# Toward Personalizing Students' Education with Crowdsourced Tutoring


**Ethan Prihar**
Worcester Polytec hnic Institute
ebprihar@wpi.edu

**Thanaporn Patikorn**
Worcester Polytechnic Institute
tpatikorn@wpi.edu

**Anthony Botelho**
Vlorcester Polytechnic Institute
abo te lho@wp i.edu

**Adam Sales**
Worcester Polytechnic Institute
asa1es@wpi.edu

eU Heffernan
Worcester Polytec hnic Inst itute
nth@wpi.edu



## ABSTRACT

As more educators integrate their curricu1a with o nli ne learn - ing, it is e asier to c rowcfaource conten t from the m. C rowd-sourced tutoring has been proven to reliably increase students' next problemcorrectness. In this work, we con.finned the find-ings of a previous study in this area, with stronger confide nce margins than previously, and revealed that only a portion of crowdsourcedcontent creators had a reliable benefit to stu-dents. Furthennore, this work provides a method to rank content creators relative toeach other, which was used tode-termine which content creators were most effective overall, and which content creators were most effecti ve for specific groups of students. When exploring data from TeacherAS-S IST, a fea ture within the ASS IS Tmen ts lear ning platform that crowdsources tutoring from teachers, wefound that while over-al l this program provides a benefit to s tude nts, some teacher& created more effective conlent than others. Despite this find-ing, we did not find evidence that the effectiveness of content reliably varied by student knowledge level, suggesting that the conten t is unli ke]y s uitab le for person alizing instruction based on student knowledge alone. These find ings are promising for the future of crowdsourced hlloring as they help provide a foundation for asses.sing the quality of crowdsourced con-tent and inve stiga ting con1ent for oppo rtuni ties 10 persona lize students· education.



## Author Keywords

Online Tutoring; Crowd Soureing; Statistical Analysis; Personalized Education





*L@5'2I, Jrnu,22 - 25,2021 Vim,d g..., Gl7ffla y.*





A CM ISBN!>°18--1-4SGW2 1S-l/2 1/0 6.
h1tps://doi0<'110.1 1451.!-41 !>-5:!'.460130



## CCS Concepts

•H um a n-ce n le red com pu tin g -t Hu1 1 1an co m pu ter in te r-a ction (H CI) ; E ln piri cJll studie'.> in H C I ;


## INTRODUCTION

1ne need for crowdsourcing within online learning platforms is growing as the user base of these platforms continues to expand and diversify [18,7]. Crowdsourcing ca n be usedeffec-tively to ge nerate new teaching materials [22] and new tutoring for students [L8]. As more platformsintegrate crowdsourcing, methods to evaluate and maintain the quality of crowdsourc.ed mate rials need to be developed to ensure students receive a high qua lity education and effective support.

In the 20 17-2018 academic year. ASSISTments. an online learning platfonn (10], deployed TeacherASSIST. TuacherAS-S lS T allowed teachers lo create Lulo ri ng i n the form of hints and explanations for problems they assigned to theirstudents. TuacherASSIST rhen redistributed teachers' tutoring ro stu-dents outside of the their class. At L@S 2020, ASSIS T ments reported that teachers created about 4-0,000 newinstancesof tutoring for abou t 26 ,00 0 d iffere nt p ro ble ms. Th ro ugh two large -sca le rand o mized controlled experiments, it was deter-mined d1at there was statistically significant improvement on the nex t prob lem co rrec tness of students who received crowd-sourced tutoring. Si nce the publication of these findings, AS-S !S T men *ts* has scaled up the dis tribution of crowdsourced content within the platform. The first part of this study uses new data,collectedfrom the 2019-2020 and 2020-202L school years to re-evaluate the fi ndi ngs of the origi nal study and con-firm that crowdsourced tutoring continues to benefit s tude nts overall.

1 1le secon d part o f this s tudy inves tigated if tl1ere was a s ig-nificant difference between the quality of different teachers' tutoring. The methodology used i.n this paper could be used in the future to determine which teacher' s content should have priority when djstributing tutoring to students in other classes.

Lastly , this study determined if there were any qualitative interactions belween the t.e.achers who created tutoring and students grouped by their knowledge-level. Personaliz.ed tea m - ing requires qualitative interactions, defined as one group of students benefiting more from one type of instruction, while a differe nt group of students benefited more from an alternative type of ins truction. TI1e learni ng science community has spent a co nsiderable amount of time investigatingtl1e impact of per- so naHzed learn i ng on s tu dents. W hil e personalized tutoring based on prior knowled ge has shown some evidence of a quali- tative interaction [20], other methods for personalization, such as teaming styles. have rarely shown co ncJusive evidence of a qualitative interaction [7 1.1k lre thod used in this study can be used to search experimental data for qualitative interactions without using a randomized c-0ntrolled trial to direc tly eval uate the presence of a particular qualitative interacLion.

Spec ifica ll y, this work seeks to address the following research questions:

I. Do tl1e find ings of the previous TeacherASS lST study s till hold whe n tes ted on new data?

2. How did lhe effec tiveness of teachers' tuto rin g co mpare 10 each other?

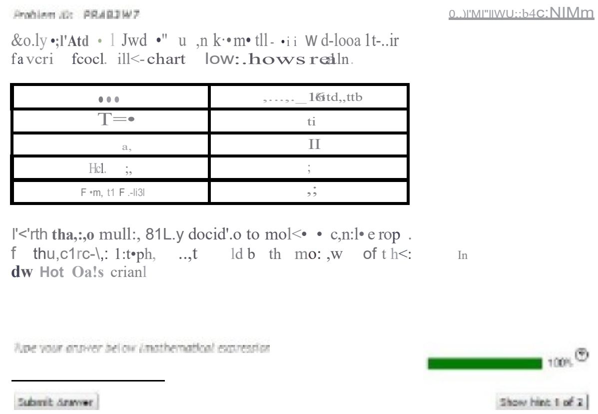
3. Was there any po te ntial to personalize the tutoring students received based on their knowledge-level?

