Creating a Corpus of Targeted Learning Resources with a Web-Based Open Authoring Tool

Turadg Aleahmad, Vincent Aleven, and Robert Kraut

Abstract—Personalizing learning to students' traits and interests requires diverse learning content. Previous studies have demonstrated the value of such materials in learning but a challenge remains in creating a corpus of content large enough to meet students' varied interests and abilities. We present and evaluate a prototype Web-based tool for the open authoring of learning materials. We conducted a study (an open Web experiment) to evaluate whether specific student profiles presented in the tool's interface increase the diversity of the contributions and whether authors tailor their contributions to the features in the profiles. We report on the quality of materials produced, the authors' facility in rating them, the effects of author traits, and the impact of the tailoring feature. Participants were professional teachers (math and nonmath) and amateurs. Participants were randomly assigned to the tailoring tool or a simplified version without the tailoring feature. We found that, while there were differences in teaching status, all three groups made worthy contributions. The tailoring feature leads contributors to tailor materials with greater potential to engage students. The experiment suggests that an open access Web-based tool is a feasible technology for developing a large corpus of materials for personalized learning.

Index Terms—Educational technology, authoring systems, personalization, technology social factors.

1 Introduction

Personalization to students' interests and identities has been shown to improve both student engagement and test scores. Fourth grade math students have higher pretest-to-posttest gains with personalized instruction and also perform significantly better on both the pretest and posttest problems [1]. Similar effects have been found with fifth and sixth grade students [2]. Personalized instruction has also been demonstrated to increase the engagement and learning outcomes of minority groups, e.g., Hispanic [3].

One challenge to the growth of personalized learning environments is that the content they present is laborious to create. Intelligent tutoring systems are very adaptive to the learner's activity, yet require 100-1,000 hours of time from skilled experts for each hour of instruction [4], [5]. Newer approaches [6] lower the total resources necessary to create a tutor, but still require careful coordination of a group to create useful tutors. Tutors such as REDEEM [7] and pSAT [8] separate logic—which requires programming—from the domain material so that nonprogrammers can customize or extend the tutor. The example-tracing feature of CTAT lowers the expertise necessary to define tutoring logic and its bulk templating feature allows simple expansion of domain material within a logic [9]. Yet, each of these require some training to use. The Assistment Builder's problem-specific authoring paradigm and Web-based interface are easy enough that novice users can develop a simple tutor for a problem in under 30 minutes [10]. Yet, all of these are limited in the dimensions by which they can personalize to the student (e.g., knowledge components learned or preference of learning style). In this paper, we describe and evaluate an open Web-based problem-specific authoring tool with a novel feature to foster personalized instruction matched to learners' interests and abilities.

The power of open authoring on the World Wide Web has been demonstrated over the last decade. 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 cohesive sources and, in general, lower their costs. Existing open authoring systems for education, such as Wikiversity or Wikibooks, create monolithic artifacts that are the same for all learners. Connexions, an open textbook authoring system, was designed to support remixing of content "modules" [11], but these are tailored to the scope of a course rather than an individual learner. The work reported here is part of a larger research program on collaborative open educational resource development around a four-phase life cycle in which system users generate, evaluate, use, and improve shared materials [12]. Here, we consider the potential for this open authoring paradigm to support individualized instruction.

Rather than encyclopedia articles or textbook modules, the artifacts created in this study are worked example problems, chosen for their value and versatility. Worked examples both instruct and help to foster self-explanation [13]. They fit easily into existing practices as an enhancement to existing intelligent tutoring systems [14], [15], as an instructional material, as a fading scaffold (by omitting some of the solution steps), or as a basic assessment (by omitting the solution altogether). A corpus of worked examples tied to personal interests and learning capacities would be a practical means of introducing personalized learning into multiple modes of use.

2 The Tool

To facilitate the creation and growth of corpora of materials, we have created a prototype Web-based authoring application designed to promote tailoring of content to learner characteristics. The version of the tool evaluated here is for worked-example problems which can be repurposed into pure assessments or instruction. The tool can also be easily adapted to make these other types of resources directly. The client-side software is built in HTML and Javascript (AJAX) and works in modern Web browsers (IE7, Safari 3, FireFox 1.5, etc.). It is running on the Web all the time and is open to anyone to contribute to at education.hciresearch.org.

