Evaluating Problem Difficulty Rankings Using Sparse Student Response Data

Ari BADER-NATAL [1], Jordan POLLACK
DEMO Lab, Brandeis University

Abstract. Problem difficulty estimates play important roles in a wide variety of educational systems, including determining the sequence of problems presented to students and the interpretation of the resulting responses. The accuracy of these metrics are therefore important, as they can determine the relevance of an educational experience. For systems that record large quantities of raw data, these observations can be used to test the predictive accuracy of an existing difficulty metric. In this paper, we examine how well one rigorously developed – but potentially outdated – difficulty scale for American-English spelling fits the data collected from seventeen thousand students using our SpellBEE peer-tutoring system. We then attempt to construct alternate metrics that use collected data to achieve a better fit. The domain-independent techniques presented here are applicable when the matrix of available student-response data is sparsely populated or non-randomly sampled. We find that while the original metric fits the data relatively well, the data-driven metrics provide approximately 10% improvement in predictive accuracy. Using these techniques, a difficulty metric can be periodically or continuously recalibrated to ensure the relevance of the educational experience for the student.

1. Introduction

Estimates of student proficiency and problem difficulty play central roles in Item Response Theory (IRT) [11]. Several current educational systems make use of this theory, including our own BEEweb peer-tutoring activities [2,8,9,13]. IRT-based analysis often focuses on estimating student proficiency in the task domain, but the challenge of estimating problem difficulty should not be overlooked. While student proficiency estimates can inform assessment, problem difficulty estimates can be used to refine instruction: these metrics can affect the selection and ordering of problems posed and can influence the interpretation of the resulting responses [6]. It is therefore important to choose a good difficulty metric initially and to periodically evaluate the accuracy of a chosen metric with respect to available student data. In this paper, we examine how accurately one rigorously developed – but potentially outdated – difficulty scale for the domain of American-English spelling predicts the data collected from students using our SpellBEE system [1]. The defining challenge in providing this assessment lies in the nature of the data. As SpellBEE is a peer-tutoring system, the challenges posed to students are determined by other students, resulting in data that is neither random nor complete. In this

[1] Correspondence to: Ari Bader-Natal, Brandeis University, Computer Science Department – MS 018. Waltham, MA 02454. USA. Tel.: +1 781 736 3366; Fax: +1 781 736 2741; E-mail: ari@cs.brandeis.edu.

work, we rely on a pairwise comparison technique designed to be robust to data with these characteristics. After assessing the relevance of this existing metric (in terms of predictive accuracy), we will examine some related techniques for initially constructing a difficulty metric based on non-random, incomplete samples of observed student data.

2. American-English spelling: A sample task domain

The educational system examined here, SpellBEE, was designed to address the task domain of American-English spelling [1]. SpellBEE is the oldest of a growing suite of web-based reciprocal tutoring systems using the Teacher's Dilemma as a motivational mechanism [2]. For the purposes of this paper, however, the mechanisms for motivation and interaction can be ignored, and the SpellBEE system and the difficulty metric used by it can be specifically re-characterized for an educational data mining audience.

2.1. Relevant characteristics of the SpellBEE system

Students access SpellBEE online at SpellBEE.org from their homes or schools. As of May 2007, over 17,000 students have actively participated. After creating a user account, a student is able to log in, choose a partner, and begin the activity.[2] During the activity, students take turns posing and solving spelling problems. When posing a problem, the student selects from a short list of words randomly drawn from the database of word-challenges. This database is comprised of 3,129 words drawn from Greene's New Iowa Spelling Scale (NISS), which will be discussed in the next section [12].[3] When responding to a problem, the student types in what they believe to be the correct spelling of the challenge word. The accuracy of the response is assessed to be either correct or incorrect. Figure 1 presents a list of the relevant data stored in the SpellBEE server logs.

