ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-663

The Design, Deployment and Evaluation of the AnimalWatch Intelligent Tutoring System

Paul R. Cohen[1], Carole R. Beal[2], Niall M. Adams[3]

Abstract. Europe and the U.S. both face the challenges of urban schools with low-achieving adolescent learners, many of whom are not proficient in the language of instruction. This paper describes the deployment and evaluation of the AnimalWatch intelligent tutoring system for mathematics in challenging classrooms. Previous studies demonstrated that AnimalWatch benefits 12-14 year-old students in relatively controlled conditions. The current study indicates that the system can help older, very low-achieving students in challenging secondary schools that serve diverse student populations.

1 Introduction

AnimalWatch was designed to help middle school students (10-14 year olds) build skills in pre-algebra topics such as number sense, computation, fractions, decimals, percentages and proportions, and rational numbers. This paper focuses on three aspects of the AnimalWatch project: designing engaging tutoring systems for young adolescent learners (Sec. 2), the challenges of deploying systems in large urban schools with diverse student populations (Sec. 3), and evaluations of the efficacy of tutoring systems (Sec. 4).

In United States schools, pre-algebra skills are introduced in Grade 6 (age 11) and covered in more depth in Grade 7, in preparation for Algebra 1, which is introduced in Grade 8 or 9. However, disparities in educational achievement are such that AnimalWatch has also recently been deployed to help high school (secondary) students catch up on the mathematics they did not master in middle school.

There is some urgency to questions about the efficacy of intelligent tutoring systems such as AnimalWatch: American students perform relatively poorly on international assessments [18]. In the large California city where AnimalWatch was most recently tested, only 36% of Grade 6 students scored at the "Proficient" level or better on the 2007 end-of-year California Standards Test-Math [17]. Ethnic gaps in achievement persist, with more White and Asian-American students scoring at a "Proficient" or "Advanced" (59% and 77%, respectively) than their African-American and Latino/a peers (36% and 27%, respectively). One in four Californian students is an English language learner. Many of these students do not complete secondary school and have limited opportunities in the labor force.

Several European countries face similar challenges in preparing young people for the workforce [12]. Although there is demand for workers with low skills in some sectors of the European labor market, the pool of young people who lack skills and qualifications for high-paying jobs is large and increasing [15]. Primary and secondary schools now enroll large numbers of children from immigrant families, many of whom are not proficient in the language of instruction and perform poorly in school [8, 9]. In both the United States and Europe, qualified teachers are retiring and it is proving difficult to recruit and train teachers who are both qualified to teach mathematics and to work with students who do not speak or read the language of instruction [11].

Technology-based instruction might help to improve students' mathematics achievement, although a recent report from the United States Institute for Education Sciences appears to cast doubt on the value of technology-based instruction, having found no benefits for classrooms that used commercial educational software products [19]. However, the software products evaluated in the study were not particularly innovative; most functioned like electronic versions of textbooks and lacked the interactivity, individualization of instruction, and rich multimedia features of instructional software being developed in research laboratories. What is needed are demonstrations that intelligent tutoring systems can be effective in the classroom. Such demonstrations should help to identify the design, deployment and evaluation factors that contribute to, and in some cases detract from, the success of these applications. This paper presents evidence that the AnimalWatch intelligent tutoring system *does* help struggling students learn mathematics, and it relates some lessons learned while deploying the software with large numbers of low-achieving students in urban schools.

2 The Design of AnimalWatch

An intelligent tutoring system should engage students and develop their problem solving skills while teaching material that is aligned to state and national curriculum standards. The system should also be easy to deploy in classrooms with low-end computers, and not require extensive training or technical support for teachers. The system should automatically collect data that will be used to evaluate how well it engages students and helps them learn the target math skills. This section discusses the design of AnimalWatch in terms of these criteria.

