Context-Based Personalized Predictors of the Length of Written Responses to Open-Ended Questions of Elementary School Students

Roberto Araya, Abelino Jiménez and Carlos Aguirre

Abstract One of the main goals of elementary school STEM teachers is that their students write their own explanations. However, analyzing answers to question that promotes writing is difficult and time consuming, so a system that supports teachers on this task is desirable. For elementary school students, the extension of the texts, is a basic component of several metrics of the complexity of their answers. In this paper we attempt to develop a set of predictors of the length of written responses to open questions. To do so, we use the history of hundreds elementary school students exposed to open questions posed by teachers on an online STEM platform. We analyze four different context-based personalized predictors. The predictors consider for each student the historical impact on the student answers of a limited number of keywords present on the question. We collected data along a whole year, taking the data of the first semester to train our predictors and evaluate them on the second semester. We found that with a history of as little as 20 questions, a context based personalized predictor beats a baseline predictor.

Keywords Written responses to open-ended questions · Online STEM platforms Text mining · Context based predictors

R. Araya (✉) · A. Jiménez · C. Aguirre
Centro de Investigación Avanzada en Educación, Universidad de Chile,
Periodista Mario Carrasco 75, Santiago, Chile
e-mail: roberto.araya.schulz@gmail.com

A. Jiménez
e-mail: abjimenez@cmu.edu

C. Aguirre
e-mail: carlosaguirre@automind.cl

A. Jiménez
Department of Electrical and Computer Engineering, Carnegie Mellon University,
Pittsburgh, PA, USA

A. Sieminski et al. (eds.), *Modern Approaches for Intelligent Information and Database Systems*, Studies in Computational Intelligence 769,
https://doi.org/10.1007/978-3-319-76081-0_12

1 Introduction

The study of the effect of the questions posed by the teacher is a subject of great importance in the teaching practice and in the preparation of teachers. Already in 1912 [1] emphasized the realization of questions as a fundamental component in teacher training. In recent years there has been an explosion of applications and educational platforms that ask questions of students. However, unlike the oral questions posed by the teacher in the classroom, the platforms in the summative and formative assessment modules mainly perform closed-ended multiple-choice questions. This is undoubtedly a great tool that allows the teacher in real time to know the progress and achievement of their students. The analysis of this type of answers has also made possible to construct computational models of the students. These models estimate the state of knowledge of each student and degree of mastery of the concepts and procedures to be taught. However, there is still little use of technology to analyze written responses to open-ended questions. While the platforms contain the possibility of introducing open answers, the analysis is still very basic. Written responses collected from students contain a wealth of information that cannot be captured by simply analyzing the answers to questions with multiple options. Here lies a huge potential for gaining a deeper understanding of student learning.

Written answers to open-ended questions depend on many factors. One of them is how the teacher formulates them. Recent advances in language technology open up a great opportunity to understand the impact of open teacher questions. So far the analysis is done mainly by hand. But this is very slow and represents a heavy workload for the teacher. The challenge then is to build tools that help the teacher analyze the answers students write on online platforms. This would allow us to get closer to knowing in real time what each student thinks and how he responds in written form. It will also allow the teacher to know how to ask to motivate their students so that they write answers that are really informative, and avoid generating extremely brief answers.

Written answers to open-ended questions also add a great value that is different from what oral answers deliver. Unlike oral answers, where typically the teacher only gets the answer of one or two students for each question, receiving written answers will get the answers of all the students. In addition, answers on a platform capture independent responses, as students do not listen to the other answers and therefore do not rely on what others say. That is, we capture personalized information that allows us to better estimate the distribution of learning. Additionally, written response reduces the problem of inhibition that complicates many students in having to speak and to respond publicly. Another advantage of written answers is a technical advantage. Greater fidelity is achieved to capture student response compared to oral responses, as no personal microphone or speech recognition system is required.