To date, we have observed over 64,000 unique (case-insensitive) responses to the challenges posed,[4] distributed across over 22,000 completed games consisting of seven questions attempted per student. Student participation, measured in games completed, has not been uniform, however. Of the challenges in the space, most students have only attempted a very small fraction. In fact, when examining the response matrix of every student by every challenge, less than 1% of the matrix data is known. An important characteristic of the SpellBEE data, then, is that the response matrix is remarkably sparse. Given that the students acting as tutors are able to – and systemically motivated to – express their preferences and hunches through the problems that they select, another important characteristic of the SpellBEE data is that the data present in the student-challenge response matrix is also biased. The effects of this bias can be found in the following example: 16% of student attempts to spell the word "file" were correct, while 66% of attempts to spell the word "official" were correct. The average grade level among the first set of students was 3.9, while for the second set it was 6.4. In Section 3.2 we

[2] In the newer BEEweb activities, if no one else is present, a student can practice alone on problems randomly drawn from the database of challenges posed in the past.

[3] In SpellBEE, the word-challenges are presented in the context of a sentence, and so of the words in Greene's list, we only use those found in the seven public-domain books that we parsed for sentences.

[4] Of these, 17,391 were observed more than once. In this paper, we restrict the set of responses that we consider to this subset. See Footnote 7 for the rationale behind this.

Figure 1. The SpellBEE server logs data about each turn taken by each student, as shown in the first list. The data in the first list is sufficient to generate the data included in the second list.

1. `time` : a time-stamp allows responses to be ordered
2. `game` : identifies the game in which this turn occurred
3. `tutor` : identifies the student acting as the tutor in this turn
4. `tutee` : identifies the student acting as the tutee in this turn
5. `challenge` : identifies the challenge posed by the tutor
6. `response` : identifies the response offered by the tutee

1. `difficulty` : the difficulty rating of the challenge posed by the tutor
2. `accuracy` : the accuracy rating of the response offered by the tutee

will present techniques designed to draw more meaningful difficulty information from this type of data.

2.2. Origin, use, and application of the problem difficulty metric

When trying to define a measure of problem difficulty for the well-studied domain of American-English spelling, we were able to benefit from earlier research in the field. Greene's "New Iowa Spelling Scale" provides a rich source of data on word spelling difficulty, drawn from a vast study published in 1954. Greene first developed a methodology for selecting words for his list (5,507 were eventually used.) Approximately 230,000 students from 8,800 classrooms (grades 2 through 8) around the United States participated in the study, totally over 23 million spelling responses [12]. From these, Greene calculated the percentage of correct responses for each word for each grade. This table of success rates is used in SpellBEE to calculate the difficulty of each spelling problem for students, whose grade level is known.

3. Techniques for assessing relative challenge difficulty

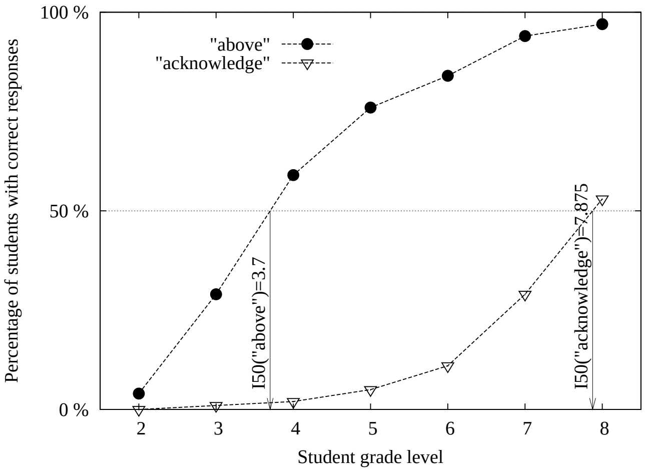
The research questions addressed in this paper focus on the fit of the difficulty model based on the NISS data to the observed SpellBEE student data. Two different techniques are involved in the calculating this fit. The first converts the graded NISS data to a linear scale. The second identifies from the observed student data a difficulty ordering over pairs of problems, in a manner appropriate for a sparse and biased data matrix. Both will be employed to address the research questions in the following sections.