## BACKGROUND

### The Value of Crowdsourclng

Th e growing popularity of online learning platforms has cre- ated a greater opportunity and a greater need for educational materials of all levels. With a greater diversity of :..t u de nts, the re arises the need to provide instruction to students of vary- ing skill levels. Crowdsourcing can he lp diversify the available tutoring and assist in personalizing lesson plans for s tuden Ls (25, 3). Crowdsourcing offers a mechilllis m to o btain rile breadth of ed uca tional co ntent required to meet the grow- ing demand of online tutoring. but posessome challen ges as we ll 125]. 1ne biggest risk from using crowdsourced materi- als is the potentiul for low quality, or misleadin g mate1ial to negatively impact students (25]. Even if the information is high-q ualily, o verly detailed tutoring, or tutoring from highly different sources cun also havea negative impact of students· lea.ming [23, 13. 12). Ways to mitigate these risks include algorithmically evaluating the quaHly of cro wdso u rced con- tent creators [21), or simply crowdsourcing content only from people tltat have been deemed qualified [16, 4, 24, 5J.

Even with theserisks, crowdsourcing has beena viable method for obtain ing info rma tion on the knowledge compo nents of different math proble ms [15], assis ting s tudents learn ing com - puter programming [2], and collecting videos ex.p lai ning how to solve mathematics problems (26, 27J. M o s t direc tl y. i n the s t udy preced i n g this work, tutoring messages created by teachers, for students c-0mpleting work in ASS lSTme nts, had an overall positive effec t ons tudents ' learning [18]. Although crowdsourcing has shown promising res ults in many situations, there is a need to c-0tl tinue to e val ua te th e methods through which crowdsonrced content is collected and validated so that as more educational platforms begin to incorporate c rowd- so urcing, they can do so effic ie ntly, effectivel y, a nd witho ut risk Lo studenIs.



F il(Urc l. Thi· ASSISl'm nls tuto r, as S«!'uc by u s tuck ul soM ug a 1lJUlt1- e malics problem.

### ASSISTments

1 1le data used in this study comes from ASSIST ments. AS- SIS T men ts[1] [ 11] is an online learning platfonn focuse d on empowering teachers via automating laborious tasks such as grading and record keeping of students, and provid ing insight to teache rs on their class's oommou wrong answers and miss concep tions o n assignment<s [ 11]. ASS IST me nts provides K- 12 mathematics problems and assignments from multiple open sourcecurricula for teache rs to choose from and assign to their students. Afte r an assignment has been assig ned lo students, stude nts complete tile assignment in the ASSISTments tuto r, shown in Figure I [18l. In the tutor, sludents receive immedi- ate feedback when they submit a response to a problem, which informs them if tl1ey are correct (9). For some problems, stu- dents can req L1es 1 t u tori ng, which is avai lable Lo them at any poin t during their completion of the probl e m, regardless of whether or not they have already at1empted the pr o ble m . Tu tor - ing comes in the form of hi nts. ex.plain in g how to solve parts of the problem. (11, 20J, exam ples of how to solve similar proble ms [8, 14], examples of incorrect responses to problems wi1h explunat ions of the error [14, l], and full solutions to problems [27, 26). 'Iwo examples of tutoring in ASSlST ments are shown in 2 [18).

Recently ASSIS Tme nts began a program called TeacherAS- SlST, in which tutoring was crowdso urced from teacherS in the form of written and video-recorded hints and explanations for solving middle-school math problems. ASSlSTmeuts col- le cted tulori ng created by teacher s who had alrea dy used the platform for their own clussrooms, and then provide d the crowdso urced hints and exp lanatio ns to stude nts. Dis tribut- ing these hints and e xplanatio ns le ad lo a positive impact on students ' learning (18). In this study, the data released from the TeacherASSIST study [19], new data from TeacherAS- SIST collected since the publicationof the previous study, and infom1ation on students' knowledge -level co llected from the AS SIST ments platform were used to investigate if any content creators' tutoring significantly out-perfonued o ther co ntent creator's tutoring, as well as deten nine if there we.re any q uali- latjve inLeractions be twee n content creators and students.

[1] lUps://www.ASSJSTme1 1ts.orgJ

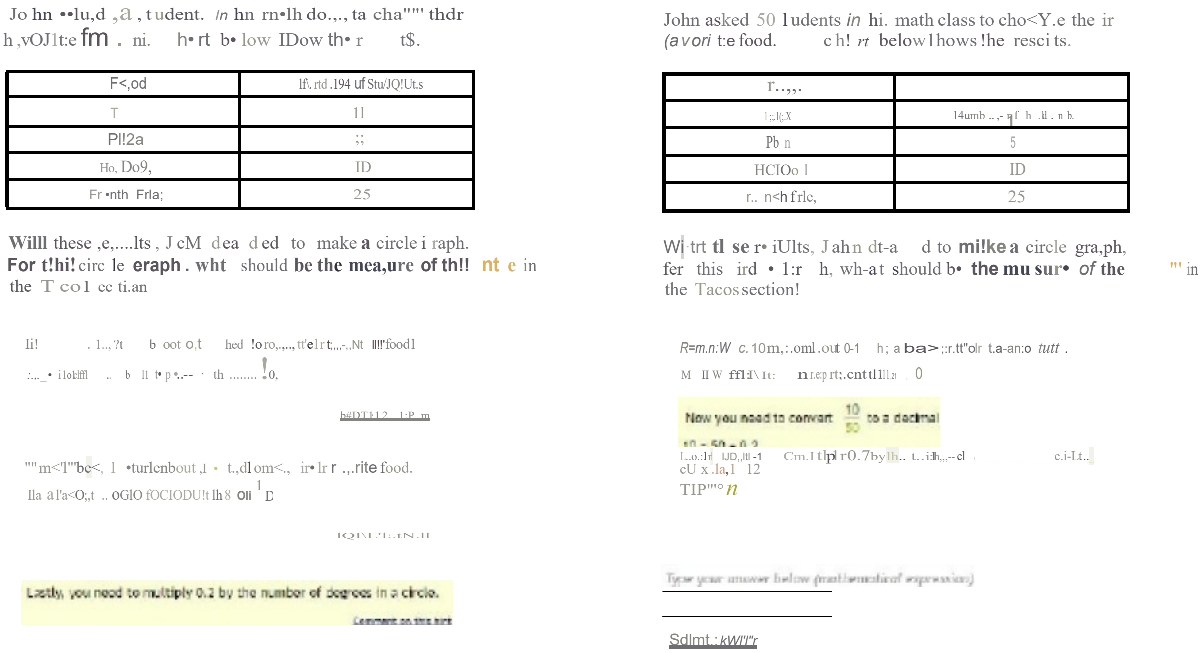



Fig ure 2. Twoi11!>1m1c.:sof l,Ulo r in g iu A SS IS [ i:o.,n ls. On lhe k-ft is a series of hinlS. On lhc right is ll filJI tlllphuli1liun of bow lo s:u lvc lhe pro blem.