Connect Mathematics with Science Following the recommendations of the National Council of Teachers of Mathematics, AnimalWatch integrates mathematics learning with authentic environmental science material [21]. AnimalWatch engages students in narratives about tracking and monitoring the status of endangered species (hence the system's name). By connecting math problem solving with science, the student encounters many examples of how mathematics can apply to real-world problems and contexts, at a point in the curriculum when many students begin to complain that math is

[1] USC Information Sciences Institute
[2] USC Information Sciences Institute
[3] Imperial College London

disconnected from their lives. Environmental science is engaging to many young adolescents, both boys and girls, and also aligns with many state frameworks for middle school science (e.g., the California Science Grade 6 and 7 curricula focus on Earth Science and Life Science, respectively).

Problem-based Instruction Prior research indicates that word problem solving is often difficult for students because it requires multiple skills beyond simple computation: the ability to understand what a problem is asking, to construct equations from text, to compute an answer, and then to evaluate it for accuracy and plausibility in the context of the problem information. AnimalWatch provides students with opportunities to develop and practice these skills. AnimalWatch includes approximately 1100 word problems, organized into narratives about endangered and threatened species. Each word problem includes an introduction with authentic background information, a graphic (image, figure or table), and a question derived from the introduction. Scientific terms in the word problems that may not be familiar to students are linked to an integrated glossary. Students enter their answer into an answer box and receive immediate feedback, including hints and explanations in a variety of media.

One-on-one Tutoring Much research suggests that the ideal learning context — the "gold standard" — is one-on-one instruction with an experienced human tutor (e.g., [5]). Human tutors present problems to help diagnose the student's sources of difficulty, choose problems within the student's "Zone of Proximal Development," scaffold the student to a successful solution, and then attribute the success to the student's effort and enhanced understanding [7, 14, 16]. However, in our partner schools, a math teacher typically works with five classes of 30 or more students each, making it extremely difficult to give students individualized instruction. A student who struggles with a particular math concept or skill can quickly fall behind as the class moves on to new material.

AnimalWatch is designed to help students build proficiency in topics that have not yet been mastered, based on a curriculum of 30 math skills. The specific sequence of word problems that is presented to an individual student is customized to his or her proficiency level, which, of course, changes during sessions with the tutor. Math topics and the difficulty of individual word problems are increased as the student demonstrates that she or he can solve problems involving a particular skill. Skills estimated to have been mastered are periodically reviewed; more specifically, if a student makes errors on a problem involving prerequisite math skills, the probability of selecting a problem involving those earlier skills will increase. Thus, the system adaptively focuses on the areas that each student most needs to practice.

When a student needs help solving a problem, clicking on the "hint" icon brings up a menu of multimedia tutorial resources, including text explanations (e.g., how to find the least common denominator), worked examples, interactive solutions, and video lessons. If the student makes a problem-solving error, text feedback about accuracy is presented, followed by an operations hint (e.g., "No, that's not quite right." "Are you sure you're subtracting?"). A third error elicits a recommendation to view the associated help resources. When the student clicks on the "help" icon, a menu window appears, showing the options available for that topic. The student can then select the type of help he or she would like to see, or can view each type in turn to review alternative solutions and explanations. A fourth error elicits the correct answer, which the student is required to enter before moving on to a new problem.

Some students may actually prefer technology-based assistance to tutoring by human teachers. In our prior work with a different intelligent tutoring system (an ITS for high school high-stakes test preparation) we found that students who described themselves as disengaged from math (and whose teachers agreed) were highly likely to access multimedia help resources in an effort to learn the material [3]. Apparently, disengaged students are willing to seek problem solving help from the computer, whereas they are reluctant to do so from their teacher or classmates. Our pilot work with AnimalWatch indicates that a similar effect may be at work, with the lowest performing students showing high rates of using multimedia help resources.

Skills Practice In addition to word problem solving, AnimalWatch includes a module that provides students with opportunities to build computational fluency and automatic retrieval of math facts. This module is based on prior research indicating that students' proficiency with basic math facts and simple computation predicts their ability to solve complex word problems [22]. When lower-level processes such multiplication are automatic, cognitive resources are available to allocate to higher-order problem solving activities such as identifying what the problem is asking (problem representation) and checking possible answers in relation to the problem context. Royer and his colleagues found that training students in basic math facts was associated with improvement on achievement test problems. The role of computational fluency is strongest when students have limited time to solve problems, for example, on high-stakes tests [24].