3.1. Linearly ordering challenges using the difficulty metric

Many subsequent studies have explored various aspects of Greene's study and the data that it produced. Cahen, Craun, and Johnson [5] and, later, Wilson and Bock [14] explore the degree to which various combinations of domain-specific predictors could account for Greene's data. Initially starting with 20 predictors, Wilson and Bock work down to a regression model with an adjusted R^2 value of 0.854.[5] Here, we not interested in

[5] The most influential of which being the length of the word.

Figure 2. Difficulty data for two words from the NISS study are plotted, and the I_{50} statistics are calculated.



predicting the NISS results, but instead are interested in assessing the fit (or predictive power) of the 1954 NISS results to observations made of students using SpellBEE over 50 years later. We will drawn upon one statistic used by Wilson and Bock: the one-dimensional flattening of the seven-graded NISS data. This statistic, which they refer to as the "location" of the word, is the (fractional) grade level at which 50% of the NISS students correctly spell word w.[6] We denote this as $I_{50}(w)$. Figure 2 illustrates how the graded difficulty data that is used to derive this statistic for two different words. The value of this statistic is that it provides a single grade-independent difficulty value for a word that can be compared directly to that of other words.

3.2. Identifying pairwise difficulty orderings using observed student data

Given the characteristics of the data collected from the SpellBEE system, identifying the more difficult of a pair of problems based on this data is not trivial. The percentage of correct responses to a challenge, the calculation used to generate the NISS data, is not appropriate here, as the assignment of challenges to students was done in a biased, non-random manner (recall the "file"/"official" example from Section 2.1.) Tutors, in fact, are motivated to base their challenge selection on the response accuracies that they anticipate. A more appropriate measure, rooted in several different literatures, is to assess pairwise problem difficulties on distinctions indirectly indicated by the students. In the statistics literature, McNemar's test provides a statistic based on this concept [10], in the IRT literature, this is used as a data reduction strategy for Rasch model parameter estimation [7], and the Machine Learning literature includes various approaches to learning

[6] Wilson and Bock calculate the 50% threshold based on a logistic model fit to the discrete grade-level data, while we calculate the threshold slightly differently, based on a linear interpolation of the grade-level data.

Table 1. While the I_{50} metric flattens the grade-specific NISS data to a single dimension, the relative difficulty ordering of most word-pairs based on the graded NISS data is the same as when based on the I_{50} scale. In this table, we quantify the amount of agreement between I_{50} and each set of grade-specific NISS data using Spearman's rank correlation coefficient. The strong correlations observed suggest that the unidimensional scale sufficiently captures the relative difficulty information from the original NISS dataset. (The number of words, N, varies by grade, as the NISS study did not show several of the harder words to the younger students.)

Grade	N	Spearman's ρ
2	2218	0.751
3	3059	0.933
4	3126	0.977
5	3129	0.974
6	3129	0.960
7	3129	0.935
8	3129	0.915

rankings based on pairwise preferences [4]. Assume that for some specific pair of problems, such as the spelling of the words “about” and “acknowledge”, we first identify all students in the SpellBEE database who have attempted both words. Given that response accuracy is dichotomous, there are only four possible configurations of a student's response accuracy to the pair of challenges. In the cases where the student responds to both correctly or incorrectly, no distinction is made between the pair. But in the cases where the student correctly responds to one but incorrectly to the other, we classify this as a distinction indicating a difficulty ordering between the two problems.[7]

It is also worth stating that in this study, we assume a “static student” model, so we are not concerned with the order of these two responses. At the cost of some data loss, one could instead assume a “learning student” model, for which only a correct response on one problem followed by an incorrect response on the other would define a distinction. Had the incorrect response been observed first, we could not rule out the possibility that the difference was due to a change in the student's abilities over time, and not necessarily an indication of difference in problem difficulties.[8]

An example may clarify. If counting the number of both directional distinctions made by all students (e.g. 12 students in SpellBEE spelled “about” correctly and “acknowledge” incorrectly, while 2 students spelled “about” incorrectly and “acknowledge” correctly), we have a strong indication of relative problem difficulty. McNemar's test assigns a significance to this pair of distinction counts. In this work, we more closely follow the IRT approach, relying only the relative size of the two counts (and not the significance.) Thus, since 12 distinctions were found in one direction and only 2 in the other, we say that we observed the word “about” to be easier than the word “acknowledge” based on collected SpellBEE student data. If distinctions were available for every problem pair,