## METHODOLOGY

### Confirming the Previous Study's Findings

The same analysis performed in the original study [ 18J was repealed using the exac t sa me co de from the previous study made available by the Open Science Fol.mda ti on [19]. New data,collectedsince the completion of the previous study up until February 2, 2021, was used to determine if the previously reported positive impact of TeacherASSIST was still present in a ne w academ ic year. 11)e new dataset contai ned 6,774 unique problems, 7,059 uniq ue tutoring mes sages . 18,420 unique students, and 500,900 answered problems. 50.426 of the answered prob le ms were answered by s tudentsin the control condition. where they were notgiven the option to request tutoring. and 450,474 of the problems were answered by students in the intent-to-tre at c-0ndit io n. in which they had the option to, but did not necessari ly request tutoring. A majority of students were placed in the treatment conditio n because the previous study found the treatment condition to havea reliable positive effect. and ASSlSTments did not want to prevent half the s tudents from receiving beneficial c rowds.ourced lutori ng. Of au the students in the new dataset, only 7.92% of them appearedin the inilial s1udy' s data as well.

In o rde r to gain more insigh t into how re liable the findi ngs. of the initial study were. a problem-level and stude nt-le vel inte nt-to-treat analysis, in which the students were considered to be in the trealmem condition if they were given the option to receive crowdso urced tutoring, regardless of whether or not they received it. and a treated analysis, where a student was considered to be in the treatment co ndi1ion only if they received crowds.ourced tulori ng , were performed. For all of these analyses, whic h were all performed in the initial study. Lhe Benja mini-Hoch berg procedure was used to co ntrol Lhe fa lse discove ry rate L6].

### Measuring the Effectiveness of Teachers

To determine the e ffec tiveness of each teach,er the data from the previous study and this s tudy were combined and tillered s uch that only the instances where a student received no tu-Lori ng , or cro wdsourced rutoring for the first time, and then immediately answered tmother problem remained. This step was. necessary Lo removecompounding and ex te nded exposure effects that would occur if students' nextp roblem correctness was used to evaluate the quality of teacher's tutoring after stu-denls had seen tulorin g from multiple teachers.. Furthenno re, any leachers whose tuloring was only seen by fewe r than 30 students was excluded, as there was insufficient data to mea-sure the effectiveness of these teachers. After data processing, 31,616 instances of a student getting one of 1,026 problems wrong, receiving tutoring from one of 1 1 di fferent teachers , and then answering one of 1,308 different problems were used in the foUow ing analysis.

llx: filtered data was used to fit a regression which predicted next problem correctness based on the s tudent, Lhe prob le m the student got wrong, the teacher who wrote the tutoring that the student saw upon ge tting the proble m wrong, and the next problem used to evaluatethe quality of the tutoring. Inaddition Lo a cco unting for compounding and extended exposure effects, thestudents, and the problems they co mpleted, were abstracted into sets of representativefeatures. T he features for students are shown in Tobie I, and the features for problems are shown in Table 2 These features were used in the model instead of unique identifiers for each student and problem for two reasons. Primarily. using features to represe nt stude nls and

| Student Features |
| --- |
| Total number of problems answered |
| Mean correctness on all completed problems |
| Mean time until first response on all completed problems |
| Mean time on task per problem |
| Mean number of attempts per problem |

Table 1. Features used to abstract students while preserving the effectiveness of teacher's tutoring.

| Problem Features |
| --- |
| Type of problem e.g., multiple choice, algebraic response |
| Mean correctness of all answers submitted for the problem |
| Mean time until first response for all students that answered the problem |
| Mean time on task of all answers submitted for the problem |
| Mean number of attempts of all answers submitted for the problem |

Table 2. Features used to abstract problems while preserving the effectiveness of teacher's tutoring.

problems makes it easier to generalize this procedure to other data from different educational platforms. Secondly, given the large number of unique students and problems, a model trained to predict next problem correctness would likely over-fit and obtain very high accuracy by recognizing unique combinations of students and problems, rather than estimating coefficients based on the teacher who created the tutoring given to the student, as intended.

Unlike the students and problems, teachers were not abstracted into representative features, as the goal of this process was to evaluate the effectiveness of the individual teachers, not the effectiveness of the different qualities of teachers. Teacher's unique identifiers were one-hot encoded for use in the model. In cases from the control condition, where students did not receive tutoring, all of the one-hot encoded teacher covariates equaled zero. By structuring the model's inputs this way, the coefficient of each teacher covariate measured how much more or less likely a student was to get the next problem correct after receiving tutoring from the corresponding teacher, and the probability of the null hypothesis for the covariate was the probability that receiving tutoring from the corresponding teacher was not better than receiving no tutoring at all. The probability of the null hypothesis was adjusted using the Benjamini-Hochberg procedure for controlling the false discovery rate [6] because each determination of the effectiveness of a teacher's tutoring was treated as a separate hypothesis. This model was used to determine which teachers' tutoring was statistically significantly better for students than receiving no tutoring.

