Learning Math With Sally: A Conversational Social Agent That Teaches Addition and Subtraction

Sho Cremers
s.a.cremers@student.tudelft.nl
(5052602)

Sukhleen Kaur
s.kaur@student.tudelft.nl
(5053307)

1 INTRODUCTION

Conversational agents allow people to interact with them using spoken language [9]. Such systems are usually categorized into two types; rule-based or data-driven. Rule-based conversational agents tend to be robust and efficient. They can collect information from the user and provide information back while remaining aware of the situation. They lack flexibility in the sense that a model for purpose X cannot be directly used for purpose Y. This is where data-driven models come in, where they use real-human dialogues to learn from. This allows them to create real responses as well as require less hand-crafting of features. But even though they are flexible, they might not be able to provide informative responses to the user. Their responses are also highly dependent on the size and quality of the database used, which may not always be nice[1]. A rule-based system is better for a specific task, and hence we use this for our conversational agent.

We created a rule-based conversational agent named Sally, who is a mathematics tutor. We created Sally to teach math to 2nd-grade students, according to the Ontario math curriculum [10]. Sally is designed to tutor addition and subtraction using both plain and word problems. Sally was built to give the user a learning experience as well as a sense of achievement. Sally is also able to detect emotion from the user and react appropriately. To assess our model's capabilities, we had participants interact with Sally in a controlled environment, after which they were asked to determine Sally's social presence using a questionnaire. The paper is organized as this; Section 2 states the motivation behind creating Sally, Section 3 describes how Sally is implemented as well as the experimental set-up, Section 4 provides the results of the questionnaire as well as an evaluation of Sally's ability to detect emotions, Section 5 discusses the limitations of Sally and finally, Section 6 rounds up the overall picture and provides some scope for future work.

2 MOTIVATION

Social agents have been used in many different areas, such as healthcare [7], therapy [2], and pedophilia detection [6]. But they have especially been used in teaching [1, 8].

It has been said that developing early mathematical skills leads to later success in math and reading [4]. Having weak math skills might lead students to stay behind their peers, and so it is essential to develop good math skills early [4]. Unfortunately, relying entirely on school teachers is not enough; these skills must continue to be developed at home. This may not always be possible as the student's parents might not be able to assist them in developing these skills [5]. Hence, the use of conversational/social agents might be beneficial in helping them improve their mathematical skills. These agents might be especially useful looking at the current COVID-19 pandemic, where many children have been constrained to study from home and need some extra help. Therefore, we created Sally to help children in need to help improve critical mathematical skills.

It has also been said that praising students on their behavior can lead to better performance [11]. Therefore, we make sure Sally is able to verbally and non-verbally provide acknowledgement for the answer provided as well say praise when the user does well, and ensure that everything is okay if the user does not.

3 IMPLEMENTATION

System Architecture

The main implementation of the conversational math tutor has been in Furhat SDK 1.21.0[2].

Dialogue Flow. When the user enters, Sally greets them and asks them whether they are ready to learn math. The user here can say yes, no, or enquire about Sally.

If the user says no, Sally tries to reassure the user that there is nothing to be scared of and that it only takes a few minutes. We added this behavior in Sally because children might be intimidated or not too keen at first, especially when things are new. Sally then asks them again if they are ready to learn math. If they say no, Sally says goodbye and returns to the initial state. If they say yes, Sally moves to the next state, which involves Sally introducing herself, the task, and asking the user their name.

If the user wishes to inquire about Sally, Sally introduces herself and tells them that she is a math tutor who will help them learn addition and subtraction. This behavior was

[1] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

[2] https://furhat.io/

added to inform the confused user of the agent. Sally then again asks whether they are ready to learn math. If the user says no, Sally says goodbye and returns to the initial state. If they say yes, Sally moves to the next state, which involves Sally introducing the task and asking the user their name.

If the user says yes, Sally then introduces herself, the task, and asks the user's name. Sally informs the user that they will be given a gold star for each question they answer correctly. This was done to motivate the user to do well. On providing the name, Sally greets them with a personal, "nice to meet you."