In starting the tool, authors first see a page explaining what a worked-example problem is and what skill to target. This page also provides a search box to look up on the Web anything they want to learn or refresh themselves on and a table of pedagogical principles to consider in creating their worked example. When they are ready to author, they click *Continue* to reach the authoring interface, shown in Fig. 1. The tailoring feature comprises the student profile shown at the top and the text guidance below it asking the author to "Please create a worked-out example to provide practice to the student above in understanding and applying the Pythagorean Theorem." Fig. 3 shows examples of other profiles. (In the control condition of the study, the profile image and the text "to the student above" are absent.) Below the guidance information is a dynamic HTML form in which they enter their worked example. They can enter a problem statement in a large *textarea* element to the left and can add a diagram or illustration of the problem using a Flash-based drawing widget to the right. The drawings are recorded in SVG format for future programmatic manipulation and native vector rendering in advanced Web browsers. Below the problem statement is the solution table where authors enter and

• *The authors are with the Human Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, 500 Forbes Ave., Pittsburge, PA 15213.*
E-mail: turadg@cmu.edu, aleven@cs.cmu.edu, robert.kraut@cmu.edu.

Manuscript received 16 Sept. 2008; revised 28 Dec. 2008; accepted 7 Jan. 2009; published online 16 Jan. 2009.
For information on obtaining reprints of this article, please send e-mail to: lt@computer.org, and reference IEEECS Log Number TLTSI-2008-09-0089.
Digital Object Identifier no. 10.1109/TLT.2009.8.

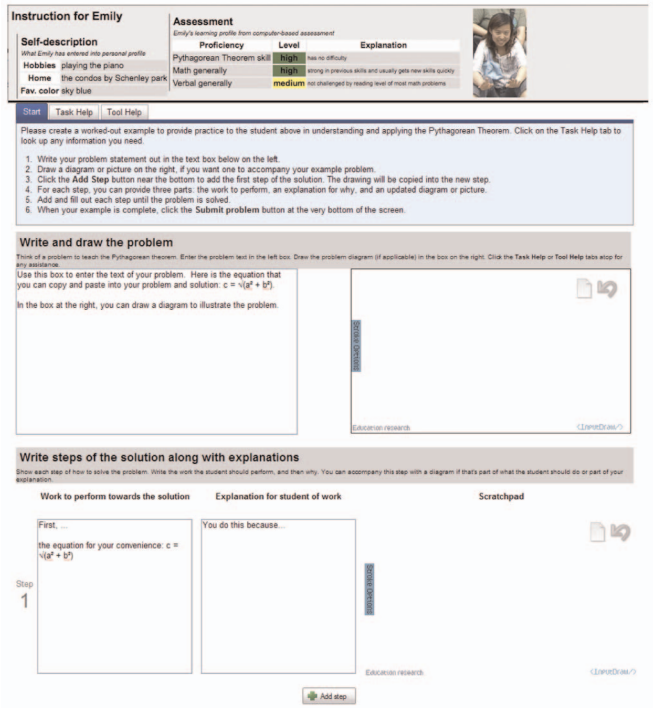

Fig. 1. Screenshot of authoring tool in profile condition.

annotate the solution steps, with columns for the work (i.e., the actual steps toward the solution), explanations of the work, and optional illustrations. Authors begin with the *Add Step* button which dynamically adds a row to the table and populates each field with starter text (e.g., "First...," "You do this because..."). Authors type out the first step of work to perform toward the solution, an explanation of why, and optionally draw an illustration. They repeat this for each step until their last, which contains the completed answer to the problem. Fig. 2 shows an example contribution authored with the tool.

Because the tool is accessible to anyone to contribute, controlling the quality of the corpus is a critical challenge. To achieve this, we have implemented (and are experimenting with) a two-pass quality check system. In the first pass, an SQL query is run to filter out any contributions that are duplicates or are not within reasonable content parameters, described below. In the second pass, humans use a simple rating tool to select the quality level of three different components of the contribution (the problem statement, solution steps, and the explanations of the solution steps) on a four-point scale specified in Table 2: Useless, Fixable, Worthy, or Excellent. The rater clicks on a button for each part to indicate its quality and then a submit button which automatically advances to the next contribution to evaluate.

3 Evaluation

We have evaluated the system in an open Web-based experiment with hundreds of contributors. To increase statistical power for the evaluation, the study controls for skill by targeting one specific skill. The skill of understanding and applying the Pythagorean Theorem was chosen for its suitability to personalization. It affords a variety of real-world scenarios to demonstrate it, providing opportunities for the author to make the problem relevant to the student. Pythagorean Theorem problems also often have a visual component, making them more difficult to generate by any automated means and thus taking advantage of the human contribution.

