Open Community Authoring of Targeted Worked Example Problems

Turadg Aleahmad, Vincent Aleven, and Robert Kraut

Human Computer Interaction Institute
Carnegie Mellon University, Pittsburgh PA 15213, USA
{turadg,aleven,robert.kraut}@cmu.edu

Abstract. Open collaborative authoring systems such as Wikipedia are
growing in use and impact. How well does this model work for the devel-
opment of educational resources? In particular, can volunteers contribute
materials of sufficient quality? Could they create resources that meet stu-
dents' specific learning needs and engage their personal characteristics?
Our experiment explored these questions using a novel web-based tool
for authoring worked examples. Participants were professional teachers
(math and non-math) and amateurs. Participants were randomly as-
signed to the basic tool, or to an enhanced version that prompted authors
to create materials for a specific (fictitious) student. We find that while
there are differences by teaching status, all three groups make contribu-
tions of worth and that targeting a specific student leads contributors to
author materials with greater potential to engage students. The exper-
iment suggests that community authoring of educational resources is a
feasible model of development and can enable new levels of personaliza-
tion.

1 Introduction

Traditionally in the development of educational resources, one person or cohesive
group produces each artifact, be it paper textbook or computer tutor. This model
can require great resources from one group, particularly for tutoring systems and
even more so for individualized tutoring. Traditional model-based tutors cost 100-
1000 hours of time from skilled experts [1,2]. Newer approaches [3] such as CTAT
[4], REDEEM [5] and ASSISTment [6] lower the expertise necessary to create a
tutor, but still require coordination of a group to create useful tutors. Like most
tutors, they are also limited in the dimensions by which they can personalize to the
student. In this paper we describe and study a prototype tutor authoring system
that uses community volunteers to create personalized instruction.

Over the last decade we have seen new development models that take ad-
vantage of the openness of the World Wide Web. Encyclopedias, web browsers,
computer operating systems, and other complex artifacts have been created by
loose networks of volunteers building on each other's contributions. These openly
developed products often meet and sometimes exceed the quality of more cohe-
sive sources and in general lower their costs. In education, there are a number of

open access initiatives such as MIT's OpenCourseWare or CMU's Open Learning Initiative. These systems are open in that they are free to use, but not in the sense of open to contributions. Wikipedia is an example of a system open to all (even anonymous) contributors and is excellent for sharing *about* something, though it disallows instructional information. Connexions [7] has been successful in supporting collaborative development of textbooks and newer site like Curriki, Wikibooks and Wikiversity are trying to adapt existing wiki software to the development of educational materials. However, these systems are all focused on traditional inert learning resources and are limited by available wiki software. Community authored intelligent tutoring systems will need semantic content structures and affordances for personalized instruction.

We propose a four phase cycle of a new system for open resource development: Generation, Evaluation, Use, and Improvement. This study is an exploration into the Generation phase. Will volunteers produce materials of sufficiently high quality that they can be put in the hands of students with little or no editorial oversight? We report results obtained with a simple web-based authoring interface for worked example problems, a tutoring resource that is very intuitive and hence can be authored by lay volunteers. Worked example problems are versatile for computer tutoring systems. They can be presented to a student as a full problem and solution, as a problem with a partial solution, or as simply a problem, as on a homework sheet. In more intelligent tutoring systems, scaffolds contained within the problem can be faded as the student masters each skill. Because worked example problems enhance problem solving in intelligent tutoring systems, they also complement what already exists [8,9]. Pedagogically, worked-examples both instruct and help to foster self-explanation [10]. Thus they are a useful educational resource, although we do not test their usefulness in this paper.

To explore the impact of open development and diverse levels of expertise, our study was open to all comers. Reasonably this would lead to a volume of content without much value and this motivated our first hypothesis. *H1: Identifying the good from the bad contributions is easy. We expect that all contributions are good, easily fixed, or easily filtered.* To assess the impact of expertise, we asked each participant whether they were math teachers, other teachers, or not teachers at all. We used this data to assess *H2: Math teachers submit the best contributions.* While math enthusiast amateurs may have the appropriate *content knowledge* and non-math teachers may have the appropriate *pedagogical knowledge*, neither will have much *pedagogical content knowledge* (as defined by Shulman [11]).