[7] We recognize that some distinctions are spurious, for which the incorrect response was not reflective of the student's abilities. Here we take a simplistic approach of identifying and ignoring non-responses (in which the student typed nothing) and globally-unique responses (which no other student ever responded, to any challenge.) Globally-unique responses encompass responses from students who don't yet understand the activity, responses from students who did not hear the audio recording, responses from student attempting to use the response field as a chat interface, and responses from students making no effort to engage in the activity.

[8] Another possible model is a “dynamic student” model, for which student abilities may get better or worse over time. Under this model, no distinctions can be definitively attributed to difference in problem difficulty.

a total of 3,129 × 3,128 = 9,787,512 pairwise problem orderings could be expressed. In our collected data so far, we have 3,349,602 of these problem pairs for which we have distinctions recorded. In the subsequent sections, we measure the fitness of a predictive model (like I_{50}) based on how many of these pairwise orderings are satisfied.[9]

4. Assessing the fit of the NISS-based I_{50} model to the SpellBEE student data

Given the NISS-based I_{50} difficulty model of problem difficulty and the data-driven technique for turning observed distinctions recorded in the SpellBEE database into pairwise difficulty orderings, we can now explore various methods to assess the applicability of the model to the data.

4.1. Assessing fit with a regression model

The first method is to construct a regression model that uses I_{50} to predict observed difficulty. Since observed difficulty is currently available only in pairwise form, this requires an additional step in which we flatten these pairwise orderings into one total ordering over all problems. As this is a highly non-trivial step, the results should be interpreted tentatively. Here, we accomplish a flattening by calculating, for each challenge, the percentage of available pairwise orderings for which the given challenge was the more difficult of the pair. So if 100 pairwise orderings involve the challenge word "acknowledge", and 72 of these found "acknowledge" to be the harder of the pair, we would mark "acknowledge" as harder than 72% of other words. A regression model was then built on this, using I_{50} as a predictor of the pairwise-derived percentage. The model, after filtering out data points causing ceiling and floor effects (i.e. $I_{50}(w) = 2.0$ or $I_{50}(w) = 8.0$), had an adjusted R^2 value of 0.337 ($p < 0.001$ for the model). The corresponding scatterplot is shown in Figure 3.[10] The relatively low adjusted R^2 value is likely at least partially a result of the flattening step (rather than solely due to poor fit.) Had we flattened the data differently, this value would clearly change. In order to obtain a more reliable measure of model fitness, we seek to avoid any unnecessary processing of the mined data.

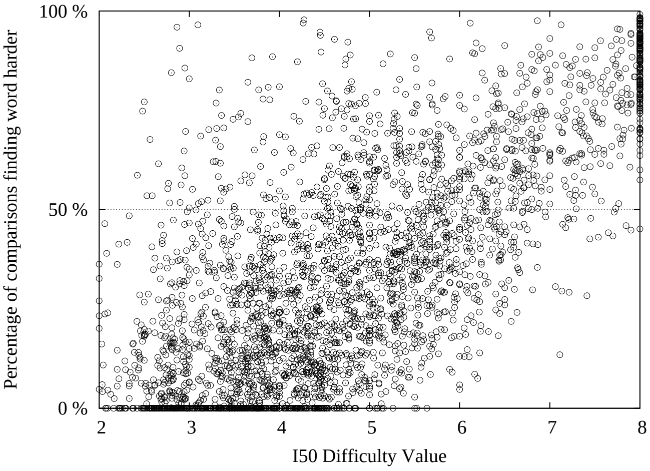
4.2. Assessing fit with as the percentage of agreements on pairwise difficulties

The second method that we explore provides a more direct comparison, without any further flattening of the student data. Here, we simply calculate the percentage of observed pairwise difficulty orderings (across all challenges) for which the I_{50} model correctly predicts the observed pairwise difficulty ordering. When we do this across all of the 3,349,602 difficulty orderings that we have constructed from the student data, we find that the I_{50} model correctly predicts 2,534,228 of these pairwise orderings, providing a 75.66% agreement with known pairwise orderings from the mined data. Remarkably, we found that the predictive accuracy of the I_{50} model did not significantly change as the

[9] Note that it is not be possible to achieve a 100% fit, as some cycles exist among these pairwise orderings.

