Quizbot: Exploring Formative Feedback with Conversational Interfaces

Bharathi Vijayakumar, Sviatlana Hoehn and Christoph Schommer

University of Luxembourg
sviatlana.hoehn@uni.lu
Esch-sur-Alzette, Luxembourg

Abstract. Conversational interfaces also called chatbots recently disrupted the Internet and opened up endless opportunities for assessment and learning. Formative feedback providing learners with a practical instruction for improvement is one of the challenging tasks in, for instance self-assessment settings and self-directed learning. This becomes even more challenging if user's personal information such as learning history and previous achievements cannot be exploited for data protection reasons or are simply not available. This study seeks to explore the opportunities of providing formative feedback in chatbot-based assessment. Two main challenges were faced: the limitations of the messenger as an interface that restricts visual representation of the quiz questions, and zero information about the user to generate adaptive feedback. Two types of feedback were investigated regarding their formative effect: *immediate feedback*, which was given after answering a question, and *cumulative feedback* detailing strengths and weaknesses of the user in each of the topics covered along with the directives for improvement. A chatbot called SQL Quizbot was deployed on Facebook Messenger for the purposes of this study[1]. A survey conducted to disclose users' perception of the feedback reveals that more than 80% of the users find immediate feedback helpful. Overall this study shows that chatbots have a great potential as an aiding tool for e-learning systems to include an interactive component into feedback in order to increase user motivation and retention.

Keywords: Formative Feedback, Educational Chatbot, Quizbot

1 Introduction

Although online education revolutionised the conventional style of learning, many of the e-learning implementations only reproduce textbooks and do not offer interactive feedback, in contrast to teacher-student communication. In addition, online education suffers from a large number of drop-outs, which is a consequence of insufficient interaction, according to the study by Muirhead and Juwah (2004) [1]. The authors recommend to develop strategies that will enhance guidance for online students, such as creating a timeline for feedback and having

[1] Try out the prototype at *https://www.messenger.com/t/2076690849324267*

a specific feedback. Previous academic publications show that students retention is higher when Facebook is used for course interaction [2]. MOOC (Massive Open Online Course) students were more engaged in Facebook groups, and they preferred interacting more in social media, which they use anyway, than through the course tools, which are dedicated to course-related communication.

However, in 2015 the preferred way of communication was no longer the social networks. At this time, big social media companies realised that more users are active in instant messengers than in social networks[2]. The messenger providers opened their APIs (Application Programming Interfaces) to facilitate the development of a special sort of interfaces to services - chat robots.

A conversational interface, or simply, a *chatbot* (a short form of *chat robot*) is a software program that usually interacts with the users in instant messengers and usually understands input in human languages, but also may use buttons and predefined replies to facilitate language understanding. The first chatbot was created many decades ago by Joseph Weizenbaum (1966) [3]. After being very unpopular for many decades, chatbots disrupted the Internet in 2015 and keep growing. Intuitively, a chatbot in a commonly used messenger has good chances to be used actively by students who are used to using messengers in their daily communication. The chatbot could help the students in self-directed learning and, first of all, provide meaningful feedback in an interactive way.

The traditional view on chatbots is still very dominant, saying that a chatbot needs to possess extraordinary conversational abilities in order to be useful. However, if used for computer-based assessment and feedback, the questions of usability and user experience combined with the opportunities to provide meaningful feedback in the dialogical setting need to be clarified. For instance, how to present various types of tasks to the learner and how to deal with the tasks that cannot be presented in a messenger for some reasons?

Prior to providing feedback, some form of evaluation of a learner's performance is needed, and assessment is one possible form of a systematic evaluation. Shute (2008) defines *formative feedback* as information communicated to the learner that is intended to modify their thinking or behaviour to improve learning [4]. Formative assessment implies obtaining the best possible evidence about what is learned, and then using this information to decide what to do next.

Assessment methods can be classified as short-term and long-term methods. Short-term methods only assess the learner's state of proficiency in a particular subject at a specific time point without the need to capture the learning process or the intermediate milestones in learning. In contrast, long-term methods would continuously monitor the learner's progress at various levels and can be associated with multiple tasks. Consequently, there is no need to store any kind of information about the learner in the short-term setting.

Short-term settings for chatbot-based assessment are especially interesting in the context of the new General Data Protection Regulation (GDPR) that provides a legal framework for the collection and processing of personal information

[2] http://uk.businessinsider.com/the-messaging-app-report-2015-11?r=US&IR=T, last checked 22.11.2018

of individuals within the European Union (EU). In addition to the potential use of personal data by chatbots, assessment data are a special case of sensitive data, that can be misused for discrimination against the learner. These issues need to be considered in the design of chatbot-based assessment solutions.