### Comparing the Effectiveness of Different Teachers

In addition to using the model from the previous section to evaluate the overall effectiveness of each teacher's tutoring, the model can also be used to compare teachers to each other. Comparing the coefficient of each teacher to determine which teacher's tutoring has a larger treatment effect is, alone, not enough to confirm that one teacher's tutoring is truly more effective that another teacher's tutoring, as the standard deviation of the difference between the teachers' effectiveness could be so large that the difference between the teachers' coefficients is statistically insignificant. However, using the variance-covariance matrix, the standard deviation of the difference between two teachers' coefficients can be calculated using Equation 1, where $var(T_x)$ is the variance of teacher x's coefficient, $var(T_y)$ is the variance of teacher y's coefficient, $cov(T_x, T_y)$ is the covariance of teacher x's and y's coefficient from the variance-covariance matrix, and $\delta$ is the standard deviation of the difference between teacher x's and y's coefficients. Then, if the difference in coefficients falls outside the 95% confidence interval calculated using $\delta$, it can be concluded that the teacher with a higher model coefficient created more effective tutoring than the teacher with a lower coefficient. This technique was used to create a map of teacher effectiveness, which could be used in the future to determine which teacher's tutoring should be given to struggling students.

$$\delta = \sqrt{var(T_x) + var(T_y) - cov(T_x, T_y)} \quad (1)$$

### Measuring the Potential for Personalized Tutoring

The method described previously for comparing the effectiveness of different teacher's tutoring was also used to explore the data for opportunities for personalized tutoring. Personalizing the tutoring different groups of students receive based on the teacher that created the tutoring would only be justifiable, in this context, if three criteria are met:

1. One teacher's tutoring is more effective than another teacher's tutoring for one group of students. This can be determined using the method described in Section 3.3, using a model trained on only data from the students in the group.
2. The other teacher's tutoring is more effective for a separate group of students. This can also be determined using the method described in Section 3.3, using a model trained on only data from the other group of students.
3. Each teachers' tutoring *is* more effective than the control condition of receiving no tutoring for students in the group that benefits the most from the corresponding teacher. This can be determined using the method described in Section 3.2 on the data from only students in one group.

These criteria qualify the core assumption of personalized education, which is that in order for all students to attain the highest level of achievement they are capable of, different groups of students need to be provided with different content. If the above criteria are met, then in the future, personalizing student's educational content based on which teacher created the content would be justified. Otherwise, it would be more beneficial to give all students educational content from the teacher whose content led to the highest improvement in next problem correctness compared to the control condition. This work explored personalizing which teacher's tutoring a student received based on the knowledge-level of the student, determined by the students' average correctness.

| Dependent Mc-11s urc | Control I\lcau | Experiment J\'IC'IU\ | t-Shll | p-Valuc- |
|---|---|---|---|---|
| Conoct First Try | 0.65 | 0.66 | -1.66 | 0.10 |
| Req1A2.Ste d Tutoring | 0.20 | 0.19 | 2.61 | 0.01 |
| Stop Out | 0.01 | 0.01 | 1.08 | 0.28 |
| Attempt Count | 1.54 | 1.54 | -0.74 | 0.46 |

TwM-3, l' ru<blem-lc d palred 1-ki.t iJJMltiuu-to-lmit im lysb on s hid.cut 1lt' -prublcm dependent l'uri1tbl,l':S The- n u mbe r o f un i que- p rob lems • 5079.

| Dependent Me1 lsure | Control l\lc u u | Experiment Me1m | t-Stnl | p-Value |
|---|---|---|---|---|
| Conoct Fir..1 Try | 0.63 | 0.64 | -2.43 | 0.02 |
| Reqm ted Tutoring | 0.20 | 0.20 | 3.22 | < 0.01 |
| Stop Out | 0.01 | O.ot | -0.26 | 0.79 |
| Atte_mpt Count | *1.59* | 1.59 | 0.52 | 0.60 |

Tahll; 4. S tudenl l,;,\el p,11irt:>d1- 1 t-s l i n ll;nlio n - lo- tre al a nal ys i s on s lud;,n t n c p rublen1 depeUdl·Ut ud i1blu Thl" nunibet of 1uli4ue students • 10340.

| Dependent Mcll.!oun.· | Cootrol Mc-11JJ | E:1-.periment Mc.u1 | r-Slal | -p VulUC' |
|---|---|---|---|---|
| C o rreot FU'St *Try* | 0.33 | 0.35 | -3.09 | <0.01 |
| ReqoostBd Tutoring | 0.55 | 0.51 | 5.10 | < 0.01 |
| StopOut | 0.02 | 0.02 | .0.49 | 0.62 |
| Anempt Count | 1.85 | 1.8'i | -0.23 | 0.82 |

Tllble 5. Pto blc-m-ll •\'d pu in "<l 1-k sl t m1k d nn...tyll<is 11 11 student ucx-l probll·m del)·lndenl VRrhiblcs. The numbn ut unlq ue p rublem.s• 2524.

| Dependent le ure | Cootrol MellJJ | Experiment Mc-ml | t•S1.11 | p , VulUC' |
|---|---|---|---|---|
| C o rrect First Try | 0.36 | 0.40 | 4 .27 | <0.01 |
| Reqoo.t,Bd Tutoring | 0.5 1 | 0.4<l | 5.70 | < 0.01 |
| StopOut | 0.02 | 0.02 | .0.94 | 0.3S |
| Attempt Count | 1.93 | 1.86 | 2.54 | 0.01 |

Tuhle 6. Slutlcut-lcd pil.in-d1- ll-sl ln-al -d analysison studl'l.ltn l: l p rob-lem dc-p,:ndeul vuriltbk s. Ti ie 11umbr of unique udenls • 3547.

## RESULTS

### The Effectiveness or Crowdsou rci ng

The resu lts of this replication of the previousstudy showed the same posilive findi ngs as the previous st udy. but with better confidence. Specifically, students who received Te.acherAS-SJST tutoring were more likely to be able to solve the next proble m correctly on their first try than students in the con-trol conditio n. When studen ts who rece ived tuloring did not succeed on their firs t aue mpt., they were not more likely 10 give up or submit many more wrong answers, and they were more likely to be able to e"e ntually solve the problem without requesting more tutoring. With this ne-.v, larger dataset, rhe effec t on the treated is large enough to be detec ted wit.h sign if-icance in the intention-lo-treat ana]ysis. Tables 3 and 4 show the results of the problem-level and student-level intentio11- to-treat analysis respectively, and tables 5 and 6 show the results of the proble m-level and student-level treated analysis respec-tively. Correct first try measures thedifference in next pro blem correctness, requested tutoring measures the differenc.e in how much tutoring students' req l es.red o n the next problem after receiving tutoring from Teac herASSIST, Stop Out meas ures. the difference in students.' co mpletio n of rhe next prob lem, and Atte mpt Co unt measures the difference in how many at-te mpts students' look lo answer the nexl problem following the tutoring tiley received from TeacberASSIST. TI1e bold *p*-va lues are the significant values after correcting for multip le hypothesis lesling with Lhe Be nja mini-H oc bberg proced ure L6 1. T hese findin gs co nfinn the previo us s tudy 's co nclusion thalTeacherASSIST has an overall posith-e effecton students' teaming.