By design, the basic math skills practice modules are distinct from the primary educational activity of solving word problems. There are twelve "skillbuilders" that test students on easy true-false problems (e.g., is 3 + 4 = 8 true or false?). Item difficulty is low, which motivates students to repeat the units (because they can achieve high scores); in turn, repetition strengthens fluency. Students also can practice solving math problems from the Math League, a popular national competition that includes practice activities completed each week by students around the country [13]. Math League items require insight or innovative solution strategies. In AnimalWatch, students may switch at any time between tutoring on word problems and these alternative activities. This provides valuable information about students' levels of engagement with the word problems.

Design for Statistical Student Modeling As noted, AnimalWatch maintains a model of each student's estimated proficiency with the target math skills. In the past, these models were fairly simple and heuristic, but recently, we have developed statistical models, particularly Hidden Markov Models of engagement [4]. As increasing numbers of students use ITSs, there will be opportunities for new kinds of data mining and statistical modeling; not only models of outcomes (e.g., improvements on tests) but models of students' learning processes. AnimalWatch is designed to support statistical modeling of students' behavior as they work with the software. It is a Web-based application that gathers moment-by-moment information about every observable aspect of students' activities and organizes the information in a temporal object store.

3 Deploying AnimalWatch

This section discusses some of the requirements for deploying ITSs — for getting them into classrooms or making them available to students in other ways.

Align with State Standards With increased emphasis on testing in recent years, teachers are reluctant to devote classroom time to activities that are not explicitly aligned with educational standards (on which annual achievement tests are based). A challenge to national or international deployment of ITSs like AnimalWatch is that states and nations have different mathematics curricula, so a math topic that is introduced in one grade in some states may be covered in different grades elsewhere. AnimalWatch is aligned with the California and Massachusetts Mathematics Standards for middle school mathematics and also with the process (i.e., problem-solving skills) standards set by the National Council of Teachers of Mathematics [21].

Identify a Role for the ITS Teachers will want to know how an ITS meshes with their own instructional activities. Originally, AnimalWatch was designed to supplement classroom activity, to review and reinforce learning of targeted topics. However, after years of meeting with teachers and leaders of community groups, several other roles have been added: AnimalWatch is used as an after-school activity in community centers, as a remedial tutor for high-school students who never mastered middle school mathematics, as a tutor in an elite program for inner-city children run by the University of Southern California, and, very recently, as a tutor for blind children.

Assessment and Learner Tracking The ability to assess students' performance and track it over time is very attractive to teachers. AnimalWatch currently includes several assessment instruments:

- **Pre- and Post-tests** These tests are completed online and scored automatically. The 30-item tests include sub-scores for computation (10 items) fractions (6 items), one-variable equations (6 items) and rational numbers (proportions, discounts, unit conversion, etc., 8 items) mapped to the California standards for Number Sense, Algebra and Functions, and Measurement and Statistics.
- **Mathematics motivation** How students perform in math reflects motivational as well as cognitive processes. Much research indicates that students' beliefs about their ability in math, the value that they place on being successful in math, and the extent to which they see math as important contribute to math achievement. The "Math Profile" is an online self-report instrument designed to assess students' math self concept, and value of math [6, 10].
- **Cognitive assessments** AnimalWatch includes three online assessments of cognitive factors that have been identified as predictors of math problem solving: A spatial cognition task based on mental rotation [23], a math-fact retrieval task based on judging the truth of simple equations as quickly as possible [22], and a Piagetian assessment of formal operational reasoning.

Resources for classroom integration AnimalWatch includes online resources to help teachers integrate the activity into their classrooms. These include a professional development manual and curriculum guide that can be viewed online or downloaded in PDF format; a users' wiki, with documentation, troubleshooting tips, discussion forum, and frequently-asked questions.