Objective With the motivation in the preceding paragraphs, the objective of this study is to *explore how can formative feedback be implemented for an educational chatbot in a short-term assessment setting.*

Method A chatbot can be made available in various chat platforms such as Facebook Messenger, Slack, Viber, Telegram, Kik and Snapchat. However, legacy platforms, such as email and SMS, or voice-based platforms, such as Alexa and Google Assistant, can be used for chatbot development, too. For the purposes of this study we chose the Facebook Messenger for the implementation because of the availability of multiple tools that facilitate deployment of a Messenger chatbot. We chose the topic of SQL database query language because of its popularity and a high number of learners. The chatbot was named SQL Quizbot because it can be easily found using the keyword *quiz*. Test items of the SQL Quizbot were designed as two-step questions: in the first step, a question related to SQL is presented to the user; in the second step, the user needs to score their own confidence on the answer. Section 3 explains the details of the design, the implementation and the methodological and technological challenges. To evaluate the feedback effectiveness we set up a user survey focusing on user's perception of the usefulness of the feedback. Section 4 explains the results of the evaluation. Finally, Section 5 makes conclusions based on the findings of this work and formulates suggestions for future development of the formative chatbot-based assessment.

Contribution This research has two main contributions: first, it shows that simple feedback, such as presentation of a correct answer, has a formative effect if presented at the right moment, which was achieved through the additional requirement for the learners to self-assess their confidence level; and second, it shows that for education and assessment practitioners, there is a fast and affordable way to create working chatbots for assessment using only free-of-charge cloud-based libraries.

2 Setting up the Frame

We see the research objective of this work as linked to mainly three domains:

1. Assessment research related to the interplay between learning, feedback and assessment methods;
2. Psychology helping to understand mutual dependencies between task properties, learner skills and confidence, and test validity;
3. Use of instant messengers and chatbots in educational settings.

The following sections discuss previous academic publications in these domains and put the research work of this paper in relation to the state of the art.

2.1 Formative Feedback, Assessment and Learning

Feedback in general has one of the most persuasive influences on learning and performance, but this impact can be either positive or negative [5]. Education research disclosed dependencies between variables such as learner level of competencies, learner motivation, task properties and feedback effectiveness [6]. In addition, adaptive feedback was found more helpful for the learner than generic feedback. Feedback is called *formative* if and only if the information provided to the learner by means of feedback is used by the learner as an instruction for action with the purpose to improve performance [7]. Although formative feedback research emphasises the importance of including information about the learner in the feedback generation, it is not possible to use or to access the knowledge about the learner in some settings. One attempt to overcome these limitations was described in [8] where user's insufficient confidence about the answer has been captured through mouse movements during the tests runtime.

Feedback in general can be provided at several conceptually different points (after or during a test/exam or during the work process) and by different actors (teacher, peers or learners themselves). If provided by an automated system, the system still takes a role of one of the listed actors: either it acts as an expert, or as co-learner or the learner uses the system for self-feedback. Prior research shows that automated feedback may be very appreciated by the students if it is well-structured and meaningful, and may be even preferred as compared to teacher's feedback [9]. The target of the feedback may be correcting inappropriate task strategies, procedural errors, or misconceptions. Any of these feedback types can be formative or not.

Because of very intensive critics of assessment and its inability to reflect the actual state of learners' proficiency level in a particular subject or skill due to such issues as test validity, the theory of *formative assessment* was developed [10]. Formative feedback in this context will be provided to the learner before, during or after a sort of assessment. The feedback needs then to be tailored to the user, task and environment in a way that facilitates the learner's improvement in a subject or a skill.

The information contained in a feedback message may contain more or less details on the error and the correct answer. The types of feedback from this perspective include presentation of the correct answer, verification (correct/incorrect), error flagging and elaborated feedback. The later may contain hints or prompts, and the learner may have a possibility to try again or even to repeat trying until a correct answer is given [4]. Goal-directed feedback provides learners with information about their progress towards the desired goal (or set of goals) rather than providing feedback on discrete responses.

Wiggins (2012) defines 7 key characteristics of effective feedback: goal-related, tangible and transparent, actionable, user-friendly (specific and personalized), timely, ongoing and consistent [11]. Furthermore, [12] shows that a dialogic perspective of feedback potentially promotes learning. The components of such a feedback are (1) providing feedback before the task is completed, (2) incorporating peer feedback into the process and (3) allowing resubmission once the

feedback has been received. Thus, feedback needs to be subject of rebuttal and not a final verdict in order to be effective. In line with this need, Narciss [13] argues that feedback messages can only have a formative effect if students have occasions to use the feedback information for regulating their learning process.

Types of feedback regarding timing that can be found in literature are immediate, delayed and postponed. Butler and Roediger (2008) show that the immediate feedback makes a positive impact when presented immediately after an incorrect response [14]. Blter et. al (2013) investigate the effect of a correct/incorrect feedback as part of generic quizzes [15]. The study shows that simple quizzes combined with simple feedback are effective in the first weeks of a course. This gives an idea about the ideal placement of this type of activity.

In line with these findings, we chose a multiple-choice quiz as an activity type and decided to provide feedback immediately after each test item. Evaluative feedback will be provided after correct answers and correct answer will be provided immediately after incorrect answers. However, the quiz will not be part of a course but an independent activity. The details are discussed in Section 4.

Novel forms of assessment always set new challenges in evaluation; see for instance [16] on psychometric multimedia-based assessment. Because of complex mutual dependencies and multiple feature combinations, it is difficult to evaluate the formative effect of automated feedback in a chatbot. A simple evaluation framework applicable for the technological proof of concept described in this article is needed. As Hattie and Timperley (2007) suggest, a feedback that enhances learning needs to answer three questions [5]: (1) Where am I going? (Feed Up); (2) How I am going? (Feed Back); (3) Where to go next? (Feed Forward). We use this three-dimensional framework as the baseline for the evaluation of the feedback in the set of user tests with the SQL Quizbot.