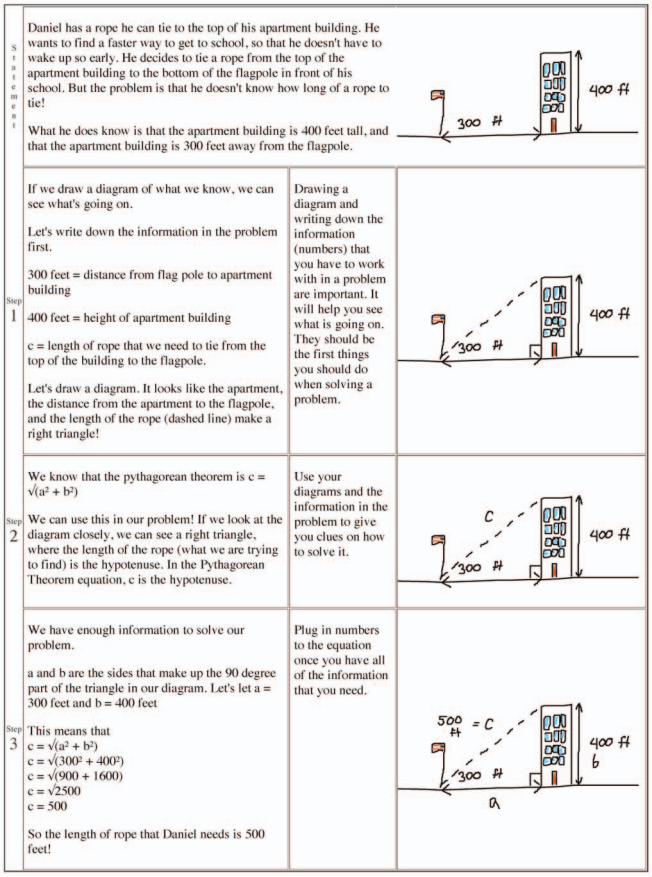

Fig. 2. Sample contribution authored with the tool.

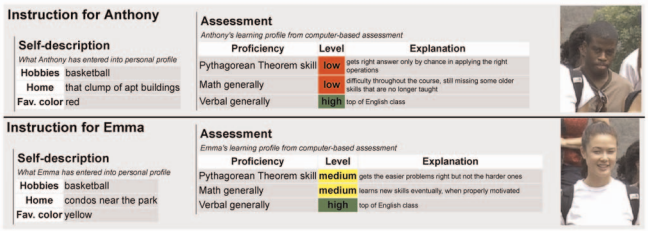

Fig. 3. Sample profiles in profile condition.

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 two hypotheses: H_F: *The software automatically filters most of the useless materials* and H_R: *Identifying the good from the bad contributions is easy with the rating tool.* To assess the impact of the authors' expertise on the quality of the contributions, we asked each participant whether they were math teachers, other teachers, or not teachers at all. We used these data to assess H_M: *Math teachers submit the best contributions*. While math enthusiast amateurs may have the appropriate *content knowledge* and nonmath teachers may have the appropriate *pedagogical knowledge*, neither will have much *pedagogical content knowledge* [16] about high school geometry.

In evaluating the tailoring feature, we hypothesized H_T: *Student profiles lead to tailored contributions*. Because being shown a specific individual to help is likely to draw out more altruistic behavior, we also expected the profiles to motivate authors [17], leading to two further hypotheses: H_E: *Student profiles increase the effort of authors* and H_Q: *Student profiles lead to higher quality contributions.*

To reliably assess the impact of the tailoring feature, participants were randomly assigned to one of two conditions. In the *profile*

TABLE 1
Count of Participants by Teacher Status and Degree of Participation

Participation level	Math teacher	Other teacher	Amateur	Total
Registered	131	170	1126	1427
Contributed	70	72	428	570
Passed filtering	25	35	220	280
Completed survey	19	30	187	236

condition, participants used the tool with the tailoring feature that presents student profiles. In the *generic* condition, this feature was removed. No profiles were shown and the words "to the student above" were stricken from the task description. In the profile condition, the profiles were varied to assess how well the feature facilitated tailoring. 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 the variation of the contributions on skill-level appropriateness. These were proficiency in the Pythagorean Theorem, proficiency in math generally, and verbal proficiency. 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 two genders. Additionally, each was assigned a favorite color to round out the description presented. Participants in the profile condition saw a new randomly selected profile for each worked-example problem they authored (e.g., one of the two in Fig. 3)

3.1 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 worked example contributions, regardless of their quality. After following the URL, they received a description of the task and a stated purpose of creating open educational materials. After consenting, they entered their e-mail, professional status, and their age. (To deter false age inputs, their was no mention of eligibility and visitors under age 18 were sent to a survey so that they would not be aware of their ineligibility.) Eligible participants would see a page describing the task in more detail and three principles of authoring worked examples. The next page presented the authoring tool. During the experiment, 1,427 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 1,130 contributions. Table 1 shows, by teacher status, the number of participants reaching each greater level of participation in the experiment.