Technical requirements Classroom teachers have many demands on their time and have limited patience for buggy software. Schools generally have only limited technical support and most have limited bandwidth. AnimalWatch is stable, works with both PC and Apple computers and browsers, and with wired and wireless networks. Nothing needs to be installed on school computers. There is no need for district technology specialists to provide technical support. AnimalWatch automatically upgrades itself. Where schools block open Internet access, we have successfully worked with districts to provide port information so that school computers can connect to the AnimalWatch site. (If necessary, AnimalWatch can be installed on one machine in a computer lab, which then acts as a server for the other machines, with the media files provided to each student computer on CD-ROM.) File compression algorithms are used to stream the video lessons and graphics to ensure adequate performance over sometimes-slow school networks. Common technical problems likely to be encountered at school sites have been identified, and "troubleshooting" tips and resources have been created in documentation for teachers (e.g., how to set the computer screen resolution). Data collection is automated. As students work with AnimalWatch, their actions with the keyboard (e.g., answers, latencies, requests for multimedia, help with problem solving, navigation between modules, etc.) are recorded and processed automatically.

Other Deployment Issues Classroom-based research can be challenging for many reasons. It can be difficult to ensure that students have equivalent time with the software. Other activities and special events frequently interrupt the school's schedule. Establishing randomized comparison groups is rarely feasible. For one thing, some classes may be assigned to a software group simply because the computer lab is available only when those classes meet. Teachers may introduce a selectivity bias by deciding to have their lowest-performing classes work with the software on the assumption that the benefits should go to the most needy students. In the schools where AnimalWatch has been deployed, student attrition and absenteeism is very high, so a sizable proportion of the sample is lost between pre- and post-test. Conversely, new students frequently arrive in classes after the AnimalWatch activity has started and students have already completed the pre-test. Students are also frequently re-assigned from one class period to another, meaning that a student may begin in an intervention group and later re-appear in a comparison group. These and other factors make it difficult to establish that a tutoring system originally developed for research purposes can also be effective in real classrooms.

4 The Efficacy of AnimalWatch

This section presents some issues that arise in testing the efficacy of Intelligent Tutoring Systems. The discussion very easily could fill a textbook on empirical methods. We refer readers to [20] for a wealth of advice on metrics, experiment protocols and analysis methods. Here, we focus on how we evaluated the efficacy of AnimalWatch.

AnimalWatch was developed for middle school students and tested initially with sixth graders. [4]. In 2005, we were approached by a charter school district in which high school (secondary) students needed remedial help with pre-algebra math. The study included Grade 9 students enrolled in Algebra 1 classes in four high schools (N = 172). The schools served primarily African-American and Hispanic students. The sample included 88 students who spoke English as their primary language, and 84 English Language Learners. Overall performance in math was poor; nearly 80% of the sample scored at the Below Basic or Far Below Basic level of the California Standards Test-Math. Teachers reported that almost half of the students were failing Algebra because they had not mastered the prerequisite skills (arithmetic, fractions and rational numbers), and requested that the students work with AnimalWatch to review this material.

Each student in the sample took the pretest, then worked with AnimalWatch, then took the posttest (described in Sec. 3). The total amount of activity with AnimalWatch varied considerably across students due to absenteeism, dropping out of school, and sessions being cancelled for higher-priority activities or emergencies at the schools. Even when students attended sessions, they did not all work exclusively on AnimalWatch problems. The number of problems that students worked on ranged from a low of 2 to a high of 88, with mean 26.8 and median 24.