The questions asked are of three types; simple, intermediate, or advanced. Each of these questions can either be a word problem ("Alex has 5 apples, Zoe has 2 apples...)" or just a plain problem ("What is 5+2?"). To assess the user's level as either beginner, intermediate, or advanced, Sally asks a simple word addition problem followed by a plain subtraction problem. If the user answers the simple word addition problem correctly, they get an intermediate plain subtraction problem. If the user answers the simple word addition problem incorrectly, they get a simple plain subtraction problem.

If the user answers the intermediate plain subtraction question correctly, the user's level is set to advanced. If the user answers the intermediate plain subtraction question incorrectly, the user's level is set to intermediate. If the user answers the simple plain subtraction question correctly, the user's level is set to intermediate as well. If the user answers the simple plain subtraction question incorrectly, the user's level is set to beginner.

Sally then gives 8 more questions to the user. From this point on, if they answer 3 questions correctly consecutively, the user is leveled up. However, if the user answers 3 questions incorrectly consecutively, they are leveled down. These 8 questions can be changed to a different number of questions as this is declared as a global variable in our system.

After the 8 questions are over, Sally congratulates the user for "their hard work" and tells them how many gold stars they receive. This verbal behavior was encoded to give the user a sense of achievement. If the user cannot, unfortunately, answer any of the questions correctly, the user gets 1 gold star for participation. This was done so that the user does not feel upset. Sally then asks the user if they had fun. If the user says yes, then Sally expresses her gladness and says goodbye. If the user says no, then Sally hopes it would be fun the next time and says goodbye.

This dialogue flow is illustrated in Figure 1.

Showing Emotion. We encode verbal behavior into Sally to keep the user motivated. We use phrases such as "Well done!" or "Wow, you are on a roll!" if they get questions correct. We use phrases like "That's okay" or "Don't worry" if the user gets the questions incorrect.

We also encode non-verbal behavior into Sally. This starts from the initial state itself, where Sally greets the user with a smile. This is done to make the user feel comfortable and to portray Sally as approachable. If the user seems to show negative emotions initially, Sally tries to reassure them. For the first question, if the user answers correctly, Sally nods to indicate that the answer is correct, along with a verbal confirmation. If the user answers incorrectly, Sally tilts its head slightly with its eyebrows up to show a bit of empathy towards the user. Throughout the rest of the interaction with Sally, if the user gets the answer correct but shows negative emotion while answering (indicating hesitation), Sally will make sure it nods to acknowledge them. If the user gets the answer incorrect and shows negative emotions, Sally will do the head tilt to show empathy towards the user.

Emotion Detection Model and Linking to Furhat. The emotion detection model used was taken from https://github.com/atulapra/Emotion-detection which is a TenserFlow 2.3.0 model and runs on Python 3.7. It is able to predict the emotions of the user in real-time using the webcam. The model makes use of Convolutional Neural Networks (CNN) to capture emotions from an image. The model uses discrete labels to label emotions. These emotion labels are; angry, disgusted, fearful, happy, neutral, sad, and surprised. We had to convert the pre-trained model to the Tensorflow SavedModel[3] format to serve as a web service. We used Ray Serving[4] for serving the model as a web service.

Each frame of the webcam feed is captured as an NumPy[5] array into a .json file. The latest file is fed into the math tutor model, which then makes an HTTP GET query of the NumPy array parameter to the emotion detection model to retrieve the user's emotion (See Figure 2). This is done on user interaction.

Question Data. We adapted 2nd-grade mathematics word problems from Khan Academy [6]. The difficulty was set using the range of questions given here as well. The simple questions were with numbers within the range of 10, the intermediate questions were with numbers within the range of 20, and the advanced questions were with numbers within the range of 100. The whole list of questions used can be found in Appendix A.