2.2 Quizzes, Feedback and Confidence

The validity of test results is an important prerequisite for meaningful feedback. In addition to the issue of test validity (did we test what we wanted to test?), assessment based on multiple-choice questions makes it easy for the students to guess the right answer. For instance, the study by Novacek (2013) shows how multiple-choice questions can be passed if the students are simply guessing the answers [17]. Consequently, the success of such a form of assessment depends on the honesty of the learner, and therefore is suitable as a self-feedback tool without any consequences for official exam grades or course scores. In such cases, an opportunity to reflect on their own confidence may create an environment for the learners in which they develop a deeper level of self-regulation.

The learner's confidence helps to assess the meta-cognitive level of the learner's knowledge. It provides an insight into what learners think they know and what they think they dont know, as opposed to assessment based on the performance. For instance, Hench (2014) highlights the use of confidence and performance details of a learner to provide feedback by first demonstrating the data on a simple linear model [18]. The model is used to provide feedback that allows students to infer either the difficulty or the degree of under- or overconfidence associated

with a specific question. The findings show that confidence indicators encourage students to reflect on their knowledge and rethink their choices.

However, there are differences in confidence judgements in students of different genders and proficiency levels. The study presented in [19] investigates gender differences in item-specific confidence judgements. The students had to judge their confidence of the answer correctness after each item. The study shows that especially undergraduate males were inappropriately overconfident when their answers were incorrect. In contrast, female subjects tend to judge their confidence more precisely. A more recent study [20, p.562], however, argues in the discussion of the state-of-the-art literature, that previous results on gender-specific differences in confidence judgement accuracy are mixed and not able to resolve the issue. The study itself [20] does not find any clear support for gender differences, either. The authors conclude that "gender differences with respect to realism in confidence judgments are unstable and that they are dependent on the knowledge domain and/or on the cognitive processes activated by the task given in a knowledge domain"[20, p.562].

Confidence can be measured at different time points related to the task: as a self-report or "online" as a post-question immediately after the task [21]. As suggested in [19], the confidence judgements are more accurate if the students need to estimate their abilities immediately after answering the respective item. The "online" confidence judgement is more closely related to measuring the ability in contrast to personality, for which self-reports suit better [21].

With regard to the scales for confidence measurements, multiple approaches can be found in the academic literature: binary (high vs. low), discrete multi-class (a small number of discrete classes), such as scale from 20 (guessing) to 100% (absolutely sure) [20, **?**], and continuous (e.g. the student needs to mark a point on a line without any scale on it which represents the infinite number of points between 0 and 1), which then is also mapped to a number of discrete classes, but the classes are defined after the measurement.

The study by [22] suggest that we have to keep practicing in order to remember what we have learned. This finding is confirmed by [23] arguing that every topic or subjects requires a set of core skills and knowledge that will be used again and again, and this forms the basis for any kind of expert knowledge a person can acquire. Because multiple-choice questions offer the learners an easy way to practice while testing their own knowledge, we see this kind of quizzes as a possible implementation of a chatbot-based self-assessment to support learning.

The study presented in [24] shows that quizzing helps learners grasp more information than re-reading. This is also called the "testing effect" or "retrieval practice". The authors considered the concepts of dynamic testing and formative assessment to improve learning.

Based on the discussion above, this research builds on the concept of confidence to capture at least some additional information about the user's knowledge during the test run-time. The "online" measurement was chosen because of the need to measure specific technical knowledge and will be performed in form of a post-item question. The measurement scale was chosen to be *low* for *not sure*,

high for *absolutely sure* and *medium* for *not sure whether I know this*. Section 3.1 explains the details.

2.3 Chatbots in Education

Contemporary intelligent tutoring systems and e-learning platforms providing automated feedback to the user are grounded in education research [13]. Automated feedback in different learning contexts has been investigated for traditional e-learning systems and artificial agents acting on websites, see for instance [25, 26]. In addition, the use of chatbots in educational settings is subject of multiple academic publications, see for example [27, 28]. Multiple different classes of chatbots in education have been introduced during these decades, such as agent, virtual character, intelligent agent, pedagogical agent, avatar and guidebot; see [29] for a review. A more broader term of *conversational interfaces* is chosen today to describe all more or less complex types of software that communicates with human users using some kind of human language [30].

Because chatbots were seen in the beginning as software that mainly has to engage in conversations with a user, a large part of all educational chatbots focuses on second-language acquisition; see for instance [31–34]. In this context, the feedback mainly concerns correction of linguistic errors and is usually grounded in Second Language Acquisition theory and is called *corrective feedback* [35, 36]. Providing feedback based on linguistic features, learner information and activity information is seen as one of the major challenges for Intelligent Computer-Assisted Language Learning [37].