Efficacy is a relationship between an *intervention* — in this case, students' work with AnimalWatch — and some *outcome* such as improved math scores. Ideally, the relationship should be positive, that is, more work with AnimalWatch should produce better outcomes. One measure of outcomes is the difference between posttest and pretest scores. This suggests a model of the form:

$$Posttest - Pretest = \beta(AW) \quad (1)$$

where AW represents work with AnimalWatch and β represents efficacy, that is, the relationship between the intervention and the outcome. Our analysis is based on a slightly different model:

$$Posttest = \beta_1 Pretest + \beta_2(AW) \quad (2)$$

The reason for this model has to do with the dual role of pretest scores. Measures of efficacy — the relationship between intervention and outcome — should be untainted by other factors that could produce good outcomes. In particular, a student's prior mathematics knowledge might influence how much work the student does with AnimalWatch and perhaps also the benefits of this work. Our strategy is to examine the *partial* relationship between the intervention and outcomes, holding the student's prior mathematics knowledge constant. Because our best estimate of students' prior mathematics knowledge is their pretest score, we need a model that allows us to examine the independent contributions of pretest score and the intervention to posttest score, as shown in Eq. 2.

There are many ways to score performance on tests. In AnimalWatch, students did not answer all the items on their pretests and posttests, so each item, for each student, was either Correct, Incorrect, or Not Attempted. Let N_C and N_I denote the number of correct and incorrect answers on a particular student's test. One measure of performance is $N_C/(N_C + N_I)$, which acknowledges that non-attempted problems do not provide information about a student's mathematics ability. Another measure is the odds ratio — N_C/N_I — but the distribution of odds ratios was skewed in AnimalWatch students. The log of the odds ratio, however, is nearly symmetric and close to Gaussian. The following analyses are for log-odds, but our qualitative conclusion hold for other measures, including $N_C/(N_C + N_I)$ (see Fig. 2 and discussion). Let $PreLO$ and $PostLO$ denote $log(N_C/N_I)$ for a student's pretest and posttest, respectively.

The student's work with AnimalWatch — the intervention — can also be measured in several ways. Recall that students worked on word problems, getting some right and others wrong, sometimes looking at hints; and had the opportunity to work on non-AnimalWatch activities, such as Skillbuilders. In AnimalWatch (and other ITSs) some students "abuse help" and do not actually try to solve problems, but click through hints and help, mechanically. We looked at several measures of the intervention, some of which gave slightly better results in the regression analyses discussed below. But we settled on an easy-to-interpret measure of amount of work the student did with AnimalWatch: the number of unique math word problems encountered by a student. Let AW denote this number.

We will treat $PreLO$ as a measure of the student's prior mathematics ability and look at the partial relationship between $PostLO$ and AW holding $PreLO$ constant. There are several ways to do this, each of which gives us some insight into the efficacy of AnimalWatch. As a simple statement of association, the partial correlation of AW and $PostLO$ holding $PreLO$ constant is 0.30 and the 95% bootstrapped confidence interval around this statistic is $[0.13, 0.44]$. Clearly there is a significant association between AW and $PostLO$ independent of $PreLO$.

Regression analysis gives similar results. Regressing AW and $PreLO$ on $PostLO$ produces a regression model that accounts for 56% of the variance in $PostLO$ ($p < .0001$). The least-squares parameter estimates for the model are $PostLO = 0.72(PreLO) + 0.012(AW) - 0.68$. T tests for all of these parameters are highly significant. Although the regression coefficients are partial, it is difficult to compare them because they are on different scales. To compare these effects, one can rescale them in terms of standard deviations of $PostLO$. This yields a model with standardized regression coefficients: $PostLO = 0.66(PreLO) + 0.21(AW)$.

The interpretation of this model is that a change of one standard deviation in $PreLO$ produces .66 standard deviations change in $PostLO$, while a change of one standard deviation in AW produces .21 units change in $PostLO$. (Standardized regression models have zero intercepts.) So, in terms of effects on $PostLO$, increasing one's level of effort with AnimalWatch by one standard deviation (approximately 19 problems) has roughly one third the effect (0.21 vs. 0.66) of being one standard deviation higher on the pretest scale. It is encouraging to see that the partial effects of $PreLO$ and AW on $PostLO$ are of the same order of magnitude. It means that posttest scores are certainly influenced by prior mathematics achievement (as measured by pretest scores), but working with AnimalWatch has a meaningful effect on posttest scores independent of pretest scores.