Moreover, written answers are a powerful tool that promotes learning. Students need to think more carefully about what they will respond to. The written response

requires a planning process and then a writing process. According to [2], writing, unlike speech has several characteristics that differentiate it and are very important in learning. Writing is a learned and artificial behavior, speech is natural and spontaneous. The written response is a technological tool, much slower (paused), and also more rigid than the oral response. Writing does not have an audience because one responds alone, no one is listening, and since it generates a tangible and permanent product, it is also much more responsible. Since writing is a representation that makes one's ideas more visible, writing requires much more work and therefore, according to [2], it achieves more learning than verbal responses. Moreover, according to [3] "Written speech is considerably more conscious, and it is produced more deliberately than oral speech …. Consciousness and volitional control characterize the child's speech from the very beginning of its development." The Common Core Standards in Mathematics contains Standards for Mathematical Practice "Construct viable arguments and critique the reasoning of others". It specifies that students should justify their conclusions, communicate them to others, and respond to the arguments of others. Writing is not only a powerful communication tool. It is a powerful reasoning tool as well. According to [4], "until I read what I have written, I do not see the holes in my logic, the missing steps, or the rambling thoughts" p. 4.

There are several reasons why it is very interesting to estimate the length of students' written responses. First, the teacher needs to know if his question is being answered and the impact he is achieving. It is ideal for the teacher to have this information in real time and thus allows him to adjust the next question. A central question is whether the question provides information that allows knowing the reasoning of the student. To achieve that, the teacher should avoid short answers such as Yes-No, True-False, one word answers or very short answers like "I do not know". For this there are several indicators that the teacher could obtain immediately: detection of very short answers, comparisons with the length of the answer to other questions, if it promotes more writing than the previous questions, comparisons of the length of the answers of students who are academically strong with those of the academically weak, if different impacts are observed according to gender, comparisons with the answers of the same question in other courses and/or levels. For example, [5] records and studies the length of student responses, concluding that with sufficient waiting time the length may increase between 300% and 700% [6].

Second, when preparing questions, it would be very useful for teachers to have a simulator tool that allows them to predict the effect they will have on the course. It would be ideal to have a predictor as accurate as possible. This would help the teacher examine alternatives and choose more effective questions.

Third, estimating the effect of questions on the length of responses is a basic component in estimating other indicators of the effects of teacher questions on student responses. For example, to estimate rates such as the number of key concepts per word, and the number of positive or negative words per word. In the literature of automatic text analysis, length is also critical feature of text complexity. For example, readability algorithms, like the Flesch-Kincaid Grade Level metric,

use length of sentences [7]. Moreover, predictive indices of essay quality are also related to length, such as syntactic complexity (as measured by number of words before the main verb) [7].

Fourth, predicting the length of answers is a powerful tool for teacher improvement. It allows the teacher to make a retrospective analysis of the type of questions he or she has been asking. According to [8] "By analyzing your questioning behavior you may be able to decrease the percentage of recall questions and increase the percentage that requires students to think". Having the length of the answers could be an essential component in building a system that automatically pre-classify the questions. There are several other taxonomies for questions [9], and for each of them it would also be interesting to predict the length of their answers for different categories.

Fifth, having a predictor is a tool for testing teachers' beliefs. According to [8] "Teachers sometimes think that if they begin to question why, explain, compare, or interpret, they are automatically encouraging their students to perform divergent or evaluative thinking operations." Is this really so or are there other factors involved?

Sixth, longer answers are indicative of a dialogical discourse of the teacher, and shorter answers are indicative of a more authoritative discourse. For example, Chin [10] says that "Classroom discourse can be analyzed in terms of its authoritative and dialogic functions [11]. In authoritative discourse, the teacher conveys information; thus, teacher talk has a transmissive function. Teacher talk often involves factual statements, reviews, and instructional questions; and students' responses to the teacher's questions typically consist of single, detached words. On the other hand, in dialogic discourse, the teacher encourages students to put forward their ideas, explore and debate points of view."

What does the teacher gain by obtaining a predictor of the length of the response per student and not only for a whole class? The study of written responses of whole classes provides important information on the impact of the teacher's questions [12]. However, by capturing each student's data and their previous behavior on similar questions, the estimator can be much more accurate and personalized. The student's historical behavior may give more information than generic information such as belonging to a course, the subject matter or content of the question, the student's gender and academic performance. There are many other components that influence the length of the responses that are not captured in those variables. These are factors that belong to the student. Knowing how the student responded when there are certain keywords can be very important and different from the response that other students of the same course, same gender and academic performance have had. For example, if in several previous questions where the word "airplane" was written a student writes long answers, then surely that same student will write a long answer in a next question with the word "airplane". This may not be the case in another student who is not motivated by the theme "airplanes". This greater accuracy and personalization is an important gain if at one point the teacher wants to stimulate more a certain student or a certain specific group of students. For example, a group that is falling behind, or a disinterested group, or a group of

students with particular interests. On the other hand, when preparing their questions, the teacher could have an estimate per student of the effect that will have.