A goal of the system is to facilitate personalized instruction. Personalization has already been shown to improve both student engagement and test scores. Fourth grade math students make better pretest-to-posttest gains with personalized instruction and also perform significantly better on both the pretest and posttest problems[12]. Similar effects have been found with 5th th and 6th th grade students [13]. Personalized instruction has also been demonstrated to increase the engagement and learning outcomes of minority groups [14, e.g. Hispanic]. To facilitate personalization, we created a feature in the tool in which authors are

prompted to target their instruction to a specific student, described through a student profile. Thus, *H3: Student profiles lead to tailored contributions.*

Because being shown a specific individual to help is likely to draw out more altruistic behavior, we believed that the profiles may motivate authors to make better contributions [15], leading to two further hypotheses. *H4: Student profiles increase the effort of authors. H5: Student profiles lead to higher quality contributions.*

The main analysis reported in this paper looks at the contributions targeted at one specific skill, in order to control for variability among skills. The skill of understanding and applying the Pythagorean Theorem was chosen for three reasons. The first is that it is difficult to learn. Data from the ASSISTment assessment system [16] show it is the most difficult skill for students to acquire. The second is that because it has a visual component, a variety of problems are more difficult to generate by machine than with non-visual math skills. The third is that it affords a variety of real-world scenarios to demonstrate it.

In this paper, we describe the web-based community authoring tools, report the volunteer contributions that were submitted through the web site, and analyze the data to see if they support the hypotheses described above. We discuss the implications for web-based community authoring of instructional materials.

2   Methods

2.1   Apparatus

The experiment took place over the internet through interactions with a newly-built prototype web application for community-based authoring. Participants followed a URL to an experiment registration form. This form explained the goal of producing open educational resources and asked for their consent in the study. Participants were asked for their age and teacher status (math teacher, other teacher, or non-teacher / amateur). Those under 18 were directed straight to a survey without participating in the experiment.

Participants were then presented a page explaining what a worked example problem is and that their task is to create a worked example problem to teach the Pythagorean Theorem. They were provided a search box to look up on the web anything they wanted to learn or refresh themselves on and some simplified pedagogical principles to remember in creating their worked example. The next screen was the authoring tool. The tabs at the top guided the user in the task. The first sentence of the Start tab was specific to the condition within the sub-experiment. In the student profile experimental condition it read, Please create a worked-out example to provide practice to the student above in understanding and applying the Pythagorean Theorem. In the control generic condition, the words to the student above were stricken. Authors wrote the text of the problem statement in the large box to the left and drew their diagrams using a simple pen tool in the box to the right. After developing their problem statement and diagram, they clicked Add Step to add each step of the solution. The solution area had three columns, one for the work for the student to perform towards the solution, one for the explanation of the work in each step.

In the student profile experimental condition, the top of the authoring tool showed a floating description of the student to target with the example-to-be-authored. The instruction to the participant also differed, as described above. The participants in the experimental condition were presented a student profile to target. Profiles were randomly selected and participants saw a new one for each problem authored. Student profiles were designed to vary on six dimensions that might differentiate the learning patterns of real students. They varied on three dimensions of skill to increase variation of the submissions on skill-level appropriateness. These were proficiency in at the Pythagorean Theorem, general proficiency in math, and verbal skill. They were also varied on cultural attributes to prompt creativity of the participants and increase the personal relevance of the examples to students. These were gender, hobbies/interests, and home environment. Four hobbies were crossed with four home environments to create 16 unique student profiles. Distributed evenly among them were four skill profiles and the two genders. Additionally, each was assigned a favorite color to round out the description presented. Figure 1 shows an example profile.

Fig. 1. Sample profile in experimental condition

Participants. The URL to participate was advertised on various web sites both related to education and not. Participants could earn up to \$12 for their contributions, regardless of their quality. During the experiment 1427 people registered on the site to participate. After seeing the task in detail most did not continue, but 570 participants did use the system to submit 1130 contributions. Table 1 shows by teacher status the number of registered and contributing participants during the experiment.

Since completion of the experiment, more participants have contributed to the site but they are not included in the analysis that follows. At the close of the experiment, the web site was disabled. At the request of people who still wanted to participate, two months later it was restored with the compensation removed. In the four months that have elapsed, 93 people have registered and submitted 93 contributions, of which 40 pass machine vetting (see below). These extras have not been further tested for quality.

Machine Coding. The submissions were analyzed by software as a first-pass filter for minimum quality. The criteria were that the problem description had a length between 50 and 1000 characters and that at least one step was provided towards the solution. This machine filtering of the 1130 raw submissions left 551 machine-vetted submissions from 281 participants. Table 1 shows by teacher status the number of participants whose contributions passed machine vetting.