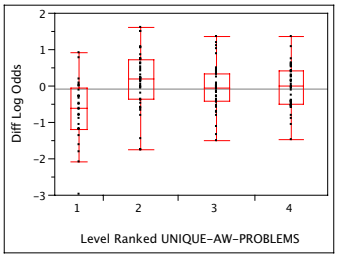

Figure 1. The relationship between $Rank(AW)$ and the residuals of the regression of $PreLO$ on $PostLO$

It is instructive to divide the AW scores into four groups, corresponding to AW scores in the first, second, third and fourth quartile of the distribution of scores, and then use this group label to organize pretest-posttest differences (i.e., $PostLO - PreLO$). One can see in Figure 1 that the biggest pretest-posttest difference was for students whose AW scores were in the second quartile. The analysis of variance associated with Figure 1 is significant ($p < .0002$), however, Tukey HSD pairwise tests found differences only between the first quartile of AW scores and all the others. One interpretation of these results is that the "dose-response" function flattens out quickly: Some amount of work with AnimalWatch will produce a pretest-posttest difference but more will not. However, students worked so little with AnimalWatch, in such chaotic classroom conditions, that we cannot

make any strong inferences about the dose-response function in regions where students work for longer in a more concentrated way.

It is clear that students who do the least work with AnimalWatch have the least improvement. Figure 2 shows two such regressions, one for $PreLO$ and $PostLO$ and the other for $N_C/(N_C+N_I)$ (also called $Correct/Attempted$). In each scatterplot, the green crosses denote students whose AW scores were in the first quartile of the distribution of AW scores. In other words, these are the students who worked on the fewest AnimalWatch problems. They fall disproportionately below the regression line, meaning that their posttest scores are lower than expected given their pretest scores. Chi-squared tests show that being in the lower quartile of AW is associated with being below the regression line for both log odds and correct/attempted scores ($p < .0001$ in each case).

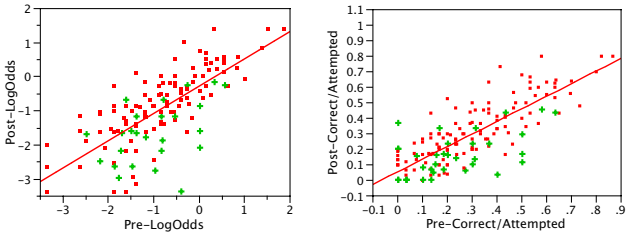

Figure 2. Regressions of pretest scores on postest scores (for log odds and Correct/Attempted scores), with the students who fall in the lower quartile of AW marked with green crosses.

5 Conclusion

Prior work with AnimalWatch under relatively controlled conditions indicated that it helped students roughly as much a small-group instruction with skilled math teachers [4]. The results of the present study indicate that AnimalWatch also helped older students in more challenging conditions: The schedules and class enrollments in the participating schools were chaotic; the student population included many learners with very low achievement in math; and many of the students were not proficient in English (the language of instruction). Even so, students who had more opportunity to work with the software showed greater pre- to post-test improvement than their peers who solved fewer AnimalWatch problems. These effects were small in absolute terms, probably because students worked with AnimalWatch for relatively little time (on average, they worked on roughly 25 word problems) and this effort was distributed over multiple sessions. With a longer-term intervention, there should be correspondingly greater improvement for struggling students.

However, it will not be easy to deploy such an intervention. One objective of this paper is to describe the conditions that many adolescents in the United States face in urban schools, and the barriers to their success in mastering basic mathematics. Our results indicate that although research-based intelligent tutoring systems can help, technology alone will not fix the problems of low achievement. Tutoring systems must be matched with teachers who are well-trained and supported in their work. Unfortunately, attrition among new teachers is high [2, 1]. All of the teachers with whom we worked in the study decided to leave the profession.