[10] The outliers in this plot mark the problems that are ranked most differently by the two measures. The word "arithmetic", for example, was found to be difficult by SpellBEE students, but was not found to be particularly difficult for the students in the NISS study. Variations like this one may reflect changes in the teaching or in the frequency of usage since the NISS study was performed 50 years ago.

Figure 3. Words are plotted by their difficulty on the I_{50} scale and by the percentage of other words for which the observed pairwise orderings found the word to be the harder of the pair. An adjusted R^2 value of 0.490 was calculated for this model. (When ignoring the words affected by a ceiling or floor effect in either variable, the adjusted R^2 value drops to 0.377.)



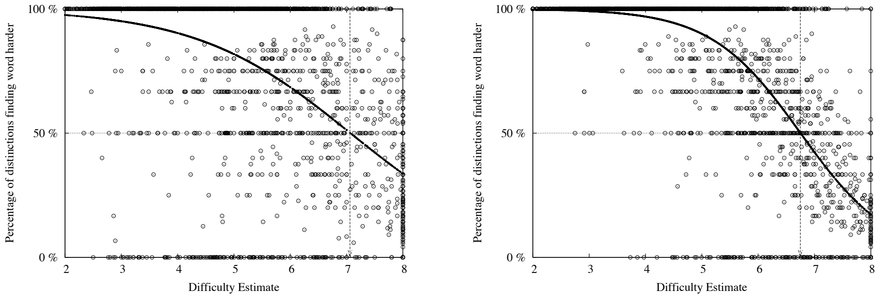
quantity of student data used for the distinction varied. 75.1% of predictions based on one distinction were accurate, while 74.7% of predictions based on 25 distinctions were accurate (intermediate values ranged from 71.0% to 77.6%). This flat relationship suggests that pairwise difficulty orderings constructed from a minimal amount of observed data may be just as accurate, in the aggregate, as those orderings constructed when additional data is available.

5. Incorporating SpellBEE student data in a revised difficulty model

We now know that there is a 75.66% agreement in pairwise difficulty orderings between the I_{50} difficulty metric derived from the NISS data and the observed pairwise preferences mined from the SpellBEE database. Can we improve upon this? We will present an approach that iteratively updates the I_{50} problem difficulty estimates using the mined data and a logistic regression model. Rather than producing a single predictive model, we construct one logistic model for each challenge, and use these fitted model to update our estimates of the problem difficulty. Applied iteratively, we hope to converge on problem difficulty metric that better fits the observed data. This process is inspired by the parameter estimation procedures for Rasch models [11], which may not be directly applicable due to the large size of our problem space.

For a given challenge c_1 (e.g. "acknowledge"), we can first generate the list of all other challenges for which SpellBEE students have expressed distinctions (in either direction.) In Section 3.2, we chose to censor these distinctions in order to generate a bi-

Figure 4. A logistic regression model is used to estimate the difficulty of the word "abandon." At left, the first estimate is based on the original I_{50} difficulty values. At right, the third iteration of the estimate is constructed based on data from the previous best estimate. The point estimate dropped from 8.0 (from I_{50}) to 7.06 (from iteration 1) to 6.81 (from iteration 3.)