Another popular domain of educational chatbots is related to programming and technical skill acquisition. For instance, [38] introduce an intelligent, adaptive learning environment for Java programming called FIT Java Tutor. It provides feedback to learners in the form of a solution for the programs whenever the learners request it. A system which assists the students in converting natural language phrases to First Order Logics is discussed in [39]. After submission of the answer in the end, the system characterises the answer in terms of completeness and accuracy to determine the level of incorrectness, based on a template. In this way, the system provides elaborated feedback to the users. Both [38] and [39] use some history information about the student to provide adaptive feedback.

Despite the huge landscape of tools that support chatbot development for messengers (rapid prototyping, natural language understanding libraries, connectors and messenger native APIs), the majority of research publications after 2015 still report about chatbots based on custom solutions (usually own university prototypes). In this work, we explore how a set of state-of-the-art tools for chatbot development can be used to implement an educational chatbot that provides formative feedback with limited information about the user. The next section explains the details.

3 Designing Formative Feedback for the SQL Quizbot

This section explains the design decisions regarding an implementation of a formative feedback component for an educational chatbot acting in Facebook Messenger without using personal user information or interaction history. Section 3.1 provides details of the quiz structure and the features used for the feedback generation. Further, Section 3.2 describes the overall software architecture of the SQL Quizbot and the tools used.

3.1 Quiz and Feedback Design

For the purposes of this research we chose the quiz to be composed of 25 multiple-choice questions that are randomly combined to a sequence of 25 test items presented to the user one by one. It is mandatory to answer every question, the user can go to the next question only after they answered the preceding one. The quiz questions will be distributed across five different technical topics: SQL Basics, Functions, Joins, Index and Stored Procedures. Every question will be bind to only one topic. As argued in Section 2, the level of learner's confidence can be used as an additional source of information to provide relevant feedback. For our solution, we chose to let the users decide and self-assess their confidence on each test question. Following this design decision, every test item consists of two parts:

1. Multiple-choice question related to one of the five topics;
2. Self-assessment scoring of the confidence level related to the answer.

The user will be informed about the requirement to judge their own confidence regarding every answer. We expect this feature to have a double effect: first, the user would be more conscious in answering questions (internal feedback), and second, in case of a mismatch between the confidence level and the correctness level (low confidence and correct answer or high confidence and incorrect answer) the user would be made more sensitive to feedback and more perceptive for corrections.

The user is not allowed to go back to the preceding questions because the impact of revisiting the question and immediate feedback is not studied in this work. Also proceeding to the next question is currently controlled by the bot. Whenever the user is away from the system, or active in another window, the quiz is automatically paused and can be continued later as per the preference of the user. This pausing is allowed because time spent on a question is not considered for the assessment and feedback. There is also an option to restart the quiz, but it is only to facilitate the user to restart when there are technical glitches.

As we argued in Section 2.2, the concept of *confidence* can be exploited for filling the gap in information about the user in order to make the feedback adaptive. We chose a grading scheme based on a confidence at three levels: low, medium and high. Depending on the correctness of the answer, a joint score

reflecting the confidence and the correctness will be assigned to the user for the respective test item Table 3.1 summarises the differences in scoring for all six cases.

Confidence	Low	Medium	High
Correct answer	1	2	3
Incorrect answer	0	-2	-6

Table 1. Scores assigned to learner per test item in case of correct and incorrect answers based on the reported confidence level

Because of the expected double-effect of the confidence judgement, the feedback based on the confidence/correctness score will consist of two parts:

1. Immediate feedback will be provided immediately after each test item.
2. Cumulative feedback over all test items will be provided at the end of the quiz.

Immediate feedback will be generated from a template and contain a different message depending on the confidence/correctness level. The feedback message will contain two parts: a didactical part and a social part. The didactical part will address the correctness of the answer and will be simply a "Correct!" statement if the user's answer is correct. The system will present the solution, if the user's answer is incorrect. The social part of the message will address the confidence level of the user combined with the correctness of the answer. Especially for the incorrect answers, the chatbot will encourage the user to learn from mistakes and show empathy using emoticons. Figure 1 provides an example of a test item and an immediate feedback. In this example, the user chose a wrong answer option, and the specified confidence level is low. The feedback message that SQL Quizbot provides to the user contains a correct answer and an encouragement to continue learning.

Cumulative feedback at the end of the quiz is generated and presented to the user. The cumulative feedback contains two types of information:

1. Conventional total correctness score with one point per correct answer and zero points for each incorrect answer. The total score indicates an overall performance.
2. Total score based on the confidence level. Table 3.1) summarises the recommendations shown to the user to address user's overconfidence and underconfidence on specific topics of the quiz.

The cumulative feedback design is based on the confidence level that the user enters after answering every question. A total score based on confidence is calculated, and it can range from -150 (if all 25 questions are answered wrong with high confidence) and a maximum score of 75 (if all answers are correct with