Table 1. Count of participants by teacher status and depth of participant

Participation	Math teachers	Other teachers	Amateurs
Registered	131	170	1126
Contributed also	70	72	428
Passed vetting also	26	35	220

Expert Coding. After the automatic filter, the remaining submissions were coded for quality by human experts. In a production version of the site, human coding would be drawn from the community. For this analysis, the two coders were a retired and a beginning math teacher. They coded using a custom web application that ensured they were blind to the sources. Submissions were coded on three criteria: quality of the problem statement (Statement), quality of the work shown (Work), and quality of the explanation of the solution (Explanation). The following are the ratings and definitions that they used:

(0) **Useless**   No use in teaching and it would be easier to write a new one than improve this one.
(1) **Easy fix**   Has some faults, but they are obvious and can be fixed easily, in under 5 minutes.
(2) **Worthy**   Worthy of being given to a student who matches on the difficulty and subject matter. Assume that the system knows what's in the problem and what is appropriate for each student, based on their skills and interests.
(3) **Excellent** Excellent example to provide to some student. Again, assume that the system knows what's in the problem and what is appropriate for each student, based on their skills and interests.

For quantitative analyses the categories were assigned the integers 0-3. The work and explanations were averaged to create a Solution quality and all three components were averaged to determine a Whole quality. Inter-rater reliability of the Statement quality was alpha=0.61, for the Solution quality was alpha=0.81, and for the Whole quality was alpha=0.78.

Ordinal variables were modeled as continuous in order to model the participant as a random effect. This requirement was due to a limitation in the statistic software available.

3   Results of Open Authoring

To test H1 we looked at the quality of all problems submitted and the work needed to classify them. Of 1130 raw submissions, 11% of whole problems (statements with solutions) were classified as Worthy, meaning that they are fit for use immediately. 39% were at least Fixable, meaning that they would be valuable with some additional effort. In general the statements were of higher quality than the solutions. 27% of statements were Worthy and 4% were Excellent as is. Figure 2 shows counts in each quality classification, plus a fifth column indicating
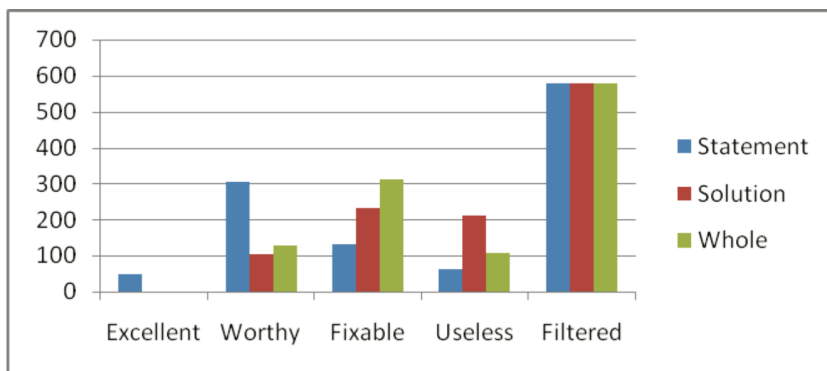
Fig. 2. Count of submissions in each quality classification

the 579 that were filtered by the machine vetting process. Classification into the Filtered category was a trivial computation. Classifying the three contribution components into the four-rating rubric took experts a median time of 36 seconds per contribution.

To test H2 we looked at the quality of each contribution as a whole, revealing no superior quality by teacher status (F(2)=1.53, p=0.22). Further analysis revealed that the effect on quality of teacher status interacted with the problem component.

Math teachers were best at writing problems statements. A comparison across teacher status showed a marginally significant effect (F(2)=2.39, p=0.093). Math teachers' contributions rated at M=1.79, followed by amateurs (M=1.45) and other teachers (M=1.45). A comparison of math teachers with the rest showed a significant effect (F(1)=4.80, p=0.015, one-tailed).

Contrary to H2, amateurs were best at writing solutions. A comparison across teacher status showed a marginally significant effect (F(2)=2.73, p=0.067). Amateurs did best (M=0.72) followed by math teachers (M=0.60) and then other teachers (M=0.48). A comparison of amateurs with the rest showed a significant effect (F(1)=4.87, p=.028).