3.2 Exit Survey

After each submission of a contribution, the participant was invited to submit another or to conclude their session with an exit survey. The survey collected information on their participation, their educational experience, their perspective on worked example problems, their regard and preferences for community authoring, and their experience using the authoring tool. Of the 570 people who made qualifying contributions, 236 also completed the exit survey.

4 Results of Open Authoring

To test H_F, the contributions were analyzed by the first-pass software filter. Of 1,130 raw contributions, 51 percent were filtered. The filtered statements were each manually coded to validate the filter. Statements that were too short (less than 50 characters) were either blank, off-topic, or overly simple like "find x." Statements that were too long (over 1,000 characters) were either proofs or contained work toward the solution and thus violated the structure. This machine filtering left 550 contributions from 280 participants and confirmed H_F, that software can automatically filter most of the useless contributions. Table 1 shows, by teacher status, the number of participants whose contributions passed machine filtering.

TABLE 2
Quality Scale Used in Coding and Analysis

Score	Quality	Definition
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.
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.
1	Fixable	Has some faults, but they are obvious and can be fixed easily, in under 5 minutes.
0	Useless	No use in teaching and it would be easier to write a new one than improve this one.

To test H_R and more fully test H_F, we looked at the quality of the remaining problems contributed and the human effort needed to classify them. In a production version of the site, human coding would be drawn from the community. For this evaluation, the two coders were a retired and a beginning math teacher. Using the streamlined rating tool described above, they each rated three parts (statement, work, and explanation) of each of the 550 contributions in a median time of 36 seconds per contribution. For further analyses, the four rating levels were assigned the integer scores 0-3, shown in Table 2. We refer to the average of ratings for the work and explanations of a given contribution as the "Solution quality" and to the average of the ratings for all three components of a contribution as the "Whole quality." Interrater reliability of the Statement quality had Cronbach's $\alpha = 0.61$ for the Solution quality $\alpha = 0.58$ and for the Whole quality $\alpha = 0.63$.

In this second-pass quality check, 23 percent of whole problems (statements with solutions) were classified as Worthy, meaning that they were fit for use immediately. Fifty-seven percent 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. Of all the statements, 55 percent were Worthy and 9 percent were Excellent as is.

To test H_M, we looked at the quality of each contribution as a whole, revealing no quality differences by teacher status ($F(2, 276.7) = 1.53, p = 0.22$). Further analysis revealed that the effect on quality of teacher status interacted with the problem component, as seen in Fig. 4.

Math teachers were best at writing problems statements, compared to other participants. A comparison across teacher status showed a marginally significant effect ($F(2, 257.2) = 2.39, p = 0.093$). 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 ($F(1, 280.5) = 4.80, p = 0.015$, one-tailed).

Contrary to H_M, amateurs were best at writing solutions. A comparison across teacher status showed a marginally significant effect with respect to Solution quality ($F(2, 287.6) = 2.73, p = 0.067$). 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 ($F(1, 283.2) = 4.87, p = .028$).

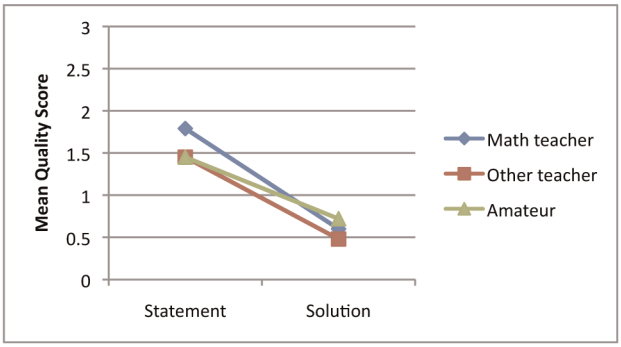

Fig. 4. Mean quality score of statement and solution by teacher status.