Experimental methods

Participants. A total of 2 subjects participated in the experiment, with a mean age of 22.5 years old ranging from 22 to 23. The subjects were one male and one female who spoke English fluently and wore eyeglasses. Participants signed

[3]https://www.tensorflow.org/guide/saved_model
[4]https://docs.ray.io/en/master/serve/
[5]https://numpy.org
[6]www.khanacademy.org

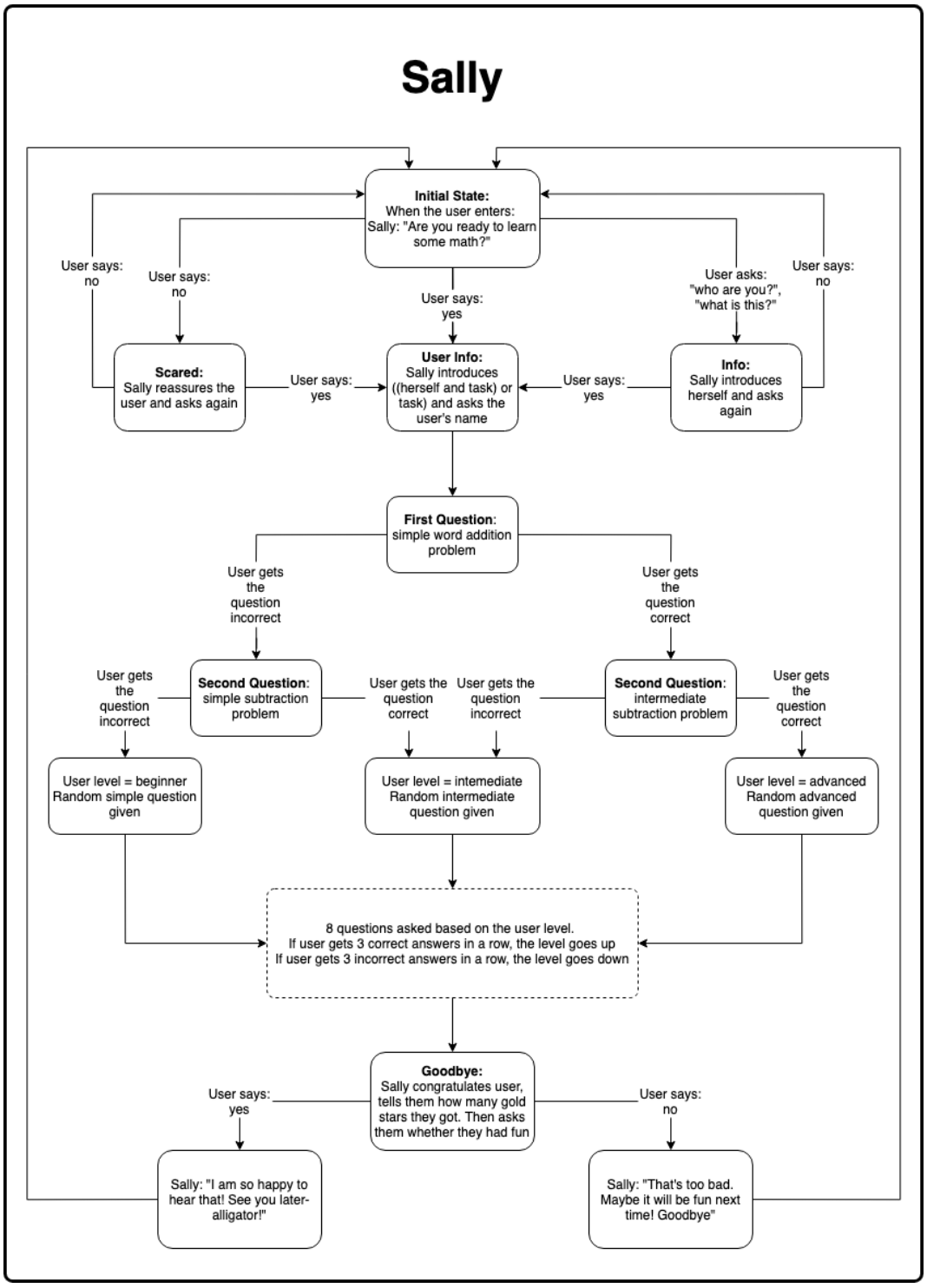

Figure 1: Sally's Dialogue Flow

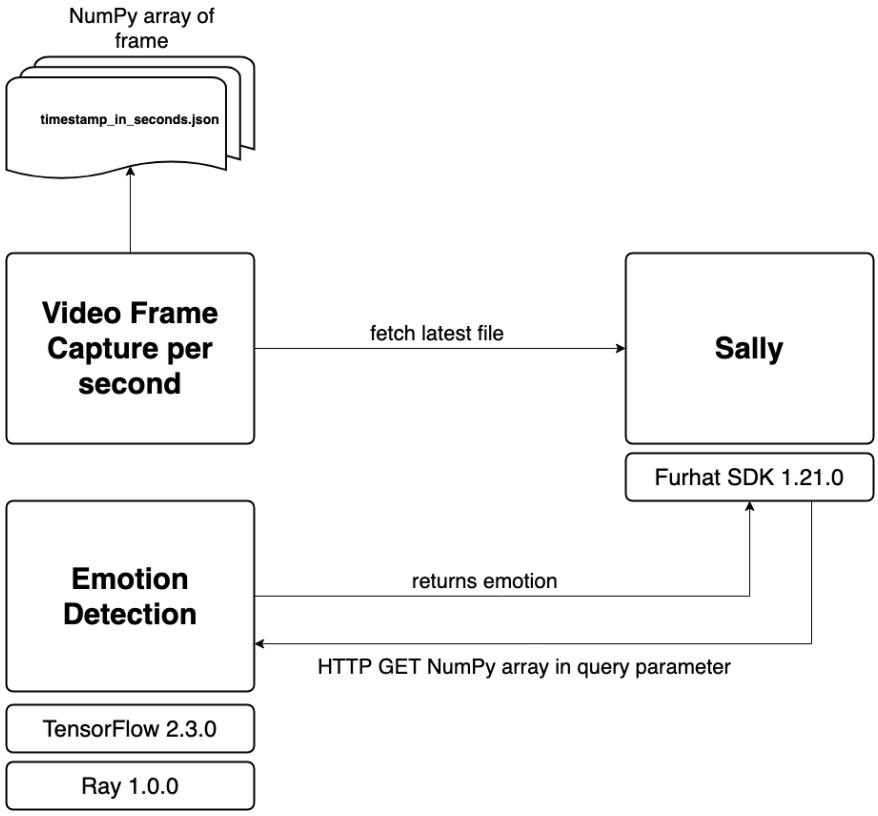

Figure 2: Emotion Detection Architecture

an electronic informed consent before the experiment. The experiment took approximately 15 minutes, and the participants received a beverage as compensation after the experiment.

Procedure. When the participant arrived, they were first told to read and sign the informed consent form. Once it was signed, participants sat in front of a MacBook Pro mid-2015 15 inches monitor. Participants also had access to paper and a pen for calculation, if needed. The monitor showed the conversational agent on the right side of the screen and the Furhat web interface with the conversation flow on the left. On the right side of the monitor, a camera was put in place to record the participant's face during the experiment, solely to annotate the emotion later.

During the experiment, participants interacted with Sally, the math tutor, for about 3 minutes. After the interaction, they filled out the social presence questionnaire by [3] to evaluate the participant's perception of the system. Seven-point Likert scale was used.