4 Discussion on Open Authoring

Through automated methods and software supports for human judges, all the contributions were rated easily, supporting H1. In a short amount of time about 1500 people registered to contribute to a commons of educational materials. While not all came with the same intentions, over half walked away after determining that they did not want to participate fully. Of the raw submissions made, over half were trivial to filter by simple automated methods. Of the remaining, a novice and a veteran teacher were able to rate each of them on three attributes in less than a minute each. About 1/10th were ready to help students learn without

needing any modification. Many more were rated as fixable, meaning that with some additional work they would be ready. Statements were the highest quality components and solutions were the most difficult parts to author well.

Teacher status had an important impact on the quality of the components of contributions. As predicted in H2, math teachers were best at authoring problem statements. Surprisingly, amateurs authored the best worked solutions. Perhaps this is because they are better able to adopt a student's perspective. Math teachers performed worse than amateurs but better than non-math teachers. Perhaps this is because their pedagogical content knowledge helps compensate (but not fully) for their expert blind spots.

Overall, it is clear that, at least for worked examples of the Pythagorean Theorem, participants of all teaching statuses were likely to make contributions of value. Math teachers do a better job at some parts of the process, but even laymen do fairly well. This is fortunate because there are many more amateurs in the world than math teachers. In this study each participant made each contribution independently, but the best resources may come from collaboration. For example, a math teacher writes a problem statement and an amateur writes the solution. Educational content systems can benefit from opening the channels of contribution to all comers.

5   Results of Student Profile

Tailoring was analyzed to test H3 and measured as the degree to which various attributes of the contributed problem matched that of a student profile. Matching took two forms: using words primed by the student profile and matching the difficulty (reading or writing) to the skill levels in the profile. The use of words in the submission was analyzed using LIWC, a word counting tool, with its default dictionary [17] plus the word piano in the music category (to go with guitar, instrument, concert, etc.). Table 2 summarizes the results for the word matching. Mentioning an attribute drew out significant increases in authoring with that attribute on almost every measure, both over the generic condition and other profiles.

To test whether authors tailor their contributions to the verbal skill of the student, we compared the verbal skill level of the student profile presented to the author with the reading level of the authored submission. The reading level was measured using the Flesch-Kincaid Grade Level Formula. [18] This indicates U.S. school reading grade level, making the problem selection easy in real learning contexts. The text analyzed is the concatenation of the problem statement and all the explanation steps. Because readability metrics aren't calibrated to math expressions, the work steps were omitted from readability analysis. Outliers were curtailed by removing the top and bottom 2.5% percentile in the distribution of Flesch-Kincaid Grade Level (leaving -1.32 to 11.71). An F-test showed the differences across profile verbal skill levels to be significant ($F(2)=2.95$, $p=0.023$, one-tailed). Table 3 shows the results of pair-wise t-tests. Additionally, it is worth noting that authors sometimes took the student's verbal skill level as a cue for the subject matter of the problem statement, as in the following submission:

Table 2. Probabilities of authoring matching an attribute

Attribute	With generic (G)	With profiles not mentioning attribute (N)	With profiles mentioning attribute (M)	F-test (G-M)	F-test (N-M)
Female pronoun	5%	4%	16%	9.68*	12.82**
Male pronoun	19%	14%	19%	0.004	1.19
Sports word	9%	9%	24%	18.01**	11.89**
TV word	4%	4%	10%	8.36*	2.63†
Music word	2%	2%	9%	6.92*	8.93**
Home word	14%	n/a	20	3.60*	n/a

†p<.10 *p<.05 **p<.001

"Shakespeare sat down one day" and had a revolutionary idea. He would write text diagonally across a page rather than horizontally! He imagined the reader's surprise when they turned the page and saw such a change of events! However, he needed to carefully plan out how much space he had to write in, if he only wanted to write one line across the diagonal of the book. Unfortunately, Shakespare had no rules. After all, he thought he would never need one being a writer! Luckily for him, he remembered that the blank books and pages he ordered were exactly 6 inches wide and 8 inches tall. How much space does Shakespare have to write in?

Table 3. Correspondence of verbal and math skill levels with the authoring interface

Verbal Skill in Profile	Sign. Diffs	Mean Reading Level of Contribution	Std Err	General Math Skill in Profile	Sign. Diffs	Probability of Using 3-4-5 Triangle	Std Err
High	A	3.78	0.24	High	A	16%	0.05
Medium	A B	3.56	0.32	Medium	A B	26%	0.05
Low	B	2.93	0.33	Low	B	27%	0.04
None (control)	B	3.20	0.16	None (control)	A B	21%	0.03

Math difficulty was measured more simply because there is no established metric available. Since all problems were on the Pythagorean Theorem, we chose to measure math difficulty by whether the problem uses only the 3-4-5 triangle, the least challenging numerical solution. An F-test showed the differences across profile general math proficiency levels to be marginally significant ($F(1)=2.35$, $p=0.063$, one-tailed). Table 3 shows the results of pair-wise t-tests.