To better understand the teacher expertise effects, we examined more features of the participants' experience as educators. Since being a professional teacher affects quality, does being a teacher longer also? We found that while Statement quality is not correlated, Solution quality declines with years in the classroom ($0.64 - 0.03$ years, $F(1, 284) = 4.99, p = 0.026$). We also looked at years tutoring outside the classroom and found no effect. Looking at whether the author tutored at all, we found that solution quality was significantly better from people who taught *math* outside the classroom than who did not ($M_{yes} = 0.68, M_{no} = 0.54, F(1, 284) = 3.95, p = 0.048$). Last, we compared across education levels and found that solution quality differed significantly. Authors with Bachelors' degrees performed better than those with high school degrees, but each degree higher than a Bachelor's led to a decrease in solution quality ($F(4, 231) = 3.28, p = 0.012$).

5 Discussion on Open Authoring

In a short amount of time, about 1,500 people registered to contribute to a commons of educational materials. Of the raw contributions, 570 made the first-pass software filter blocked leaving 550, of which 109 were judged useless by human experts. The software filter saved human raters from seeing 84 percent ($\frac{blocked}{all\ useless}$) of useless contributions, confirming H_F. 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, confirming H_R. About one-fourth of the contributions the raters saw were ready to help students learn without needing any modification. More than half were rated as Fixable, meaning that they would be ready to use with some additional work, which, in an open system, could be performed by anyone. Statements were the highest quality parts 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 H_M, math teachers were best at authoring problem statements. Surprisingly, amateurs authored the best worked solutions. Further, the quality of solutions declined with years spent as a teacher and years spent in school (after a Bachelor's degree). This can be explained by the "expert blind spot" hypothesis [18] that the more expert someone is in a domain, the more unaware they are of the difficulties that novices have. That math teachers performed worse (on solutions) than amateurs but better than nonmath teachers adds further weight to this idea. It may be that they used their pedagogical content knowledge in geometry to help compensate (but not fully) for their expert blind spot.

Additionally, it seems that *tutors* of math outside the classroom have less of this blind spot, either through less domain expertise or greater pedagogical content knowledge. Interestingly, there was no observed difference in quality by the number of years spent tutoring, so if it is due to pedagogical content knowledge, it may develop quickly. If so, an explanation may be that a tutor gets direct feedback from a tutee on her explanation while a teacher in front of a classroom has that feedback only in the aggregate of many students, if at all.

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. Educational content systems can benefit from opening the channels of contribution to all comers.

TABLE 3
Probabilities of Contribution Matching an Attribute

Attribute	Generic, no profile (G)	Profile omits attribute (N)	Profile mentions 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

†$p < .10$, *$p < .05$, **$p < .001$

6 Results of Tailoring Feature

The tailoring feature of the tool was evaluated experimentally. To test H_T, the amount of tailoring was measured as the degree to which various attributes of the contributed problem matched those of the particular student profile for which the contribution was made. Matching took two forms: We measured the frequency of words (presumably) primed by the student profile and we evaluated to what degree the difficulty of the contribution (math and verbal) matched the skill levels in the given profile. First, we evaluated whether the frequency of words related to gender and interest (sports, TV, music, and home situation) differed depending on the corresponding attributes in the student profile for which the contribution was written. The use of words in the contribution was analyzed using LIWC, a word counting tool, with its default dictionary [19] plus the word "piano" in the music category (to go with "guitar," "instrument," "concert," etc.). Table 3 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. For example, use of a female pronoun in the problem statement was 5 percent without a profile (G) and 4 percent with a male profile (N) but 16 percent with a female profile (M). Both G-M and N-M pairs were significantly different. In contrast, a male pronoun was present in 19 percent of problems, when shown a male profile (M) or no profile (N), suggesting that authors already have a male in mind without viewing a profile.

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 contribution. The reading level was measured using the Flesch-Kincaid Grade Level Formula [20]. This formula assesses US school reading grade level for a given text, making it easy to match a worked example contribution to the reading level in a student's record. The text analyzed is the concatenation of the problem statement and all the explanation steps. Because

TABLE 4
Correspondence of Verbal and Math Skill Levels with the Authoring Interface

Verbal Skill in Profile	Sign. Diffs	Mean Reading Level	Std Err
High	A	3.78	0.24
Medium	A B	3.56	0.32
Low	B	2.93	0.33
None (control)	B	3.20	0.16

(a)

General Math Skill in Profile	Sign. Diffs	Probability of Using 3-4-5 Triangle	Std Err
High	A	16%	0.05
Medium	A B	26%	0.05
Low	B	27%	0.04
None (control)	A B	21%	0.03