It is also a tool that allows the teacher to predict not only the average behavior of the class but the entire spectrum. That is, this type of tool could predict the dispersion of responses and, even more, predict the distribution of class responses. It is also important to distinguish between predicting the distribution of response of a generic class and the distribution of response of a specific class. By having a custom model, you can predict the distribution of personalized responses per class. This ability to estimate and predict considering the personal history of each student and even more so in real time is exclusive of computer systems. It is something that requires the use of a platform.

2 Methods

In this article, we report the analysis of a year's use of the open-ended question feature of Conecta Ideas, an online cloud-based STEM platform [13]. This platform was used in 13 low-SES elementary schools in Santiago, Chile. Students attended lab classes twice a week. The platform was used both for math and science classes. Teachers and lab coordinators tracked the students' progress in real time using their smartphones. A real time, early warning system highlights which students are having most difficulty completing the tasks. The Conecta Ideas platform also includes features that encourage the more advanced students to cooperate with and help their peers. Students who receive support from the teacher, lab coordinator or their peers rate the quality of the help they receive. If the majority of students are having difficulties, the teacher can freeze the system and explain certain key concepts.

In each session, the students solved between 10 and 30 multiple-choice questions. The schools had been using this platform for 5 years. Since the beginning of 2015 the teacher has also been able to pose one or two open-ended questions during each session, in addition to the multiple-choice questions. The goal of this was to have students reflect on the contents and to encourage student metacognition. The open-ended question feature was implemented in 2015 and was completely new for teachers. However, not all teachers started using it immediately. Even though it was introduced as a metacognitive tool, the teachers initially started by asking mainly calculation questions or questions that only required the students to identify a fact. Following a couple of meetings reviewing questions together and discussing alternative ways of promoting reasoning and argumentation, the teachers started to change the type of questions they asked, in favor of truly open-ended questions. An example of a question is "Pedro has to buy 4 pencils, each one costs 150. How much money did he spend? Explain how you arrived at the result". And an example of an answer is "I summed to get to the result and the result is 600". We present the analysis of the data gathered during 2016. We present evidence from 25 classes. These classes ranged between fourth and eighth grade and were all from 12

low-SES schools based in Santiago, Chile. We define several models to predict the length of each student's responses. All models use the Q & A information for the first semester to predict the length of answers to questions to be asked in the second semester. The Baseline model for a student in a subject is the average length of all that student's responses in that subject in the first semester.

The following models are introduced with the goal to reduce the universe of questions characterizing them with only 200 words. The predictors are based on the list L of the 200 most frequent non-stop words in the questions posed on first semester. For example, the 20 most frequent words are: apple, tens, which, Pedro, chocolates, explain, why, obtain, answer, how many, hundreds, flowers, María, how much, subtract, candies, cover, trays, mathematics, Francisco. The list L defines the context component of the predictors. For each word w of the list L and for each student s, the average length of all answers to questions in the first semester containing that word is calculated. We call this number the *impulse response* of that word for that particular student and we write it down as $h(w, s)$. If for a word a student did not answer any questions in the first semester that included that word, then the value of the impulse response of that word for that student is No_Info. Thus, every student is characterized by 200 impulse responses.

The Context Based-1 model $CB1(q, s)$ predicts for each student s the length of the answer to each question q posed on the second semester to the student. This prediction is obtained by computing the impulse response $h(w, s)$ for each of the words w within the subset $L(q, s)$ of the words that are in the question q that belong to L, and that are in questions answered on the first semester by s. Then Context Based-1 model $CB1(q, s)$ is defined as the average of these impulse responses. That is

$$CB1(q,s) = \frac{1}{Card(L(q,s))} \sum_{w \varepsilon L(q,s)} h(w,s) \tag{1}$$

If for a question all the impulse responses are No_Info, then *CB1* is No_Info. The *CB1* model could be better than Baseline since it uses more historic data. It is important to emphasize that this predictor does not use any data for the semester 2 at all.