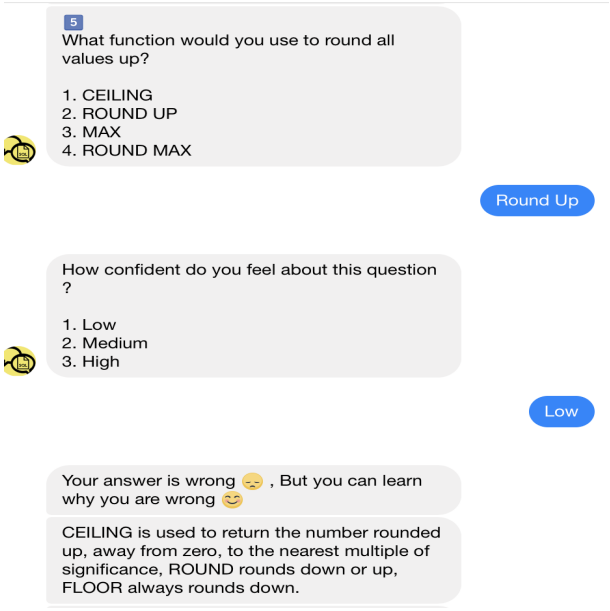

Fig. 1. An example of a test item and immediate feedback

high confidence). A total confidence score greater than the normal score indicates that the user performed really well overall in the SQL quiz with a genuine exhibition of confidence, leading to a positive feedback or appreciation. A total confidence score lower than the medium score indicates that the user misjudged the level of confidence in most of the questions which were answered wrongly, leading to a feedback for improvement. This suggests the scope for improvement in the knowledge regarding the subject and also, clearly distinguishes what the user knows and what he/she was guessing. The summary about how a feedback message is generated using confidence level is specified in Table 2.

In both the above discussed scenarios (i.e positive feedback and feedback for improvement), there may be some questions in specific subtopic/competency performed well compared to other subtopic. Because every question is mapped to a topic, the user is informed about the performance in every topic.

After finishing the quiz, it was mandatory for the user to participate in a survey targeting the effect of the feedback that was presented to the user. Table 3 presents the survey questions and the results.

The credibility of the system which gives the feedback is always questioned. If a chatbot is involved in the assessment process, the users need to trust it to accept the feedback. So by asking the user to enter the confidence level, we involve them in the feedback design. Whenever negative feedback is presented to a user, the user might understand that their decision is part of the feedback and may not get offended (compare to [40]). In order to reduce the negativity during the quiz, if the user answers three questions continuously incorrect, the

Confidence level	Correct answer	Incorrect answer
Low	Good score but the user underestimates her/his abilities.	The user didn't score well and but identified correctly that there is scope of improvement.
Medium	Good score but there is lack of meta-knowledge.	The user didn't score well and also, didn't identify correctly that there is a need for improvement.
High	Good score and judged appropriately that he knows the answer.	The user didn't score well is inappropriately overconfident.

Table 2. The interpretation of the confidence scores provide a basis for the different feedback messages

SQL Quizbot sends the user a message encouraging and cheering up the user. The user is provided with sufficient time to read the feedback before proceeding to the next question.

3.2 Implementation

We used Chatfuel[3] for designing the conversational interactions for the Quizbot. The entire interaction with the user takes place in Facebook Messenger, since Chatfuel allows easy connection to Facebook. However, Chatfuel does not store any data and cannot do complex manipulations using the data. For all back-end processing such as storing the quiz summary after answering a question and generating the cumulative feedback based on confidence is done using Firebase[4]. Simple tasks like navigating to the next question and storing the survey results in Google sheets are done using Integromat[5].

Every question in the quiz is composed as the flow of static text cards for the question along with the choices in a vertical arrangement. The Chatfuel generates dynamic blocks or cards based on the JSON response using the API links connected to Chatfuel platform.

We chose Firebase for managing user answers and scores because of its flexibility, scalability, possibility to store data in JSON format, its simplicity of integration with Chatfuel and its popularity in small-scale development. In addition, it is not required in Firebase to create database schema prior to development because, therefore it allows changes to the schema during the phase of development. Cloud functions run back end code in response to events triggered by Firebase features and HTTPS requests. The code is stored in Google cloud and

[3] https://dashboard.Chatfuel.com
[4] https://Firebase.google.com
[5] https://www.Integromat.com/

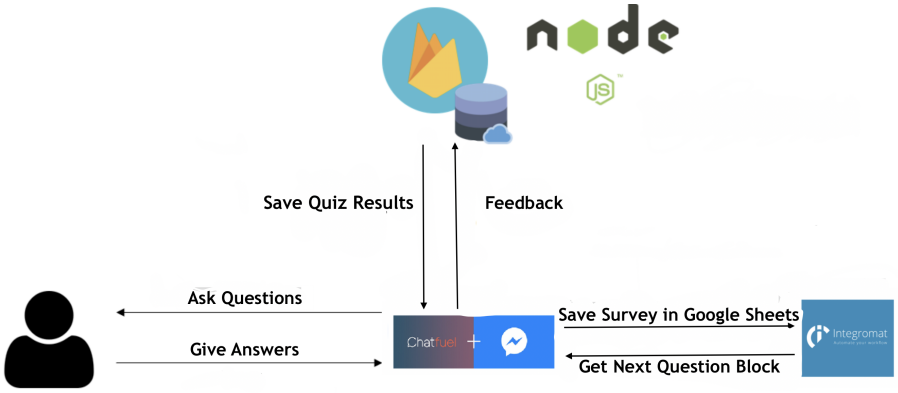

Fig. 2. Workflow

runs in a managed environment. The integration of Admin SDK together with Cloud Functions provides a webhook for the Chatfuel.