(b)

The same letters are not statistically different. (a) Matching to verbal skill. (b) Matching to math skill.

readability metrics are not calibrated to math expressions, the work steps were omitted from readability analysis. Outliers were curtailed by removing the top and bottom 2.5 percent percentile in the distribution of Flesch-Kincaid Grade Level, leaving a range -1.32 to 11.71. An F-test showed the differences across profile verbal skill levels (modeled as continuous) to be significant ($F(1, 330.1) = 4.69, p = 0.016$, one-tailed). Table 4a shows the results of pairwise 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 contribution, as in the problem statement that begins, "Shakespeare sat down one day and had a revolutionary idea. He would write text diagonally across a page rather than horizontally [...]."

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 significant ($F(1, 519.3) = 3.62, p = 0.029$, one-tailed). Table 4b shows the results of t-tests between each comparable pair.

The effect of the tailoring tool on author effort was also analyzed to test H_E. It was measured by both the length of each contribution and the time spent on it by the author. Authors in the generic control condition wrote an average of 766 characters per contribution compared to 847 characters in the profile condition, a marginally significant difference ($F(1, 266) = 2.35, p = 0.063$, one-tailed). Most of that difference is accounted for by the problem statements. Participants in the profile condition wrote 23 percent longer problem statements ($M_{generic} = 204$, $M_{profile} = 250$), a significant difference ($F(1, 247.2) = 8.61, p < 0.01$). But there was no significant difference in the time spent authoring problem statements. For the solution portion, no significant differences were observed either in time spent, characters type, or steps added.

Effects on future effort were also analyzed using responses to the exit survey. The 10 five-point agreement Likert items from the Community section of the survey (Table 5 with items marked *(R)* reversed) were combined to form a scale (-2 to 2) of regard for community authoring (Cronbach's $\alpha = 0.74$). There were no main effects of the experimental manipulation, but it had a significant interaction with teacher status ($F(2, 227) = 5.88, p = 0.0032$). Fig. 5 shows that the profiles that raised math teachers' mean regard for community did not affect amateurs, and actually lowered regard for community among nonmath teachers. This interaction effect holds for each of the questions in the scale individually.

TABLE 5
Exit Survey Items on Community

If my contributions became part of an active authoring community,
I would read feedback provided by my peers.
I expect most feedback would be helpful.
I would provide feedback to other authors on their contributions.
I would take feedback into account when making more contributions.
If I agreed with some feedback, I would go back and improve the contribution.

What is the quality of what a voluntary community can produce?
If many teachers contributed to a community system, it would come to have useful materials.
The best contributions would not be as good those in textbooks. (R)
The best contributions would be better than those in textbooks.
The average contribution would not be as good those in textbooks. (R)
The average contribution would be better than those in textbooks.

Quality was analyzed by experimental condition to test H_Q. 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 contributions. (For the whole contribution, $F(1, 263.1) = 0.0, p = 0.998$.)

7 Discussion on Student Profiles

Confirming H_T, all features of the profile display were accounted for in the problems contributed. 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 the likelihood of including a female in the problem statement. Without a profile, males were used in 19 percent of problem statements and females in just 5 percent. (The rest used only "it" or no pronouns.) Female student profiles bring female pronoun usage up to 16 percent, 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 their contributions to the student's skill level in both math and reading. Contributions made for students with high and low reading skill differed in terms of reading difficulty by almost a grade level. Contributions for profiles with high general math skill level were one-third less likely to make use of simple 3-4-5 triangle problems.

Supporting H_E, participants shown profiles of students wrote problem statements that were 25 peercent longer. It is perhaps odd then that they did not 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

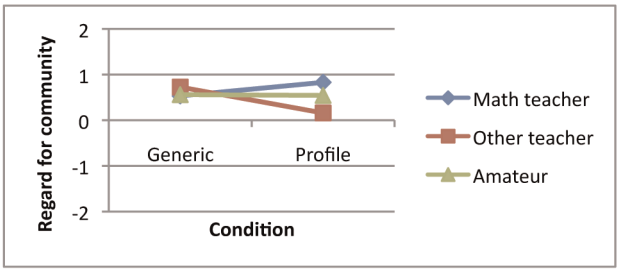

Fig. 5. Regard for community by professional status and experimental condition.

statements in the profile condition are so much longer suggests that the profile prompts ideas that are more involved.

That profiles would lead to contributions of higher quality on an absolute scale H_Q did not bear out. Instead, the contributions maintained quality. In other words, the tailoring came at no cost to the generic quality of the contributions.