6 Acknowledgments

The AnimalWatch project is supported by award R305K050086 from the United States Institute of Education Sciences. We would like to thank our project colleagues Wesley Kerr, Jean-Philippe Steinmetz, Erin Shaw and Sinjini Mitra, and the participating schools for their assistance.

REFERENCES

[1] National Education Association. Meeting the challenges of teacher recruitment and retention., 2003.
[2] G. Barnes, E. Crowe, and B Schaefer. The costs of teacher turnover in five school districts., 2007.
[3] C. R. Beal, L. Qu, and H. Lee, 'Classifying learner engagement through integration of multiple data sources.', in *Proceedings of the 21st National Conference on Artificial Intelligence. Boston MA.*, (2006).
[4] C. R. Beal, E. Shaw, and M. Birch., 'Intelligent tutoring and human tutoring in small groups: An empirical comparison.', in *Artificial intelligence in education: Building technology-rich learning contexts that work*. Amsterdam: IOS Press., (2007).
[5] B.S. Bloom, 'The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring.', *Educational Researcher*, **13**, 4–16, (1984).
[6] M. Boekaerts, 'The on-line motivation questionnaire: A self-report instrument to assess students' context sensitivity', in *Advances in Motivation and Achievement, Vol. 12: New Directions in Measures and Methods*, eds., P. R. Pintrich and M. L. Maehr, pp. 77–120. Elsevier Science, (2002).
[7] A. L. Brown, S. Ellery, and J. Campione, 'Creating zones of proximal development electronically', in *Thinking practices: A symposium in mathematics and science education*, eds., J. Greeno and S. Goldman. Erlbaum, (1994).
[8] G. Christensen and P. Stanat, 'Language policies and practices for helping immigrants and second-generation students succeed', in *Transatlantic Task Forces on Immigration and Integration.*, (2007).
[9] M. Crul. Pathways to success for the children of immigrants., 2007.
[10] J. Eccles, A. Wigfield, R. D. Harold, and P. Blumenfeld, 'Age and gender differences in children's self and task perceptions during elementary school.', *Child Development*, **64**, 830–847, (1993).
[11] European Trade Union Committee for Education. Europe needs teachers, 2005.
[12] National Center for Education Statistics. Comparative indicators of education in the united states and other g-8 countries: 2006., 2007.
[13] The Math League. Math league website: http://www.mathleague.com/.
[14] M. R. Lepper, M. Woolverton, D. Mumme, and J. Gurtner, 'Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors.', in *Computers as cognitive tools*, eds., S. P. Lajoie and S. J. Derry, pp. 75–105. Erlbaum, (1993).
[15] S. McIntosh and H.. Steedman. Low skills: A problem for europe., 2000.
[16] D. C. Merrill, B. J. Reiser, M. Ranney, and J. G. Trafton, 'Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems.', *Journal of the Learning Sciences*, **2**, 277–305, (1992).
[17] California Department of Education. Dataquest query results retrieved from http://dq.cde.ca.gov/dataquest, 2007.
[18] U.S. Dept. of Education. No child left behind is working., 2006.
[19] U.S. Department of Education Institute of Education Sciences. Effectiveness of reading and mathematics software products: Report from the first student cohort, 2007.
[20] U.S. Department of Education Institute of Education Sciences. What works website: http://ies.ed.gov/ncee/wwc/twp.asp, 2008.
[21] National Council of Teachers of Mathematics. Principles and standards for school mathematics., 2000.
[22] J. M. Royer, L. N. Tronsky, Y. Chan, S. J. Jackson, and H. Merchant, 'Math fact retrieval as the cognitive mechanism underlying gender differences in math test performance', *Contemporary Educational Psychology*, **24**, 181–266, (1999).
[23] S. G.. Vandenberg. Mental rotation test, 1971.
[24] J. J. Walczyk and D. A. Griffith-Ross, 'Time restriction and the linkage between subcomponent efficiency and algebraic inequality success', *Journal of Educational Psychology*, **98**, 617–627, (2006).