nary value representing the difficulty ordering. Here we will make use of the actual distinction counts in each direction. For each challenge with which pairwise distinctions for c_1 are available, we note our current-best estimate of the difficulty of c_2 (initially, using I_{50} values), and note the number of distinctions indicating that c_1 is the more difficult challenge. We can then regress the grouped distinction data on the problem difficulty estimate data to construct a logistic model relating the two. For some c_1, if the relationship is statistically significant, we can use it to generate a revised estimate for the difficulty of that challenge. By solving the regression equation for the c_2 problem difficulty value for which 50% of distinctions find c_1 harder, we can calculate the difficulty of a problem for which relative-difficulty distinctions are equally likely in either direction. This provides a revised estimate for the difficulty of the original problem, c_1. We use this procedure to calculate revised estimates for every challenge in the space (unless the resulting logistic regression model is statistically not significant, in which case we retain our previous difficulty estimate.) This process can be iteratively repeated, using the revised difficulty estimates as the basis of the new regression models. Figure 4 plots this data for one word, using the difficulty estimates resulting from the third iteration of the estimation.

A second approach towards incorporating observed distinction data into a unified problem difficulty scale is briefly introduced and compared to the other metrics. Here, we recast the estimation problem as a sorting problem, and use a probabilistic variant of the bubble-sort algorithm to reorder consecutive challenges based on available distinction data. Initially ordering the challenge words alphabetically, we repeatedly step through the list, reordering challenges at indices i and $i + 1$ with a probability based on the proportion of distinctions finding the first challenge harder than the second.[11] After "bubbling" through the ordered list of challenges 200,000 times, we interpret the rank-order of each challenge as a difficulty index. These indices provide a metric of difficulty (which we refer to as $ProbBubble$), and a means for predicting the relative difficulty of any pair of challenges (based on index ordering.)

[11] If distinctions have been observed in both directions, the challenges are reordered with a probability determined by the proportion of distinctions in that direction. If no distinctions in either direction have been observed, the challenges are reordered with a probability of $p = 0.5$. If distinctions have been observed in one direction but not the other, the challenges are reordered with a fixed minimal probability ($p = 0.1$).

Table 2. Summary table for the predictive accuracy of various difficulty metrics. For each metric, the percentage of accurate predictions of pairwise difficulty orderings is noted. The accuracy of the I_{50} metric is measured against all of the 3,349,602 pairwise orderings identified by student distinctions. The accuracy of the data-driven metrics ($I_{50}rev.1$ and $ProbBubble$) are based on the average results from a 5-fold cross-validation, in which the metrics are constructed or trained on a subset of the pairwise distinction data and are evaluated on a different set of pairwise data (the remaining portion.)

Difficulty Model	*Predictive Accuracy*
I_{50}	75.66%
$I_{50}rev.1$	84.79%
$ProbBubble$	84.98%

Table 3. Spearman's rank correlation coefficient between pairs of problem difficulty rank-orderings ($N = 3129$, $p < 0.01$, two-tailed.)

Metric 1	Metric 2	Spearman's ρ
I_{50}	$I_{50}rev.3$	0.677
I_{50}	$ProbBubble$	0.673
$I_{50}rev.3$	$ProbBubble$	0.908

Given the pairwise technique used in Section 4.2 for analyzing the fit of a difficulty metric for a set of pairwise difficulty orderings, we can examine how these two data-driven models compare to the original I_{50} difficulty metric. Table 2 summarizes our findings. Here we observe that the data-driven approaches provide an improvement of almost 10% accuracy with regard to the prediction of pairwise difficulty orderings. As was noted earlier, cycles in the observed pairwise difficulty orderings prevent any linear metric from achieving 100% prediction accuracy, and the maximum achievable accuracy for the SpellBEE student data is not know. We do note that two different data-driven approaches, logistic regression-based iterative estimation and the probabilistic sorting, arrived at very similar levels of predictive accuracy. Table 3 uses Spearman's rank correlation coefficient as a tool to quantitatively compare the three metrics. One notable finding here is the extremely high rank correlation between the $ProbBubble$ and $I_{50}rev.3$ data-driven metrics.