### Measuring the Effectiveness of Teachers

Using the method described in Section 3.2. The next problem correctness of students after receiving a teacher·s tutoring was compared to receiving no 1u1oring.A coefficient meas uring the impact of each teacher's tutoring on students' next prob-le m correctness, a p-valuede noting the probab ili ty that this coefficient is statis tically eguivalent IO a null trea tment effect, and the total numberof students who viewed the tuioringfrom each teacher were calculated and are shown in Table 7. If a teacher' s row *is* bold , Lhis indica tes 1Jrnt the ir tutoring had a sta tistically significant impac t on next problem coneclness after adjusting for multiple hypothesis testing.

Interestingly, e ven though rece iving cro wd ourced ttiloring had an overall poi.itiveeffeclon students' next problem oorrec.l-nes.s, only fo ur of the 11 teac hers' tutoring had a s.tatis ticaJ1y s ignificant positive effect Additionally, one teacher's tutoring had a statistically significantly negative impact on st udent's next problem correctness. This demonslr'<ltes a potential bene-fit lo evalua ting the qua lity of each conte nt creator's tutoring as it is not necessarily the case that when crowdsourcedcon-tent ii- overall beneficial, eac h co nte nt creator by themse lves is providing a benefit. Ln the future of TeacherASSlST, and in othe r c row dsourcing endeavors, only dii-trib uting co ntenl from teachers whose tulOring has a reliable positiveeffect., and tutol'ing fro m teac hers whose tutoring is s.ti U of ambiguous benefit, would likely le ad to higher nexl problem correctness for students.

| TeachreJO | View Count | Cootlicient | p-Value |
|---|---|---|---|
| No Tutoring | 2.289 | | |
| A | 95 | 0.0629 | 0.112 |
| **B** | 222 | •0.0724 | **0.1144** |
| C | 11,202 | 0.0147 | 0.118 |
| D | 5.340 | 0.0301 | **0.00S** |
| E | 76 | 0.0573 | 0.189 |
| •• | 3.671 | 0.0449 | **<0.001** |
| G | 5,76'.1 | 0.0271 | **0.008** |
| H | 911 | 0.03!16 | 0.007 |
| 1 | ,1452 | .0.0184 | 0.197 |
| *l* | 544 | 0.0046 | 0.819 |
| K | 51 | .0.0061 | 0.914 |

T11ble 7. The imp1tcl, s tat ist ic,il igulfk"llnt'.t, a11d >'k·\\ rUUlll of 1tC'b teacher's futonng 011 students· neA problem con eciness.

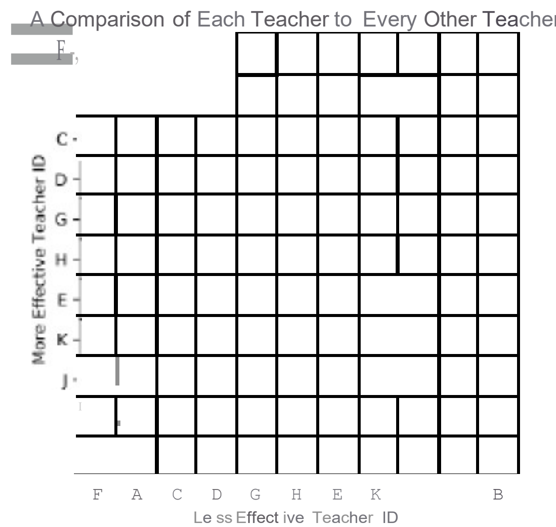
This evaluation of leachers' effectiveness co uld also be used as professiona l development for the teachers themselves. ff a teacher's tutoring is no! leading Lo a s tatistically :.ignificant increase in stu dents' next proble m correctness, the c rowd-sourcing platform could alert these teachers that their tutoring could use improvemenl and provide them with examp les of other teacher's tutoring that had been shown to be e ffec tive. Then, after the teacher updates their tulori ng , 1he platform could re-evaluate their effec tiveness and report back to the teac he r. This interaction with teachers could also encourage teacherS lha t are creating highly effective tutoring Lo create more tutoring by reporting how ma ny s tudents have received their tutoring, and to what exte nt their tutoring has helped students beyond their classroom.

### Co m p a r i n g the Eff ect ive n e s s o f Differe nt T e a c h ers

Us in g the method described in Section 3.3,the effectiveness of each teacher's tutoring wa. compared to every other ieacher's tutoring. Figure 3 Shows the instances, in gree,n whe n the tu-toring from the teacher labeled on the row. was more effective than tt-e tutoring from lhe teacher labeled on lhe col11mn. A grey cell indicates that the row teacher did not crea te more e ffec tive tuloring than the co lumn teacher. For clarity, the teachers were sorted by how many olher teachers their tutor-ing was more effective than. If all the teachers could be put in order from most to leasl effective tutoring, then Figure 3 would have entirely green cells above the diago nal. However, this is clearly not th e case. Due to **the** variancein the effectiveness of teacher·s tutoring. no teacher's tutoring is significantly better or worse than every other teachers' tutoring.

Figllfe 3 s hows so me clear examples of teachers whose tutor-ing is more effective than some of the other teachers' tutoring, for examp le, teacher F, and teachers whose tutoring is less ef-fectivethan moslother teachers' tllloring, for example, teacher B. Figure 3 also shows examples of teacher's whose varia nce in the effectiveness of their tutoring is very high, forexample, teacher K. This high variance results in no teacher significantly outperfor ming teacher K's tutoring, and teacher K's tutoring not significa nlly outperforming any other teacher's tuloring. Teacher **K** demonstrates the need to take into aroount the vari-ance of the difference between teachers' effectiveness. One cannot assert that one teacher's tutoring is more effective than another teacher's tutoriug using t11e mode l coe fficients alone.

Co mparing teacher's tutoring can be used to choose between potentia] tutoring for students when more than one option is available, but care must **be** taken. if imp lementing thisat scale, to not ignore tutoring from content creators with high variance in the effectiveness of their tutoring. **ll** could be that these con-te nt creators are new to the platfonn. and have **either** created only a few instances of tutoring, or lheir tutoring has not had a lot of exposure yet. Content cre.ators with high variance should be given the benefit of the doubl, and only when a teacher's tu-toring is statistica)]ysignificantly better than another teacher's tutoring sbould the more effective tutoring be chosen for the student When using this model to selecl which tuloring 10 give the student, the student's next problem correctness should not be included in any statistical analysis that relies onr andom sam pling.