The second context based model is the Context Based-2 model *CB2*. For each student s and question q, $CB2(q, s)$ also uses the impulse responses of all the words on the question that are within the list of the 200 most frequent words on the first semester. Context Based-2 is defined as the weighted average of these impulse responses. The weights used are the same as the frequency of those words in the first semester. It is then divided by the sum of the active weights. That is, for each student s and every question q, *CB2* is always a convex combination of impulse responses. Thus if we denote $f(w)$ the frequency of the word w on all responses on the first semester, then

$$CB2(q,s) = \frac{\sum_{w \varepsilon L(q,s)} f(w) h(w,s)}{\sum_{w \varepsilon L(q,s)} f(w)} \tag{2}$$

If for a question all the impulse responses are No_Info, then *CB2* is No_Info. *CB2* could be better than Baseline and Context Based-1 since it uses more historic information. This predictor does not use any data from the second semester.

The third model is Context Based-3 model *CB3*. This estimator is similar to *CB2*, but the weights are different. They are the Kolmogorov-Smirnov (KS) statistics of the impulse responses of the words *w* within the list of the 200 most frequent words on the first semester. They are obtained by classifying students according to their average length of response on the second semester. Two classes are defined: students with long answers and students with short answers. Students with long answers are those with average length response above the median of all students' average lengths. The rest are students with short answers. For a given word *w,* the Kolmogorov-Smirnov statistic *KS(w)* is the maximum distance between the empirical cumulative distribution of the impulse response *h(w, s)* of the students *s* with long answers and the empirical cumulative distribution of the impulse response *h(w, s)* of the students *s* with short answers. *KS(w)* ranges from 0 to 1. Words *w* with *KS(w)* close to 1 are words w in questions that can discriminate between students whose answers are long from those that are short. As in the previous predictor, for each student and for each question, *CB3* is a convex combination of impulse responses. Thus,

$$CB3(q,s) = \frac{\sum_{w \varepsilon L(q,s)} KS(w) h(w,s)}{\sum_{w \varepsilon L(q,s)} KS(w)} \tag{3}$$

If for a question all the impulse responses are No_Info, then *CB3* is No_Info. *CB3* might be better than Baseline, *CB1* and *CB2*. This predictor does use data from semester 2, so that to build it there would have to be separate basis of construction and testing. However, for years to come the model could use the KS of previous years.

The fourth model is the Context Based-4 model *CB4*. This model is obtained by searching for weights *p(w)* of the words *w* for the linear combination of impulse responses and an intercept *b* that minimizes the mean square error between the model prediction and the responses of all students in a sample of questions from the second semester. Thus

$$CB4(q,s) = \mathrm{b} + \sum_{w \varepsilon L(q,s)} p(w) h(w,s) \tag{4}$$

It is important to stress that weights *p(w)* and intercept *m* do not depend on each student or question. This model uses information from the second semester, and therefore the weights and intercept are calculated in a construction sample. It then sets its performance in an independent test sample.

To evaluate the performance of the models we use three metrics. First, we use the mean quadratic error to the questions of the second semester belonging to the test base. This is the average of the squares of errors in the length of responses between what the model predicts and the actual length. Second, we also use a less demanding metric that only considers the average length of each student's responses in the second semester. This means that we evaluate the model according to whether it is able to approach the average length of response, and no response per response. This metric is the square root of the average of the quadratic errors of the students, and where for each student the error is the difference between the predicted average of the lengths of the student's answers and the average of the real lengths of his answers. Third, we use the Kolmogorov-Smirnov distance metric. In this case we only measure the ability of each model to classify students according to their average length of response in the second semester will be below or above the median of the students averages. That is, if it predicts that the student will be a student with short answers or long answers.

3 Results

In each of the three metrics we will analyze the behavior of the metrics in the subpopulation of students who have answered at least a certain amount of questions in the first semester. The reason for this is that we expect that for those students with more historical information, that is to say that they have answered more questions in the first semester, then the models should make better predictions and thus errors should fall.

Figure 1 shows the square root of the mean square error for all the second semester questions in the test sample. It is observed that for students with the highest number of questions answered in the first semester the errors of the models improve to less than half of the error for all the population. These differences are in some cases statistically significant, but due to the low number of students who have answered many questions in the first semester then it is necessary to consider a larger population. Since the Baseline, Context Based-1, and Context Based-2 predictors do not use second-semester information, we can use the full base to make error estimates.

Figure 2 shows the mean square error in the total base and shows that Context Based-2 has smaller error than Baseline for students with more than 20 questions.