The Integromat is used in the SQL Quizbot for two scenarios: first, for navigating from one question to next question block in Chatfuel; and second, for saving the survey results from every user to Google Sheets.

4 Evaluation

This section describes the evaluation of the SQL Quizbot with regard to the participants' perception of the usefulness of the feedback. As mentioned in Section 3.1, every user was asked to participate a survey at the end of the quiz. Section 4.1 explains the details of the experimental setup and the population. In Section 4.2 we discuss the findings.

4.1 Experiment Design and Main Findings

The participants of the study were acquired through researchers' personal channels. The population included representatives from diverse backgrounds (students, IT professionals, academic staff and business analysts) with almost equal male-female ratio. The participants received a personal invitation to join the SQL Quizbot on Facebook Messenger upon their prior agreement to participate the study. The participants were allowed to use the Messenger on any device and at any time they prefer. Although the chatbot offers the option to rerun the quiz, only the first fully accomplished attempt counts for this study. One important observation is that none of the users required any demonstration or explanation on how to use the chatbot, the technology adoption happened on the fly because of the familiarity with messengers and quizzes.

Overall 47 persons agreed to participate the study. Out of these 47, seven persons quit the quiz either in the beginning or in the middle of the quiz. Out

of these seven persons, two persons could not continue because of technical problems, and the remaining five did not show interest in finishing the quiz. In total, more than 90% of the users completed the quiz and were then asked the questions presented in Table 3.

The results show a high grade of acceptance and a positive attitude towards a chatbot-based assessment from all the users. In all the survey questions, either very helpful or helpful was selected as survey answers by more 80% of the users. Overall, only less than 4 users had felt that the bot was either unhelpful or none leading to a negative feedback. Table 3 shows the distribution of the answers to each survey question.

Survey Question	**very helpful**	**helpful**	**unhelpful**	**none**
How would you rate the quiz in chat format?	47%	41%	9%	3%
How would you rate the feedback based on confidence?	39%	49%	9%	3%
How would you rate the immediate explanation on wrong answers ?	47%	41%	9%	3%
How would you rate the Confidence Indicators in making you rethink or reflect on your answers?	37%	47%	9%	6%
How would you rate the overall experience with the Quizbot?	64%	30%	0%	6%

Table 3. Survey Questions

In addition to these scored questions, the users had a possibility to leave an open comment and share their experience with the developer. About 30% of the users (13 out of 40) chose to do this. The content of the comments confirms a high grade of user acceptance of this kind of technology: seven of the 13 qualitative answers were just praise. Six remaining comments provide meaningful directions for further development and include suggestions such as providing the correct answer in the feedback even if the user's answer is correct but the confidence level is low; providing a final overview of all question numbers and the correctness of the response; making the length of the quiz variable. Overall, the SQL Quizbot was perceived as cool and fun, which also shows a positive user's attitude towards artificial learning companions.

4.2 Results and Discussion

As was announced in Section 2.1, we use the three feedback criteria to evaluate the SQL Quizbot (Feed Up, Feed Back, Feed Forward). In addition we address the issues of technology adoption and novelty, and technical advantages and limitations.

Feed Up Although goal setting was not the purpose of the SQL Quizbot, it needs to be taken into consideration that the same feedback has different effect for users with different goals in using the Quizbot. Factors such as professional relevance and assessment purpose can be captured in a pre-quiz in order to make feedback more helpful. However, it was a design decision in the beginning of the study not to use any personal information.

Feed Back The need to reflect on their own confidence after each answer was expected to make the users reflect on their own knowledge of SQL. The findings confirm this hypothesis: the users were made sensitive toward the presentation of correct answers when they were overconfident or under-confident. In this way, the feedback generated by the chatbot was only effective in connection to the self-feedback based on this meta-knowledge. However, as discussed in Section 2.2 complex relationships between confidence and different personality features do exist and need to be investigated in a separate study.

Feed Forward This research confirms an earlier finding that making a mistake can feel rewarding when the brain is given the opportunity to learn from its mistakes and assess its options [41]. Although users who made multiple errors reported that they felt frustrated and would prefer to get a bit of encouragement from the chatbot, they also reported that they found the immediate feedback helpful. This subjective perception of the feedback's helpfulness may not be confirmed in a post-test. However, the subjective perception of the helpfulness is a great motivator to learn and to keep using the tool. In addition, the cumulative feedback in the end of the quiz provides explicit instructions on topics to be revised.

Although the novelty effect might be one of the reasons for the positive users' perception of the SQL Quizbot, a sustainable adoption may be very likely for this kind of assessment for the following reasons: first, the learners are familiar with the concept of the quiz, there is nothing conceptually new for them in a quiz; second, a quiz has by itself a game principle, and earning a higher score may motivate the users to use the chatbot again and again; third, users' expectations towards the Quizbot's conversational capabilities are managed in a way that the user would normally not expect extraordinary language understanding from the bot, and therefore not become disappointed; and finally, the interactive feedback component of the bot in a messenger showed a motivating effect and positively influenced users' intention to learn. Nevertheless it needs to be investigated in a separate long-term study whether this forecast is true.