FigllreJ. A map comparing tlteeffectivene5Sof ditl e re11t teac hers' tutor• ing.

Thache r's co uld also benefit from a plalfonn thal co mpares their effecthe nes.s to othe r teachers. For professional develop-ment. teachers could be paired with a mentor and mentee. The mentor would be a teacher with statistic.ally be11er tutoring than them, and the menl:ee wo uld **be** a teacher with siatistically wo.-se tutoring than them. This would give teacheni the oppor-tunjty to learn and teachothers, and garner community support for the platfonn. Top performers could be rewarded with no-toriely within the platform, and enco urage d lo co ntinue to make content Considering how heavily crowdsourcing relies 011 user engageme nt, working the analysis of teachers' effec-tiveness into different methods of engaging exjsting users and drawing in new users is an importan t step in the crowdsourciog process.

### **Measuring** the Poten tial for Person alized Tutoring

Lastly, using the method res cribed in Sectio n 3.4, it was in-vestigated if personalizing which leacher's tutoring s tudents received based on students' kuowledge-levels would likely have had a positive impact on sruden ts' next problem correct• ness. To groupstudents by knowledge-level, the data was split into two datasets, The high-knowledge student data contained 18,139 instances of st udents whose average correctness was above average and tile low-knowledge student data contained 1, 3 475 in lances of studenls whose average correctness was below average. To detennine which teachers met Criteria I and 2 from Section 3.4: one teacher's tuto.ring is moreeffective Lba n another teacher's tutoring for one group of sludenLs, and, the other teacher's tutoring is more effective for a separate group of sludents, the same method used in Section 3.3 was used on each grmþ of stude nls. The results of these co mpar-isons are shown in Figure 4. Figure 4 shows that there is no evidence lo support tl1e clai m thal perSOna li zing the tutoring students received would have led to an increase in next prob-le m correctness . While so me teachers, like te ac her E, were very effective for low-knowledge stude nts, and some teacher, like teacher B, were particularly ineffective for high knowledge students, there were no teac hers that met Crileria 2 and 3, in other words, the same teacher's tutoring was likely 10 have lhe highest positive impact on all students'nextproblem correctness regardless of the student's knowledge-level.

T his rigorous process used to deierrnine if there is truly a benefit to personalized tutoring could be used for more than just detemtining if student's tutoring can be personalized based on their knowle dge -leve l and who created the tutoring. This process could be used on a per-problem basis. For each problem. an analysis co uld be performed to evaluate which of the avaiJable crowdsourced tutoring messages would be most likely to positively impact students' next problem correctness based on trai ts of the students. Doing this analysis on a per-problem basis wo11ld require mllCh more data, bulasplatformsexpand and cmTicula increase their integration with online leaming. this may become a viable option. Additionally, if socioeconomic and demographic informationon studen ts is available, then this process could be used to personaliu tutoring fors tudents based on their gender or race. It is particularly important 01 pay attention to how personalizationeffects minority students. If the e ffectiveness of whatever inte rvention being deployed is being measured by how it effects au students on average, then in the same way that this stud y fo und that crowdsourced tutoring was overall beneficial. but some teacher's tutoring had a negative impact on next problem correctness, an intervention may be beneficial overall, but also be detrimental to minority sl1.1dents. Being aware of howeach groop of students is e ffected by an inte rvention will allow researchers to maintain fair interventions that help all students achieve their full potential.

## LIMITATIONS AND FIITURE WORK

Altho11gh the reslllts of this st11dy are promising, there are limitations to this wo rk. In order to co mpare teachers' tutoring, sludents and problems had to be represented with features. While these features adequately modeled students and proble ms we ll enough 10 account fo r the variations in problem difficulty and student performa nce, these features are nol necessarily the best features to use. The features used in our models could only predict next problem correctness with an ROC AUC of 0.7L It is unlikely that rhe features we had available captured I00% of the variance in pro ble ms and s tudents. and therefore including more, or different features **for** problems and students could increase the reliability in the measurements of the e ffective ness of teacher's tutoring by increasing the model's accuracy.

1n addition to potential improvements to thestudent and proble m feat ures, features for teachers could also be used to group teachers similar ro how students were grouped in Section 4.4. Features of teachers could be used Lo investigate if certain groups of teachers tend to outperform other groups and could be used for personalizationsimilarly to how individual teacherS were compared in this work. Additionally, if ce r l.ain features were indicative of a teacher"s ability to create particularly e ffect ive tu.ta rin g, this information could be used to advise teachers and other conte nt creators.

In this work, statistical analysis was used to determine which teachers' tutoring was mosteffective. While this method could be used 10 select which tutoring Lo pro vide lo s tud e nts based on which teacher is overall most effective. an online learning platfonn could also use reinforcement learning toselect which of multiple instancesof tutoring 10 provide to a stu dent based on the same features of problems and students used in lhis wor k. Contextua l bandi t algorithms L28 ] use context, which in this case are features of students and problems, to take one of multiple actions, which in this case are rhe actions o f providing one of many different instances of tutoring to a student Then they receive .i reward, which in this case would be the student's next problem correctt1ess, and adjust their decision making process lo take the actfon that *is* most likely to lead to the hjghest reward, While using a contextual bandi t algorithm prevents one from doing the same kind of experimen tal analysis perfonned in this work, it provides a method to aigorithmicaJJy detemline and offerthe best tutoring available to students.

Allhough no conclusive evidence of qualitative interactions between teachers' tutoringand students knowledgewere found in this work, the potential for persona lized learning should continue to be explored. More specific or altema tive student features co uld be creaied evaluated for q ualitative interactio ns the same way 1hat knowledge-level was used in this work. It is possible that even within the datasetused in this work, there are q ualitative interactions between groups of students that were not able lo be considered. For example, this work hadno knowledge of s tuden ts' state tes t scores , home environments, demogrnphic infonnation, or socioeconomic status. All of these factors could influence what tutoring is most effective for each student and reveal the opportunity Lo persona lize students' education.