Effort was analyzed to test H4. It was measured by both the length of each contribution and the time spent on it by the author. While authors in the generic control condition wrote an average of 766 characters per contribution, authors in the student profile condition wrote 847 characters, a marginally significant difference ($F(1)=2.35$, $p=0.063$, one-tailed). Most of that difference is accounted for by the problem statements, for which the generic condition wrote 204 characters versus 250 characters with profiles. This 23% increase in length was significant

(F(1)=8.61, p=0.01). A similar analysis of time spent authoring (normalized) revealed no significant differences (t(235)=-0.15, p=0.56).

Quality was analyzed to test H5. The quality of the statement, the solution, and the whole were compared between the experimental and control conditions. F-tests showed no effects of the student profiles on the quality of submissions.

6   Discussion on Student Profile

All features of the profile display were accounted for in the problems submitted. Participants were more likely to mention a particular hobby when shown it in the profile. They were also more likely to make mention of some home environment (a feature of every profile). Particularly striking is the increase in likelihood of including a female in the problem statement. Without a profile, males were used in 19% of problem statements and females in just 5%. (The rest used only it or no pronouns.) Female student profiles bring female pronoun usage up to 16%, almost on par with males. Male pronoun usage is clearly the default of most authors since the usage without any profile is just as high as with a male profile. Furthermore, male pronoun usage was not much suppressed by the female profiles.

Participants shown the student profiles also tailored the skill level of their contributions. High and low reading levels differed by almost a grade level. Submissions for profiles with high general math skill level were one third less likely to make use of simple 3-4-5 triangle problems. Participants shown profiles of students wrote problem statements that were 25% longer.

It is perhaps odd then that they didn't spend significantly more time on these statements. One explanation is that the time typing is negligible compared to the time required to generate an idea. That the statements in the profile condition are so much longer suggests that the profile prompts ideas that are more involved.

One hypothesis on the student profiles condition did not bear out: that profiles would lead to contributions of higher quality on an absolute scale. Instead the contributions maintained quality. In other words, the tailoring came at no cost to the generic quality of the contributions.

7   General Discussion and Conclusions

This study looks at the feasibility of an open development model for developing resources for tutoring systems and a particular design feature to solicit to draw more and better work from contributors. We found that while over half the contributions were useless there were some gems. Importantly, it required little effort to separate the wheat from the chaff, confirming H1. Both professional educators and amateurs contributed a large portion of useful materials. Contrary to the prediction H2, contributions from math teachers were not superior to those from others. This is encouraging because there are many more people who aren't math teachers than who are.

Math teachers did write the best problem statements but amateurs wrote the best solutions. This suggests a model in which math teachers contribute the problem statements and amateurs write the solutions. In general, it suggests that users of different aptitudes and abilities be directed to different tasks within the collaborative authoring system, a solid design implication.

The student profile feature of the interface successfully drew out personalized resources (H3). On every attribute the profile increased the likelihood of targeting it. The profiles also drew out more effort on the part of participants (H4). While the profiles didn't measurably improve the quality of their contributions (H5), it is important to note that they did not impair them either.

An important limitation of the study is that there are no measures yet of how these contributions actually aid learning. The expert ratings were taken as proxies for the utility in real learning contexts, but the true test will be using the materials to teach real students and measure their gains versus alternative materials. One potential pitfall is that the personalizing details in the tailored resources distract students from learning. Of course, the improvements to their motivation might offset this. A real-world study is necessary to answer these questions.

Another key limitation of the findings here is the ecological validity of paying participants for their contributions. The problem is not that participants were incented to contribute. One can imagine a future system with incentives such as peer status or competitions with non-monetary awards. (e.g. [19]) Certainly, volunteers are always motivated by some incentive, external or internal. How though do contributions differ under more ecologically valid incentives? Because participants were paid for any contribution, there is good reason to believe that real world volunteers would be more dedicated and likely to produce higher quality materials on average.

This study focused on phase one, Generation, of a proposed cycle of development. It also touched upon phase two, Evaluation, through the finding that experts can evaluate each contribution in about half a minute. We will turn next to the fourth phase, Use, in order to begin answering the above limitations. We will first develop a personalized homework system that creates homework assignments for students based on their real personal profiles. This will also require expanding the Generation aspects of the system to facilitate a wider variety in the skills taught. Farther in the future, we will explore the potential for developing interactive tutoring around the worked example problem artifact, through dynamic scaffolding of its subcomponents.