Emotion detection analysis. During the user's interaction with the conversational agent, the system recorded the predicted emotion according to the model with a timestamp every time the system entered the interaction state. From the user's face recording during the interaction, the facial emotion of the user at the timestamps was annotated by the group. The predicted emotion labels and the annotated labels were used to compute the model's emotion detection accuracy during user interaction. Emotion captured within 1 second from the previous detection was removed for the analysis.

4 RESULTS

Emotion Detection

The emotion was detected 23 times on participant 1 and 22 times on participant 2. The average accuracy of emotion detection between the participants was 0.47 (SD = 0.17).

Questionnaire

Table 1 shows the results of the questionnaire. It shows the response of each of the participants for each question. They are grouped by their dimensions, and the percentage Inter-Reliability (%IRR) for each dimension is given as well. As you can see, the two participants had the most agreeance on perceived message understanding (%IRR=67%) and none

on attentional allocation, perceived emotional independence, and perceived behavioral interdependence (%IRR=0%).

5 DISCUSSION

Sally was modeled to take the initiative over the user. Such a system's strengths are that it is very robust and limits the interaction to go out of context. Such behavior can be seen from the questionnaire results, in which the users rated highly on perceived message understanding. This shows that Sally was able to respond appropriately by restricting the user's responses. The weakness that arises from such a system is that they are not flexible. It can especially be a problem with a system that is meant to interact with children, as children tend to be talkative and have shorter attentional span. Another problem with inflexibility is that children might find the interaction with the system boring quickly if the users are restricted to only answering Sally's questions.