## CONCLUSION

In this follow upstudy, providing tutoring through TeacherASSIST continued ro reliably increase students' next problem correc tness, an indication that crowdsourced lutoring within the ASS[ST1nents platform has a positive impact on students' learning. Due to many schools' recent transition to partially or fully remote leamlng, more data was available this year lhan in previous years, which allowed this study to find a reliably positive effect on students' leaming even in an inlent-to-treal analysis where not every student chose to view the tutoring available to tl1em. Furthermore, when investigating the impact of each teacher's tutoring separate ly.only four of the 11 teachers had a reliably positive impact on students, andone teachers' tutoring had a reliably negative impact. This findi ng could be used in the future to select which teacher's tutoring to provide to students based on how reliable a teachers' tutoring has been in the past Asonline tutoring platforms grow and continue to incorporate crowdsourcing techniques, it will be important to inc lude metrics for evaluati ng the quality of crowd.so urced materials and the means lo algorithmicallyselect the most effective content. As the corpusof crowdsourced tutoring grows, the most effective content can also be explo red for similarities to each other. Empirically evaluating whal makes tutoringe ffective has the poiential to improve curre nt methods for creating tutoring, andenhance ting pedagogy.

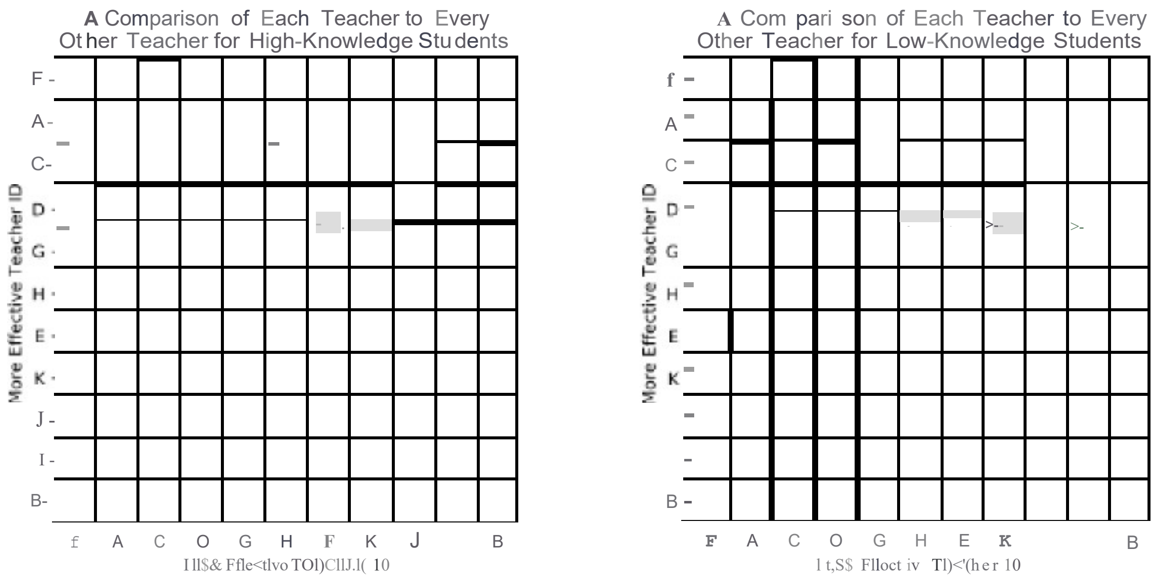



Fig ure 4. A mu p ccmiparlng lhc d Tl ·tivl'!l.eSS of dilTcre ut M 1d tcrs' tut w-in i= Stpuratelf for higt1und low kil<Mled ge s tud i:n ts.

Alth o 11g h n o evide nce of the benefit of perSonal ized ed uca tio n wa s fou nd in lllis st udy, Lhere is slill lhe polential for other quali ties of tutoring and the stude nts tha t receive the tuto ri ng to have an impact on what kind of tutoring is mosl e ffec tive . Futurework can explore for more opportunities to personalize students' education using the same me thod in th is s tudy , or look lo contexlual bandit algorith ms to find opportunities for personalization. Tlu ough continued efforts, crowdsourcing has the po te ntial to advance pedagogy and provide students with a more equitable education.


### ACKNOWL EDGMENTS

We would like to rhank multiple NSF grants (e.g., 1917808 , 19 31523 , 194 023 6, 19 177 13, 1903304 , 1822 830, 17592 29, 1724889, 163678 2, 15354 28, 1440753 . 13167 36, 125 2297 , 1 10948 3 , & DRL- 103 1398) , as well as th e US Depar tment of Educa tion for three different fund- ing lines; a} the lnstinite for Education Sciences (e.g., IES R305A l70 137 , R305 A l70243 , R30SA 18040 1, R305A120125, R305A180401, & R305Cl00024), b) the Grad- uate Assistance in Areas of National Need program (e.g., P200Al80088 & P200Al50306). and c) the EIR. We also thank the Office of Naval Research (N00014-18- l-2768), Sc hm idt F ulures, an d a n ano nymous philanthropic founda- tion.


### REFERENCES


(1) Deanne M Adams, Bmce M McLaren, Kelley Durkin, Richard E Mayer, Belhany Rittle-Johnson. Se iji lso tani, and Martin Van Velsen. 2014. Using erroneous examples to im prove mathematics learning with a web-based tutoring system. *Computers in Human Behavior* 36 (2014), 401-4 11.

12] Dhiya Al-Jumei ly, Abir Hussain, Mohammed Alghamdi, Chelsea Dobbins, and Jan Lunn. 2015. Educational cro wd.-;ourc i ng to support Lhe le,arni ng of co mputer programming. *Researchand practiceir1 techllology enhan ced I.earn in g* 10, I (20 15), 1- 15.

[3] Car los Eduardo Barbosa. Vanessa Janni Epelbaum, Marcio Antel.io, Jonice Oliveira, and Jano Moreira de S ouza. 2013. Crowdsourcing environments in E-lear ning scenario: A classification based on educational aud collaborat io n criteria. In *20 /3 IEE Elnternatio,ial Conference on Systems, Man, and Cybernetics.* IEEE. 687- 692.

[4] Tiffany Barne s and John Stamper. 2008. Toward automatichin t generation for logic prooftutoring using historical student d ata. In *Intemariona.l Conference on Imelligent TutoringSystems.* Springer, 373- 382.