The biggest challenge in implementation of the quiz questions was determined by the limitations of the messenger as interface. Although instant messengers provide their own widgets (e.g. carousel) that can be used for creating of new types of questions, they make it sometimes difficult to transfer commonly known question types into messenger window. More specifically, long texts in task descriptions are not appropriate in a messenger window, and automatically generated short replies are not well-displayed if their number is higher than three. Therefore, the question of user experience in chatbot-based assessment may be worth a more detailed study.

The existing implementation of the Quizbot has the advantage that even persons without coding skills can create such quizbots. However, the implementation not flexible because it has a fixed number of test items. All the questions are added as static text cards along with answer options in the form of quick replies in Chatfuel. So there will be 25 blocks for accommodating the quiz questions alone in the dashboard of this project. This makes the project in the Chatfuel dashboard difficult to interpret for a person without an explanation. Firebase cloud function is used only for storing the quiz results and getting the immediate and cumulative feedback. If we plan to do any changes to the number of questions, introduce a new subtopic, or display the choice as a list of images, it requires changes in both Chatfuel and Firebase, which is not well maintainable. In addition, any change to the Firebase function needs to be built and deployed again to make the changes come into effect. A more flexible and scalable approach would be an advantage.

Although elaborated feedback may be helpful in many settings, instant messenger interface requires concise information presentation and short text messages. This adds a new challenge to the feedback and formative assessment research requiring that formative feedback, which is elaborated, adaptive and personalised, needs to be in addition very concise and interactive.

Because knowing what one knows and what one does not know has certain implications for the learning behaviour, confidence judgment in chatbot-based quizzes is one of the options how we can help students to achieve an appropriate confidence calibration and optimise their studies.

5 Conclusions and Future Work

This paper presented a research study focusing on the design and implementation of formative feedback in a chatbot acting in an instant messenger. Although prior research states that elaborated feedback is usually more effective, our study shows that simple feedback such as presentation of a correct solution has a formative effect if presented at the right time, namely, when the students are made sensitive to their mismatch in confidence and correctness results. This also shows the effectiveness of a reflection on meta-knowledge.

The use of artificial teaching assistants such as SQL Quizbot was difficult in the past because of the implementation effort that was needed for the custom solution. It was shown in this study that on the practical level it is quite fast and easy to implement a working prototype for a proof of concept using only free-of-charge cloud services without the need to write complex program code. This result may encourage other researchers in education and assessment research to continue these investigations.

This study generated more open questions than answers. We make a short outline of possible future research directions based on the presented findings.

Because existing question databases such as Moodle already contain a huge number of potential test items for chatbot-based assessment, it needs to be tested in practice, which of them can be transferred into messenger with which effort.

In this way, guidelines for user experience for chatbot-based assessment can be formulated. As a continuation, automated test-item generation from text books and open learning resources similar to [42] can be adapted to be presented in a messenger.

In order to make the quiz more interactive and competitive, the Quizbot can be implemented in team-work environment such as Slack. Team members may play with the bot against each other or against the bot and compete in their expert skills. A deployment of the same Quizbot on multiple messengers is easily possible with the current tool support.

Although the SQL Quizbot explicitly excluded the use of the learner's personal information, future personalisation can help in providing a more effective feedback. If implemented as a learning long-term companion that assists the learner for many weeks of a course, the chatbot can store intermediate quiz results and track learning progress. In addition, long-term data capturing and personalised user modelling in the educational context may help to get deeper insights in the process of learning, which in turn may initiate new and modify traditional psychology and learning theories.

References

1. Muirhead, B., Juwah, C.: Interactivity in computer-mediated college and university education: A recent review of the literature. Journal of Educational Technology & Society **7**(1) (2004) 12–20
2. Zheng, S., Rosson, M.B., Shih, P.C., Carroll, J.M.: Understanding student motivation, behaviors and perceptions in moocs. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. CSCW '15, New York, NY, USA, ACM (2015) 1882–1895
3. Weizenbaum, J.: ELIZA - a computer program for the study of natural language communication between man and machine. Commun. ACM **9** (1966) 36 – 45
4. Shute, V.J.: Focus on formative feedback. Review of Educational Research **78**(1) (2008) 153–189
5. Hattie, J., Timperley, H.: The power of feedback. Review of Educational Research **77**(1) (2007) 81–112
6. Lipowsky, F.: Unterricht. In: Päd. Psychologie. Springer (2015) 69–105
7. Wiliam, D.: Embedded formative assessment. Solution Tree Press (2011)
8. Höhn, S., Ras, E.: Designing formative and adaptive feedback using incremental user models. In: ICWL2016, Springer (2016) 172–177
9. Denton, P., Madden, J., Roberts, M., Rowe, P.: Students' response to traditional and computer-assisted formative feedback: A comparative case study. British Journal of Educational Technology **39**(3) (2008) 486–500
10. Black, P., Wiliam, D.: Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education) **21**(1) (2009) 5
11. Wiggins, G.: Seven Keys to Effective Feedback. Educational Leadership **70**(1) (September 2012) 10–16
12. Espasa, A., Guasch, T., Mayordomo, R., Martnez-Melo, M., Carless, D.: A dialogic feedback index measuring key aspects of feedback processes in online learning environments. Higher Education Research & Development **37**(3) (2018) 499–513