Another weakness of the current system is that Sally can only tell if the answer is correct or incorrect. It will be much more useful for actual students if the system can also explain why the answer was wrong. For example, if the user answered the subtraction word problem wrong by adding the numbers instead, the system should explain that the values need to be subtracted rather than added in such a problem.

One thing in the system that did not work very well was the emotion detector. The two subjects who participated in the current study had eyeglasses, which might have caused the detector to perform worse than usual. Nevertheless, a better emotion detection model needs to be used for such a system.

From the questionnaire results, we can see that the system did not perform well on perceived affect understanding, although we tried to incorporate and show Sally's emotion to the users. One of the reasons for this may be that the emotion detection model had low accuracy, and hence the system was not able to show the proper behavior according to the user's emotion. The contradiction between the user's emotion and the system's reaction might have caused the users not to understand Sally's emotion. Another reason for a low perceived affect understanding may be that the user's emotion was not continuously updated during the interaction. The system only retrieved the user's emotion when the system went into the interaction state, which may have caused a time delay between when the user's emotion was captured from when the system had to behave according to the user's emotion. The perceived emotional independence also scored low on the questionnaire, but this is most likely because the user could not understand Sally's emotion very well, and so the system could not influence the user's emotion.

Finally, the word questions showed repetition. The number of available questions might have been too small to make sure the questions were different. On top of that, certain questions did not fit well with the values. Take, for example, the following problem; "Ann and Rachel love ice cream with cherries on top. Ann puts X cherries on hers. Ann had Y fewer cherries than Rachel. How many cherries does Rachel have?". If the user is at the advanced level, the used values may be too large for a realistic situation, causing slight confusion. Hence, the system should also correctly identify which word problems should be used for which specific values.

6 CONCLUSION

In the current paper, we presented the architecture and the implementation of a conversational agent that helps children to learn addition and subtraction. The system was also connected to an emotion detector that predicts users' emotions from their facial expressions. Two participants interacted with the agent and filled the social presence questionnaire afterward. The user's emotion was detected during the interaction with the system, and it was later used to analyze the prediction of the emotion detection model.

The results of the emotion detection analysis showed that the model could not capture the user's emotion accurately. The questionnaire's responses showed that the conversational agent did well on co-presence, attentional allocation, and perceived message understanding but scored low on perceived affect understanding and perceived emotional independence. However, the system was built to tutor children with the mathematical level of a 2nd-grader [10]. Since the participants in this experiment were adults, direct conclusions cannot be made about how actual target users of the system would have thought about the interaction with the agent.

Future work for improving the system could be to use a data-driven approach for creating word questions. As mentioned in Section 5, the current system had issues with repeating the same word questions, as well as values of the problem not being reasonable. Using a data-driven system at least for forming problems, we can expect to create a larger variety of them while being coherent. Another useful addition to the system would be to be able to model the user's thinking process in order to figure out why the user made a mistake and be able to explain the solution accordingly. Finally, the system should be able to improve the emotion prediction of the user by not only using facial expression but also incorporating verbal cues by the user.

Table 1: Results of the questionnaire with the percentage Inter-Rate Reliability (IRR) for each category.

