weaknesses and identifying content understanding and misconceptions. Cognitive information generated from students' task performance in these assessments has significant implications for guiding instruction and enhances individualized learning.

References

Beland, A., & Mislevy, R. J. (1996). Probability-based inference in a domain of proportional reasoning tasks. *Journal of Educational Measurement, 33,* 3-27.

Bezuk, N., & Cramer, K. (1989). Teaching about fractions: What, when, and how? In P. Trafton (Ed.), *National Council of Teachers of Mathematics 1989 Yearbook: New Directions For*

Elementary School Mathematics (pp. 156-167). Reston, VA: National Council of Teachers of Mathematics.

Brown, G., & Quinn, R. (2006). Algebra students' difficulty with fractions: An error analysis. *Australian Mathematics Teacher*, *62*(4), 28–40.

Chi, M., & Koeske, R. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology, 19*(1), 29–39.

Hanson, S., & Hogan, T. (2000). Computational estimation skill of college students. *Journal for Research in Mathematics Education*, *31*, 483–499.

Larson, C., & Choroszy, M. (1985). *Elementary education majors' performance on a basic mathematics test*. Retrieved from http://www.eric.ed.gov/.

Lee, J., & Corter, J. E. (2011). Diagnosis of subtraction bugs using Bayesian networks. *Applied Psychological Measurement, 35*(1), 27-47.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Mahwah, NJ: Erlbaum.

Mislevy, R.J., Almond, R.G., Yan, D., & Steinberg, L.S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K.B. Laskey & H.Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (437-446). San Francisco: Morgan Kaufmann.

Mislevy, R.J., & Riconscente, M.M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Erlbaum.

National Center for Educational Statistics (2000, August). *NAEP 1999 trends in academic progress: Three decades of student performance*. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main1999/2000469.pdf.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* San Mateo, CA: Morgan Kaufmann.

Steen, L. (2007). How mathematics counts. *Educational Leadership*, *65*(3), 8–14.

Suydam, M. N. (1978). Review of recent research related the concepts of fractions and of school learning. *Journal of Special Education, 9*(3), 281-291.

VanLehn, K., & Martin, J. (1998). Evaluation of an assessment system based on Bayesian student modeling. *International Journal of Artificial Intelligence in Education, 8,* 179-221.

Wang, Y. (2005). A GA-based methodology to determine an optimal curriculum for schools. *Expert Systems with Applications, 28*(1), 163–174.

Wu, H. (1999, October). *Some remarks on the teaching of tractions in elementary school*. University of California, Berkeley. Retrieved from http://www.math.berkeley.edu/~wu/fractions2.pdf.