Figure 3 shows the evolution of the second metric. This metric is the prediction error in the averages of responses per student. It is again observed that for students with the highest number of questions answered in the first semester, the errors of the models improve over the Baseline-1 model, but now the error of the optimal model does not improve over others. This is because we are now observing a different metric. We do not measure the difference response to response but the difference of averages. These differences are in some cases statistically significant, but due to the

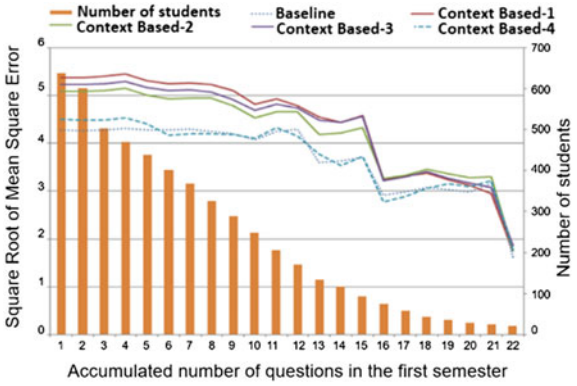

Fig. 1 Square root of the mean square error by question on the testing set

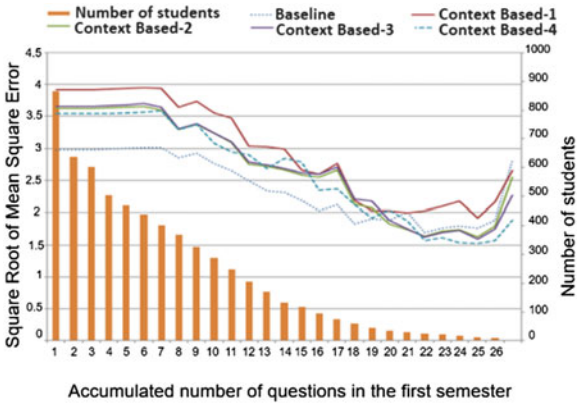

Fig. 2 Square root of the mean square error by question on the complete data base

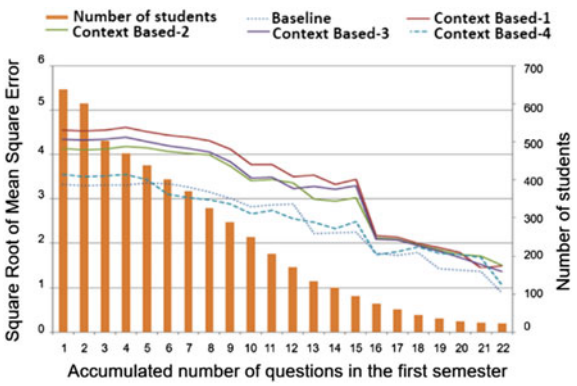

Fig. 3 Square root of the mean square error by student on the testing set

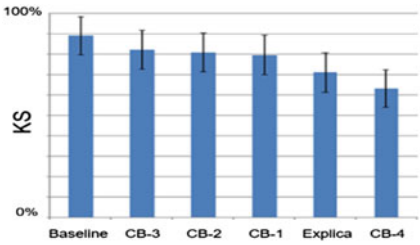

Fig. 4 Kolmogorov-Smirnov (KS) distance by student on the testing set

low number of students who have answered many questions in the first semester then it is necessary to consider a larger population.

Figure 4 shows the distances KS for the different predictors. Context Based predictors are abbreviated by CB. The KS of predictors based on a single word are also included. It is observed that the predictor based on the impulse responses of "explain" has a large KS, although less than the Baseline and Context Based-3, Context Based-2 and Context Based-1 predictors.

4 Conclusions

In this paper we have analyzed written responses of students from various elementary school classes to open-ended questions. These are questions that teachers pose to students on an online STEM platform. In each session the students answer one or two open questions. One of the teacher's goals is for students to reflect on the multiple-choice exercises they have been doing and respond by writing justifications of their results and strategies. A critical problem is to get students to write as long as possible, since at that age they are beginning to write and still write very brief sentences. Research on science and mathematics learning [8, 4, 14] suggests that writing helps students to learn. Moreover, the longer the students write, then more opportunities students have to reflect and learn. For example, it is known that the waiting time is key and that there is an optimal time that gets students to write more [5].