	Factor Items	Participant 1	Participant 2
Co-presence (%IRR=17%)	I noticed Sally.	Strongly Agree	Agree
	Sally noticed me.	Strongly Agree	Agree
	Sally's presence was obvious to me.	Strongly Agree	Strongly Agree
	My presence was obvious to Sally.	Strongly Agree	Agree
	Sally caught my attention.	Strongly Agree	Agree
	I caught Sally's attention.	Strongly Agree	Agree
Attentional Allocation (%IRR = 0%)	I was easily distracted from Sally when other things were going on.	Neither agree not disagree	Disagree
	Sally was easily distracted from me when other things were going on.	Strongly Disagree	Somewhat Agree
	I remained focused on Sally throughout our interaction.	Somewhat Agree	Agree
	Sally remained focused on me throughout our interaction.	Strongly Agree	Agree
	Sally did not receive my full attention.	Disagree	Strongly Disagree
	I did not receive Sally's full attention.	Disagree	Strongly Disagree
Perceived Message Understanding (%IRR = 67%)	My thoughts were clear to Sally.	Somewhat Agree	Agree
	Sally's thoughts were clear to me.	Strongly Agree	Strongly Agree
	It was easy to understand Sally.	Strongly Agree	Strongly Agree
	Sally found it easy to understand me.	Somewhat Agree	Somewhat Agree
	Understanding Sally was difficult.	Strongly Disagree	Strongly Disagree
	Sally had difficulty understanding me.	Somewhat Agree	Somewhat Agree
Perceived Affect Understanding (%IRR = 33%)	I could tell how Sally felt.	Disagree	Disagree
	Sally could tell how I felt.	Neither agree not disagree	Disagree
	Sally's emotions were not clear to me.	Agree	Somewhat Disagree
	My emotions were not clear to Sally.	Neither agree not disagree	Somewhat Agree
	I could describe Sally's feelings accurately.	Disagree	Neither agree not disagree
	Sally could describe my feelings accurately.	Neither agree not disagree	Neither agree not disagree
Perceived Emotional Independence (%IRR = 0%)	I was sometimes influenced by Sally's moods.	Strongly Disagree	Somewhat Disagree
	Sally was sometimes influenced by my moods.	Neither agree not disagree	Strongly Disagree
	Sally's feelings influenced the mood of our interaction.	Neither agree not disagree	Disagree
	My feelings influenced the mood of our interaction.	Somewhat Agree	Strongly Disagree
	Sally's attitudes influenced how I felt.	Agree	Somewhat Agree
	My attitudes influenced how Sally felt.	Somewhat Agree	Strongly Disagree
Perceived Behavioral Interdependence (%IRR = 0%)	My behavior was often in direct response to Sally's behavior.	Somewhat Agree	Agree
	The behavior of Sally was often in direct response to my behavior.	Somewhat Agree	Disagree
	I reciprocated Sally's actions	Disagree	Somewhat Agree
	Sally reciprocated my actions.	Neither agree not disagree	Somewhat Agree
	Sally's behavior was closely tied to my behavior.	Neither agree not disagree	Disagree
	My behavior was closely tied to Sally's behavior.	Neither agree not disagree	Disagree

REFERENCES

[1] Ivon Arroyo, James M. Royer, and Beverly P. Woolf. 2011. Using an Intelligent Tutor and Math Fluency Training to Improve Math Performance. *International Journal of Artificial Intelligence in Education* 21, 1–2 (2011), 135–152. https://doi.org/10.3233/JAI-2011-020

[2] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4, 2 (June 2017), e19. https://doi.org/10.2196/mental.7785

[3] Chad Harms and Frank Biocca. 2004. Internal consistency and reliability of the networked minds measure of social presence. (2004).

[4] Barbara Harris and Dana Petersen. 2017. Developing Math Skills in Early Childhood. https://files.eric.ed.gov/fulltext/ED587415.pdf. (Accessed on 11/08/2020).

[5] Sara A. Hart, Colleen M. Ganley, and David J. Purpura. 2016. Understanding the Home Math Environment and Its Role in Predicting Parent Report of Children's Math Skills. *PLOS ONE* 11, 12 (Dec. 2016), e0168227. https://doi.org/10.1371/journal.pone.0168227

[6] Carlos Laorden, Patxi Galán-García, Igor Santos, Borja Sanz, Jose María Gómez Hidalgo, and Pablo Garcia Bringas. 2013. Negobot: A Conversational Agent Based on Game Theory for the Detection of Paedophile Behaviour. In *Advances in Intelligent Systems and Computing*. Springer Berlin Heidelberg, 261–270. https://doi.org/10.1007/978-3-642-33018-6_27

[7] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Y S Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (July 2018), 1248–1258. https://doi.org/10.1093/jamia/ocy072

[8] Guido Makransky, Philip Wismer, and Richard E. Mayer. 2019. A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation. *Journal of Computer Assisted Learning* 35, 3 (2019), 349–358. https://doi.org/10.1111/jcal.12335 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcal.12335

[9] Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface*. Springer International Publishing. https://doi.org/10.1007/978-3-319-32967-3

[10] Ontario. 2020. The Ontario Curriculum, Grades 1–8: Mathematics – Curriculum Context, 2020. https://assets-us-01.kc-usercontent.com/fbd574c4-da36-0066-a0c5-849ffb2de96e/dab22a67-d9e8-4c42-a2a7-8c98cf1bbbb1/Math_Curriculum%20Context_AODA.pdf.