6. Conclusion

The findings from the research questions posed here are both reassuring and revealing. Although the NISS study was done over 50 years ago, much of its value seems to have been retained. The NISS-based I_{50} difficulty metric was observed to correctly predict 76% of the pairwise difficulty orderings mined from SpellBEE student data. Many of the challenges for which the difficulty metric achieved low predictive accuracies corresponded with words whose cultural relevance or prominence has changed over the past few decades. The data-driven techniques presented in Section 5 offers a means for incorporating these changes back into a difficulty metric. After doing so, we found the predictive accuracy increased approximately 10%, to the 85% agreement level.

The key technique used here to enable the assessment and improvement of problem difficulty estimates works even when not all students have attempted all challenges or when the selection of challenges for students is highly biased. It is data-driven, based on identifying and counting pairwise distinctions indicated indirectly through observations of student behavior over the duration of use of an education system. The pairwise distinction-based techniques for estimating problem difficulty information explored here is a part of a larger campaign to develop methods for constructing educational systems that require a minimal amount of expert domain knowledge and model-building. Our BEEweb model is but one such approach, the Q-matrix method is another [3], and most the IRT-based systems discussed in the introduction are, also. Designing BEEweb activities only requires domain knowledge in the form of a problem difficulty function and a response accuracy function. The latter can usually be created without expertise, and the former can now be approached, even when collected data is sparse and biased, using the techniques discussed in this paper.

References

[1] Ari Bader-Natal and Jordan B. Pollack. Motivating appropriate challenges in a reciprocal tutoring system. In C.-K. Looi, G. McCalla, B. Bredeweg, and J. Breuker, editors, *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED-2005)*, pages 49–56, Amsterdam, July 2005. IOS Press.

[2] Ari Bader-Natal and Jordan B. Pollack. BEEweb: A multi-domain platform for reciprocal peer-driven tutoring systems. In M. Ikeda, K. Ashley, and T.-W. Chan, editors, *Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS-2006)*, pages 698–700. Springer-Verlag, June 2006.

[3] Tiffany Barnes. The q-matrix method: Mining student response data for knowledge. Technical Report WS-05-02, AAAI-05 Workshop on Educational Data Mining, Pittsburgh, 2005.

[4] Klaus Brinker, Johannes Fürnkranz, and Eyke Hüllermeier. Label ranking by learning pairwise preferences. *Journal of Machine Learning Research*, 2005.

[5] Leonard S. Cahen, Marlys J. Craun, and Susan K. Johnson. Spelling difficulty – a survey of the research. *Review of Educational Research*, 41(4):281–301, October 1971.

[6] Chih-Ming Chen, Chao-Yu Liu, and Mei-Hui Chang. Personalized curriculum sequencing utilizing modified item response theory for web-based instruction. *Expert Systems with Applications*, 30, 2006.

[7] Bruce Choppin. A fully conditional estimation procedure for rasch model parameters. CSE Report 196, Center for the Study of Evaluation, University of California, Los Angeles, 1983.

[8] Ricardo Conejo, Eduardo Guzmán, Eva Millán, Mónica Trella, José Luis Pérez-De-La-Cruz, and Antonia Ríos. Siette: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, 14:29–61, 2004.

[9] Michel C. Desmarais, Shunkai Fu, and Xiaoming Pu. Tradeoff analysis between knowledge assessment approaches. In C.-K. Looi, G. McCalla, B. Bredeweg, and J. Breuker, editors, *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED-2005)*. IOS Press, 2005.

[10] B. S. Everitt. *The Analysis of Contingency Tables*. Chapman and Hall, 1977.

[11] Gerhard H. Fischer and Ivo W. Molenaar, editors. *Rasch Models: Foundations, Recent Developments, and Applications*. Springer-Verlag, New York, 1995.

[12] Harry A. Greene. *New Iowa Spelling Scale*. State University of Iowa, Iowa City, 1954.

[13] Jeff Johns, Sridhar Mahadevan, and Beverly Woolf. Estimating student proficiency using an item response theory model. In M. Ikeda, K. Ashley, and T.-W. Chan, editors, *Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS-2006)*, pages 473–480, 2006.

[14] Mark Wilson and R. Darrell Bock. Spellability: A linearly ordered content domain. *American Educational Research Journal*, 22(2):297–307, Summer 1985.