In a previous paper [12] we have analyzed the impact in complete classes on the length of the answers due to the presence of key words in the teacher's questions. We analyzed the effect on the average of the entire class. However, in each class there are many differences in the length of the answers between students, and in calculating the average of the class long answers are compensated with short answers. We had found that certain words have an important impact on the average response length. In this work, unlike the previous one, we study the individual behavior of each student. By knowing the history of the lengths of a student's responses, it is possible to be more precise in predicting the length of answers to future questions. For this purpose we first calculate stimulus response per word. It is the length of that student's response when the word is in a question. This is the

effect size of each word when it appears in the questions. If we think of a word as an input or impulse, the computed average is the response to that impulse for that student. It is similar to the stimulus-response or impulse response analyzes in control systems. Based on the stimulus response for each student of each of the 200 selected non stop-words, we explore several predictors.

We use the history of 865 elementary school students from 25 classes exposed to open questions posed on an online STEM platform. In particular, we analyze four different personalized models with context information about the question, characterized by a list of 200 keywords. We collected data along a whole year, taking the data of the first semester to train our models and evaluate them on the second semester. We found that for students that have answered more questions on the first semester, the performance of one of the proposed context based models beats a baseline predictor. Moreover, given the obtained trend of error as function of the number of question on the first semester, it is probable that with a history of more questions, the performance of the context based predictors will be much better than the baseline predictor. With more questions we could also include more keywords and study the robustness of the predictors.

These results suggest that if there is sufficient history of a student's responses, then predictors based on a reduced set of keywords compete favorably with a baseline predictor. This baseline is the historical average of student response lengths in the previous semester. The baseline contains only general information, not information about the effect on the student of each word. By incorporating the impulse response of each word the prediction can be improved. The results achieved here are very preliminary because they are based on data from a few hundred students of elementary schools. As a next step we are gathering more information to increase the number of students and to increase the number of historic questions answered per student. We are also planning to explore the impact of clustering similar words [15] and reduce further the list of 200 key words.

Acknowledgements Funding from PIA-CONICYT Basal Funds for Centers of Excellence Project FB0003 is gratefully acknowledged and to the Fondef D15I10017 grant from CONICYT.

References

1. Stevens, R.: The question as a measure of efficiency in instruction: a critical study classroom practice. Teach. Coll. Contrib. Educ. **48** (1912)
2. Emig, J.: Writing as a mode of learning. Coll. Compos. Commun. **28**, 122–128 (1977)
3. Vygotssky, L.: Thought and Language. MIT Press (1986)
4. Urquhart, V.: Using Writing in Math to Deepen Student Learning. McREL (2009)
5. Rowe, M.: Wait time: slowing down may be a way of speeding up! J. Teach. Educ. 43–49 (1986)
6. Shahrill, M.: Review of effective teacher questioning in mathematics classrooms. Int. J. Human. Social Sci. **3**(17) Sept (2013)

7. McNamara, D.; Graesser, A.: Coh-Metrix: an automated tool for theoretical and applied natural language processing. In: Applied Natural Language Processing: Identification, Investigation and Resolution, pp. 188–205. IGI Global (2011)
8. Blosser, P.: How to ask the right questions. The National Science Teachers Association (2000)
9. Tofade, T., Elsner, J., Haines, S.: Best practice strategies for effective use of questions as a teaching tool. Am. J. Pharma. Educ. **77**(7) Article 155 (2013)
10. Chin, C.: Teacher questioning in science classrooms: what approaches stimulate productive thinking? J. Res. Sci. Teach. **44**(6), 815–843, Aug (2007)
11. Scott, P.: Teacher talk and meaning making in science classrooms: A Vygotskian analysis and review. Studies in Science Education, 32, pp. 45–80 (1998)
12. Araya, R., Aljovin, E.: The effect of teacher questions on elementary school students' written responses on an online STEM platform. In: Andre, T. (ed.) Advances in Human Factors in Training, Education, and Learning Sciences, vol. 596, pp. 372–382. Springer, Cham (2017)
13. Araya, R., Gormaz, R., Bahamondez, M., Aguirre, C., Calfucura, P., Jaure. P., Laborda, C.: ICT supported learning rises math achievement in low socio economic status schools. LNCS, vol. 9307, pp 383–388 (2015)
14. Winograd, K.: What fifth graders learn when they write their own math problems. Educ. Leader. **64**(7), 64–66 (1992)
15. Pennington, J.; Socher, R.; Manning, C.: GloVe: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014). http://www.aclweb.org/anthology/D14–1162