[11] David J. Royer, Kathleen Lynne Lane, Kristin D. Dunlap, and Robin Parks Ennis. 2018. A Systematic Review of Teacher-Delivered Behavior-Specific Praise on K–12 Student Performance. *Remedial and Special Education* 40, 2 (Jan. 2018), 112–128. https://doi.org/10.1177/0741932517751054

A QUESTION BANK

Addition Word Problems

(1) Alex has X apple(s). Zoe has Y apple(s). How many apples do they have together?
(2) David's pet snake is X metre(s) long. Monica's pet snake is Y metre(s) longer than David's. How long is Monica's pet snake?
(3) Eli has two bookshelves. On the first bookshelf he has X book(s). On the second bookshelf he has Y book(s). How may books does he have all together?
(4) Joey went to the zoo today and saw X monkey(s). That is Y fewer than he saw last Saturday. How many monkeys did Joey see last Saturday?
(5) Abby has X lollipop(s). Her brother Ben has Y more lollipop(s) than she does. How many lollipops does Ben have?
(6) Jack swam X lap(s). He swam Y fewer lap(s) than Lea. How many laps did Lea swim?
(7) Ann and Rachel love ice cream with cherries on top. Ann puts X cherry/cherries on hers. Ann had Y fewer cherry/cherries than Rachel. How many cherries does Rachel have?
(8) King Arthur had X gold coin(s) and Y silver coin(s). How many coins did he have in all?
(9) Eli had too many fish in his old fish tank. He got a new fish tank and moved X fish from the old tank to the new tank. He still had Y fish in his old tank. How many fish were in Eli's old fish tank at the start?
(10) A giant and a dragon live next door to each other. The giant's house is X metre(s) tall. His house is Y metre(s) shorter than the dragon's house. How tall is the dragon's house?
(11) Momma chipmunk had some acorns. Her babies ate X of the acorn(s). Then she ate the Y of the acorn(s) that were left. What is the total number of acorns that Momma chipmunk had?

Subtraction Word Problems

(1) The King gave Jerry X gold coin(s). He gave Elaine X fewer coin(s) than he gave Jerry. How many coins did Elaine get?
(2) Sue is X year(s) old. Her brother George is X year(s) old. How many years older is Sue than George?
(3) Ross has X dinosaur(s). Ted has X dinosaur(s). How many fewer dinosaurs does Ted have than Ross?
(4) Fiona fought a dragon with X head(s). Then she fought another dragon with X fewer head(s) than the first dragon. How many heads did the second dragon have?
(5) James has X pencil(s). Dave has X pencil(s). How many fewer pencils does James have than Dave?
(6) Yuri wrote X book(s). Hiro wrote X book(s). How many more books did Yuri write than Hiro?
(7) Yesterday Ali caught X bug(s). Today he caught X bug(s). How many fewer bugs did Ali catch yesterday the today?
(8) Sam rode a bull for X second(s). Maddie rode a bull for X fewer second(s) than Sam did. How long did Maddie ride her bull?
(9) Phoebe jumped X metre(s). Aria jumped X metre(s). How much farther did Phoebe jump than Aria?
(10) Ross's class has X crayon(s). At the end of the year only X crayon(s) remain. How many crayons have been used?
(11) In a basketball game Ross scored X fewer point(s) than Chandler. Chandler scored X point(s). How many points did Ross score?
(12) Vic used X marker(s) to draw his pictures. His friend Susan drew her picture with X fewer marker(s) than Vic. How many markers did Susan use?
(13) Sparky the dragon was born with X spike(s). He grew several more spikes as he got older. Now sparky has X spike(s). How many new spikes did Sparky grow?
(14) There are X camper at Fun Camp. Only X camper brought their sleeping bags. How many campers did not bring sleeping bags?
(15) Mary has a collection of shark teeth. When her uncle returns from vacation he brings her X more shark tooth/teeth. Now Mary has X shark tooth/teeth. How many did Mary have before her uncle's gift?