LSI Ti ffa ny Hames and John Stamper. 2010. A utomatic hin t generation for logic proof tutoring using historical data. *Journal of &11,cationalTecli1wlogy& Society* 13. 1 (2010 ), 3.

[6] Yoav Be njamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple te sting. *Journal of the Royal statistical society: series B (Metlw dological)* 57, 1 (J995 ), 2 89 - 3 00 .

[7J Anthony F Botelho and Ne il T Heffernan. 2019. Crowd. ourcing Feedback to Suppor t ToacherS and Studeols. *Design Recommer1darions for Imelligent Tutoring Systems 7* (20 19 ), 101- 108.

[8] Tessa HS Eysink Ton de Jo ng, Kirslen Berthold. Bas KoUoffel. Maria Opfermann,and Pieter Wouters. 2009. Learne r performa nce in multimedia learning arrangements: An analysis across i1 1stn1 ctio nal approaches. (2009).

19] Mingyu Feng and Neil T Heffernan. 2006. Informing teachers live about student learning; Reporting in lhe assislrnenl syslem. *Technologyh istruction Cognition andLeam{ng* 3, 1/2 (2006). 63.

[ to] Neil T Heffernan and Cristina LindquistHeffernan. 2014a. T he ASSlSTments ecosys tem: Building a platfonn that brings scienlists and teachers togelher For minimaJly invasive research on human le arni ng and teachfag. *ltuernarional Journal of Artificial lmelligence in Educat ion* 24, 4 (2014), 470-497.

[I I] Neil T Heffernan a11d C ristina Lindquist He ffernan. 2014b. The ASSlSTments Ecosystem:Building a platfom1that brings scientists and teachers together for minimaJly invasive research on humanlearning and teachfag. *lmernarional Journalof Artificial lmelligence in Educariorl*24, 4 (2014), 470-497.

1121 Raymon d W Kulhavy, Mary T White , Bruce W Topp, Ann L Chan, and James Adams. 1985. Feedb ack co mplexity and corrective efficiency. *Contemporary edll carionai psychology* I0, 3 (198 5), 285-291.

1131 Mary Ellen Lepionka. 2oog _ *Writirlg a11d develop1i1g your college 1e.x1book: a comprehensive guide to textbook authorship and high.er edl,cation publish1i1g.* AHantic P.ath Publishing.

[14] Bruce M McLar,en Tamara vIIJI Gog, C raig Ganoe, Michael Karabinos, and David Yaron. 2016. The efficiency of worked examples compared toerroneous examples, tutored problem solving, and problem solving in computer-based learning environments. *Computersin Himum Behavior* 55 (2016), 87- 99.

(15) Steven Moore. Huy A Nguyen, and John Stamper. 2020. Evaluating Crowdsourcing andTopic Modeling in Generating Knowledge Components from Explanations . In *l me rnatior1t11 Co n ference on An ificial lrrrelligenc e in Education.* Springer. 39S--4IO.

[16) KorinnOslrow and Neil Heffernan. 2014. Testing the multimedia principle in the real world: a comparison of video vs. Text Feedback in authentic middle school math ass ignments. In *Educational DaJa M{ni r,g2014.*

[17) Hamid Pashler. **Mark** McDaniel, Doug Rohrer, and Robert Bjork. 2008. Learning styles: Concepts and evidence. *Psychological science inthe public interest 9,* 3 (2008), 105- 119.

[18) Thanaporn PaUkorn and Neil T Heffernan. 2020a. Effect.iveness of Crowd-SourcingOn- Demand Assistancefrom Teache rs in Online Learning Platforms. In *Proceedin gs of the Seventh ACM Conference on Leaming at Scale.* 115- 124.

1191 Thanaporn Patikorn and Neil T. Heffeman. 2020b. Release of TeacherASSIST Dataset # I. (2020). DOI: **http:/ / ch.** doL **org/ lt . 17685/ 0SF. 10/ ECPSF** Accessed: 2020-05-15.

[20J Leena M Razzaq and Neil T Heffernan. 2009. To Tutor or Not to Tutor: That is the Question.. In *AJED.* 457-464.

I 2 11 Paul Ruvolo, Jacob Whitehill, and Javier R Move llan. 2013 . Exploiting commonality and interaction effects in crowdsourcing tasks using late nt factor models. In *Neural lnformatio,i Processing Systems. Workshop on Crowdsourc1i1g: Theory, Algori1hms a11d Application.s* Citeseer.

[22) Catharyn C Shelton and Leanna M Archamba ult. 2019. Who are onlineteacherpreneurs and whatdo they do? A surveyof conten t creators on TeachersPayTeachers.com *Jounzal of Research on Technology in Edl,cation* SL, 4 (2019), 398-4 14.

[23J Valerie JSh ute. 2008 . Focus on formative feedback . ***Review****of educational research*78, I (2008) , 153- 189 .

[24J John Stamper, Michael Eagle, Tiffany Barnes, and Marvin Croy. 2013. Experimental evaluationof automatic hint generdlion fora logic tutor. *llllem aJio ,ral Journalof Arrificial Intelligence in Educatio11* 22, 1-2 (2013), 3- 17.

l25J Daniel S Weld, Eytan Adar, Lydia B ChiJton, Raphael Hoffmann, Eric Horvil.2., Mitche ll Koc h, James A Landay, Christopher H Lin, and Mausam Mausam. 2012. Personalized Online Education-A Crowdsourcing Challenge.. In *HCO MP @ AAAI.* Cil.eseer.

[26) Jacob Whitehill and Margo Seltzer. 2017. A Crowdso urcing Approach 10 Collecl jng Tulorial Videos-TowardPersonalized Leaming-al-Scale. In *Proceedings ofrhe Founh (2017) ACM Co11fere11ce* 011 *Learning@ Scale.* ACM, 157- 160.

[27J Joseph Jay Williams, Juho Kim, Atma Rafferty. Samuel Maldonado, Kn ysztofZ Gajos, Walter *S* Lasecki, and Neil Heffernan. 2016. Axis: Generating explanations at scale with leamersourcing and machine learning. In *Proceedi,igs of the Third(2016)ACM Conference on Leaming@ Scale.* ACM, 379- 388.

[28] Li Zho u. 2015. A survey on contextual multi-anned bandits. *arXiv prepri111a r Xiv:J508,03326*(2015).