This study has positively, if partially, demonstrated the feasibility of an open development model for resources for tutoring. Volunteers regardless of professional expertise are able to make useful contributions and features of the authoring interface can incline contributions to have different features and make instruction more socially inclusive.

References

1. Anderson, J.R.: Rules of the mind. Erlbaum, Hillsdale (1993)
2. Murray, T.: Authoring intelligent tutoring systems: An analysis of the state of the art (International Journal of Artificial Intelligence in Education) 10, 98–129
3. Murray, T., B.S.A.S. (eds.): Tools for Advanced Technology Learning Environments. Kluwer Academic Publishers, Amsterdam (2003)
4. Aleven, V., Sewall, J., McLaren, B.M., Koedinger, K.R.: Rapid authoring of intelligent tutors for real-world and experimental use. In: Proceedings of the 6th IEEE International Conference on Advanced Learning Technologies (ICALT 2006), pp. 847–851. IEEE Computer Society, Los Alamitos (2006)
5. Ainsworth, S., Fleming, P.: Evaluating a mixed-initiative authoring environment: Is redeem for real? In: Proceedings of the 12th International Conference on Artificial Intelligence in Education, pp. 9–16. IOS Press, Amsterdam (2005)
6. Turner, T.E., Macasek, M.A., Nuzzo-Jones, G., Heffernan, N.T.: The assistment builder: A rapid development tool for ITS. In: Proceedings of the 12th Annual Conference on Artificial Intelligence in Education, pp. 929–931 (2005)
7. Baraniuk, R., Burrus, C.S., Hendricks, B., Henry, G., Hero, A., Johnson, D.H., Jones, D.L., Nowak, R., Odegard, J., Potter, L., Reedstrom, R., Schniter, P., Selesnick, I., Williams, D., Wilson., W.: Connexions: Education for a networked world. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, ICASSP (2002)
8. McLaren, B.M., Lim, S.J., Gagnon, F., Yaron, D., Koedinger, K.R.: Studying the effects of personalized language and worked examples in the context of a web-based intelligent tutor. In: Ikeda, M., Ashley, K.D., Chan, T.-W. (eds.) ITS 2006. LNCS, vol. 4053, pp. 318–328. Springer, Heidelberg (2006)
9. Schwonke, R., Wittwer, J., Aleven, V., Salden, R., Krieg, C., Renkl, A.: Can tutored problem solving benefit from faded worked-out examples. In: European Cognitive Science Conference, pp. 23–27 (2007)
10. Renkl, A., Atkinson, R.K.: Learning from examples: Fostering self-explanations in computer-based learning environments. Interactive Learning Environments 10(2), 105–119 (2002)
11. Shulman, L.S.: Those who understand: Knowledge growth in teaching. Educational Researcher 15(2), 4–14 (1986)
12. Ku, H.Y., Sullivan, H.: Student performance and attitudes using personalized mathematics instruction. Educational Technology Research and Development 50(1), 21–34 (2002)
13. Anand, P.G., Ross, S.M.: Using computer-assisted instruction to personalize arithmetic materials for elementary school children. Journal of Educational Psychology v79(n1), 72–78 (1987)
14. López, C., Sullivan, H.: Effect of personalization of instructional context on the achievement and attitudes of hispanic students. Educational Technology Research and Development 40(4), 5–14 (1992)

15. Small, D.A., Loewenstein, G.: Helping a victim or helping the victim: Altruism and identifiability. Journal of Risk and Uncertainty 26(1), 5–16 (2003)
16. Feng, M., Heffernan, N.T., Koedinger, K.R.: Predicting state test scores better with intelligent tutoring systems: Developing metrics to measure assistance required. In: Proceedings of the 8th International Conference on Intelligent Tutoring Systems, pp. 31–40. Springer, Berlin (2006)
17. Pennebaker, J., Francis, M., Booth, R.: Linguistic inquiry and word count: LIWC. Erlbaum Publishers, Mahwah (2001)
18. Kincaid, J., Fishburne, R., Rodgers, R., Chissom, B.: Derivation of new readability formulas for navy enlisted personnel, 8–75 (1975)
19. Cheng, R., Vassileva, J.: Design and evaluation of an adaptive incentive mechanism for sustained educational online communities. User Modeling and User-Adapted Interaction 16(3-4), 321–348 (2006)