The profiles did have a curious effect on a measure of regard for community, a possible indicator of future participation. Amateurs were not affected by the profiles but teachers were. The profile feature led math teachers to value peer feedback more highly and trust in the quality of community-generated learning materials. In contrast, teachers of other subjects came to think less of peer feedback and of community-generated materials. While this may be due to different dispositions of math and other teachers, it may also be simply because math teachers saw it as a valuable tool in their work and other teachers thought it distracted from theirs. The explanation for this interaction remains an open question.

8 Limitations

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 community-authored 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 had an incentive to contribute. One can imagine a future system with incentives such as peer status or competitions with nonmonetary awards (e.g., [21]). 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. It is worth noting that, since completion of the experiment, additional participants have contributed to the site without an incentive. At the close of the experiment, the Web site was disabled but at the request of people who still wanted to participate, two months later it was restored for free contributions. In the months that have elapsed, 93 people have registered and submitted 93 contributions, of which 40 pass machine filtering.

We are addressing the above limitations by creating a production system in which materials are both authored, used, evaluated, and improved. We are planning an open-source open-content platform for collaborative authoring in different domains. We will manipulate and study the extrinsic (e.g., money and social credit) and intrinsic (e.g., fun) motivations of authors and may assess the learning impact of materials.

9 Conclusions

We evaluated whether open authoring and profile-based tailoring might be a way of addressing a significant obstacle to a highly individualized instruction, namely, the fact that a large pool of differentiated instructional materials is needed.

Our first main conclusion is that the results support the feasibility of open authoring of instructional materials targeted at highly specific instructional objectives. We confirmed that quality control of the contributed materials is feasible through simple means. Automated filtering of the least valuable content was trivial, and teachers using our rating tool did not have to expend much effort to separate the wheat from the remaining chafe. Importantly, both professional educators and amateurs contributed a large portion of useful materials. Contrary to our expectation, contributions from math teachers were not superior to those from others. This finding bodes well for the viability of open authoring to support math learning because there are many more people who are not math teachers than who are.

Math teachers did write the best problem statements but amateurs wrote the best solutions. This finding suggests a model for community authoring 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. That additional tutoring experience led to greater solution quality while classroom teaching experience led to less invites the speculation that tutoring is a better way to build pedagogical content knowledge than classroom teaching is. This is worthy of further study.

A second main conclusion to follow from this work is that community authoring efforts can be directed toward producing individualized materials. The tailoring feature of our authoring tool, in which authors are shown specific student profiles, successfully led to tailored materials. The profiles led to more highly tailored materials. On every attribute, the profile increased the likelihood of targeting it, compared to authoring without profiles. The profiles also drew out slightly more effort on the part of participants. While the profiles did not measurably improve the quality of contributions, they did not impair them either. Thus, the feature provides measurable gains in individualization without measurable impairments to the quality of the contributions. The tailoring feature also perhaps increased likelihood of future efforts from math teachers by causing them to hold community authoring in higher esteem. Curiously, the tailoring feature had the opposite effect on nonmath teachers. This unexpected interaction with teaching domain suggests a factor to consider in designing and evaluating education technologies.

This study has positively, albeit partially, demonstrated the utility of a Web-based open authoring system for personalized learning resources. Participants, regardless of professional expertise, are able to make useful contributions. A relatively simple student profile feature is successful in eliciting contributions tailored to cultural (interests and environment) and cognitive (math and verbal) attributes of different learners. Thus, open authoring, combined with student profiles, helps overcome a significant obstacle to large-scale individualization of learning materials, namely, the need for a large pool of individualized materials.

Acknowledgments

The authors would like to acknowledge the suggestions of the reviewers. The photo shown in the student profile included in this paper came from Flickr user *jenrock* under a Creative Commons Attribution-Noncommercial 2.0 Generic license. This work was supported in part by Graduate Training Grant awarded to Carnegie Mellon University by the US Department of Education (#R305B040063). The research reported here was supported by the Institute of Education Sciences, US Department of Education, through "Effective Mathematics Education Research" program grant #R305K03140 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the US Department of Education.

REFERENCES

[1] H.-Y. Ku and H. Sullivan, "Student Performance and Attitudes Using Personalized Mathematics Instruction," *Educational Technology Research and Development,* vol. 50, no. 1, pp. 21-34, 2002.