13. Narciss, S.: Designing and evaluating tutoring feedback strategies for digital learning. Digital Education Review (23) (2013) 7–26
14. Butler, A.C., Roediger, H.L.: Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. Memory & Cognition **36**(3) (Apr 2008) 604–616
15. Bälter, O., Enström, E., Klingenberg, B.: The effect of short formative diagnostic web quizzes with minimal feedback. Computers & Education **60**(1) (2013) 234–242
16. De Klerk, S., Veldkamp, B.P., Eggen, T.: The psychometric evaluation of a summative multimedia-based performance assessment. In: International Computer Assisted Assessment Conference, Springer (2015) 1–11
17. Novacek, P.: Confidence-based assessments within an adult learning environment. Int. Ass. for Development of the Information Society (2013) 403–406
18. Hench, T.L.: Using confidence as feedback in multi-sized learning environments. In Kalz, M., Ras, E., eds.: Computer-Assisted Assessment. Research into E-Assessment, Springer International Publishing (2014) 88–99
19. Lundeberg, M.A., Fox, P.W., Punćcohaŕ, J.: Highly confident but wrong: Gender differences and similarities in confidence judgments. Journal of educational psychology **86**(1) (1994) 114
20. Jonsson, A.C., Allwood, C.M.: Stability and variability in the realism of confidence judgments over time, content domain, and gender. Personality and Individual Differences **34**(4) (2003) 559–574
21. Burns, K.M., Burns, N.R., Ward, L.: Confidencemore a personality or ability trait? it depends on how it is measured: A comparison of young and older adults. Frontiers in psychology **7** (2016) 518
22. Ericsson, K.A., Krampe, R.T., Tesch-Römer, C.: The role of deliberate practice in the acquisition of expert performance. Psychological review **100**(3) (1993) 363–406
23. Christodoulou, D., Wiliam, D.: Making Good Progress?: The Future of Assessment for Learning. Oxford University Press (2017)
24. Roediger, H.L., Karpicke, J.D.: The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science **1**(3) (2006) 181–210 PMID: 26151629.
25. Petersen, K.A.: Implicit Corrective Feedback in Computer-Guided Interaction: Does Mode Matter? PhD thesis, Georgetown University (2010)
26. Wilske, S.: Form and Meaning in Dialog-Based Computer-Assisted Language Learning. PhD thesis, University of Saarland (2014)
27. Kerly, A., Hall, P., Bull, S.: Bringing chatbots into education: Towards natural language negotiation of open learner models. Know.-Based Syst. **20**(2) (2007) 177–185
28. Kane, D.A.: The role of chatbots in teaching and learning. In: E-Learning and the Academic Library: Essays on Innovative Initiatives, UC Irvine (2016) 1–26
29. Soliman, M., Guetl, C.: Intelligent pedagogical agents in immersive virtual learning environments: A review. In: MIPRO 2010 Proceedings of the 33rd International Convention, IEEE (2010) 827–832
30. MacTear, M., Callejas, Z., Griol, D.: The Conversational Interface: Talking to Smart Devices. Springer (2016)
31. DeSmedt, W.H.: Herr Kommissar: An ICALL conversation simulator for intermediate german. In Holland, V.M., Sams, M.R., Kaplan, J.D., eds.: Intelligent language tutors: Theory shaping technology. Routledge (1995)
32. Lu, C.H., Chiou, G.F., Day, M.Y., Ong, C.S., Hsu, W.L.: Using instant messaging to provide an intelligent learning environment. In: International Conference on Intelligent Tutoring Systems, Springer (2006) 575–583

33. Jia, J.: CSIEC: A Computer Assisted English Learning Chatbot Based on Textual Knowledge and Reasoning. Know.-Based Syst. **22**(4) (2009) 249–255
34. Höhn, S.: A data-driven model of explanations for a chatbot that helps to practice conversation in a foreign language. In: Proceedings of SIGDIAL 2017 Conference, ACM (2017)
35. Lyster, R., Ranta, L.: Corrective feedback and learner uptake. Studies in second language acquisition **19**(01) (1997) 37–66
36. Lyster, R., Saito, K., Sato, M.: Language teaching. Oral corrective feedback in second language classrooms. **46** (2013)
37. Amaral, L.A., Meurers, D.: On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. ReCALL **23**(01) (2011) 4–24
38. Gross, S., Pinkwart, N.: Towards an integrative learning environment for java programming. In: 2015 IEEE 15th International Conference on Advanced Learning Technologies (ICALT). (July 2015) 24–28
39. Perikos, I., Grivokostopoulou, F., Hatzilygeroudis, I.: Assistance and feedback mechanism in an intelligent tutoring system for teaching conversion of natural language into logic. International Journal of Artificial Intelligence in Education **27**(3) (2017) 475–514
40. Ryan, T., Henderson, M.: Feeling feedback: students emotional responses to educator feedback. Assessment & Evaluation in Higher Education **43**(6) (2018) 880–892
41. Palminteri, S., Khamassi, M., Joffily, M., Coricelli, G.: Contextual modulation of value signals in reward and punishment learning. In: Nature communications. (2015)
42. Ras, E., Baudet, A., Foulonneau, M.: A hybrid engineering process for semi-automatic item generation. In: International Computer Assisted Assessment Conference, Springer (2016) 105–116