[2] P.G. Anand and S.M. Ross, "Using Computer-Assisted Instruction to Personalize Arithmetic Materials for Elementary School Children," *J. Educational Psychology,* vol. 79, no. 1, pp. 72-78, Mar. 1987.

[3] C. López and H. Sullivan, "Effect of Personalization of Instructional Context on the Achievement and Attitudes of Hispanic Students," *Educational Technology Research and Development,* vol. 40, no. 4, pp. 5-14, 1992.

[4] J.R. Anderson, *Rules of the Mind.* Erlbaum, 1993.

[5] T. Murray, "Authoring Intelligent Tutoring Systems: An Analysis of the State of the Art," *Int'l J. Artificial Intelligence in Education,* vol. 10, pp. 98-129, 1999.

[6] *Tools for Advanced Technology Learning Environments,* T. Murray, S. Blessing, and S. Ainsworth, eds. Kluwer Academic Publishers, 2003.

[7] S. Ainsworth, "Using a Single Authoring Environment across the Lifespan of Learning," *J. Educational Technology and Soc.,* vol. 10, no. 3, p. 22, 2007.

[8] S. Ritter, J. Anderson, M. Cytrynowicz, and O. Medvedeva, "Authoring Content in the PAT Algebra Tutor," *J. Interactive Media in Education,* vol. 98, no. 9, p. 2, 1998.

[9] V. Aleven, B.M. McLaren, J. Sewall, and K.R. Koedinger, "Example-Tracing Tutors: A New Paradigm for Intelligent Tutoring Systems," *Int'l J. Artificial Intelligence in Education,* to appear.

[10] T.E. Turner, M.A. Macasek, G. Nuzzo-Jones, and N.T. Heffernan, "The Assistment Builder: A Rapid Development Tool for Its," *Proc. 12th Ann. Conf. Artificial Intelligence in Education,* pp. 929-931, 2005.

[11] R. Baraniuk, C.S. Burrus, B. Hendricks, G. Henry, A. Hero, D.H. Johnson, D.L. Jones, R. Nowak, J. Odegard, L. Potter, R. Reedstrom, P. Schniter, I. Selesnick, D. Williams, and W. Wilson., "Connexions: Education for a Networked World," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '02),* 2002.

[12] T. Aleahmad, V. Aleven, and R. Kraut, "Open Community Authoring of Targeted Worked Example Problems," *Proc. Ninth Int'l Conf. Intelligent Tutoring Systems,* pp. 216-227, 2008.

[13] A. Renkl and R.K. Atkinson, "Learning from Examples: Fostering Self-Explanations in Computer-Based Learning Environments," *Interactive Learning Environments,* vol. 10, no. 2, pp. 105-119, 2002.

[14] B.M. McLaren, S.-J. Lim, F. Gagnon, D. Yaron, and K.R. Koedinger, "Studying the Effects of Personalized Language and Worked Examples in the Context of a Web-Based Intelligent Tutor," *Proc. Eighth Int'l Conf. Intelligent Tutoring Systems (ITS '06),* pp. 318-328, 2006.

[15] R. Schwonke, J. Wittwer, V. Aleven, R. Salden, C. Krieg, and A. Renkl, "Can Tutored Problem Solving Benefit from Faded Worked-Out Examples?" *Proc. European Cognitive Science Conf.,* pp. 23-27, 2007.

[16] L.S. Shulman, "Those Who Understand: Knowledge Growth in Teaching," *Educational Researcher,* vol. 15, no. 2, pp. 4-14, 1986.

[17] D.A. Small and G. Loewenstein, "Helping a Victim or Helping the Victim: Altruism and Identifiability," *J. Risk and Uncertainty,* vol. 26, no. 1, pp. 5-16, 2003.

[18] M.J. Nathan and A. Petrosino, "Expert Blind Spot Among Preservice Teachers," *Am. Educational Research J.,* vol. 40, no. 4, pp. 905-928, 2003.

[19] J. Pennebaker, M. Francis, and R. Booth, *Linguistic Inquiry and Word Count: LIWC.* Erlbaum Publishers, 2001.

[20] J. Kincaid, R. Fishburne, R. Rodgers, and B. Chissom, "Derivation of New Readability Formulas for Navy Enlisted Personnel," Research Branch Report 8-75, Naval Air Station Memphis, Millington, Tennessee, Feb. 1975.

[21] R. Cheng and J. Vassileva, "Design and Evaluation of an Adaptive Incentive Mechanism for Sustained Educational Online Communities," *User Modeling and User-Adapted Interaction,* vol. 16, nos. 3-4, pp. 321-348, 2006.