ORIGINAL PAPER

CrossMark

Personalisation and automation in a virtual conversation skills tutor for children with autism

Marissa Milne[1] · Parimala Raghavendra[2] · Richard Leibbrandt[3] · David Martin Ward Powers[1]

Abstract

Existing research involving children with autism suggests that autonomous (self directed) virtual humans can be used successfully to improve language skills, and that authorable (researcher controlled) virtual humans can be used to improve social skills. This research combines these ideas and investigates the use of autonomous virtual humans for teaching and facilitating practice of basic social skills in the areas of greeting, conversation skills, listening and turn taking. The Social Tutor software features three virtual human 'characters' who guide the learner through educational tasks and model social scenarios. Participants used the software for 10–15 min per day, 3–5 days per week for 3 weeks, with data collected before software use commenced, immediately after use ended, then again 2 and 4 months after this. The software evaluation revealed that the social tutor was generally well-received by participants and caregivers, with data showing a clear difference in theoretical knowledge of social skills between the experimental (social content) group and the control (placebo content) group. This paper focuses on the automated assessment, dynamic lesson sequencing, feedback, and reinforcement systems of the social tutor software, the impact these systems had on participant performance during the software evaluation, and recommendations for future development. Thus, while data reflecting overall performance of the social tutor system is provided for context, the main data presented is limited to that specifically relating to the performance and user perceptions of the aforementioned systems.

Keywords Automated assessment · Personalised education · Autism · Social skills · Conversation skills · Embodied virtual agents

✉ Marissa Milne
marissa.milne@flinders.edu.au

Parimala Raghavendra
parimala.raghavendra@flinders.edu.au

Richard Leibbrandt
richard.leibbrandt@flinders.edu.au

David Martin Ward Powers
david.powers@flinders.edu.au

1. College of Science and Engineering, Flinders Digital Health Research Centre, Flinders University, GPO Box 2100, Adelaide 5001, South Australia

2. College of Nursing and Health Sciences, Disability and Community Inclusion, Flinders University, GPO Box 2100, Adelaide 5001, South Australia

3. College of Medicine and Public Health, Flinders University , GPO Box 2100, Adelaide 5001, South Australia

1 Introduction

People with autism spectrum disorders experience difficulties with social skills and can find understanding the nonverbal cues and social behaviours of other people challenging [22]. This makes building friendships and other appropriate relationships difficult, which can lead to isolation, social anxiety, and depression, impacting their overall wellbeing. Further, many young people with autism report an affinity for technology and exhibit high technology usage patterns [24,31]. Using virtual humans to teach social skills to children with autism harnesses this preference for technology and provides a tool that can support the development of social skill knowledge and behaviour, ultimately aiming to improve individuals' everyday wellbeing.

Existing research involving children with autism suggests that autonomous (self directed) virtual humans can be used successfully to improve language skills [6] and authorable (researcher controlled, 'Wizard of Oz' style) virtual humans

can be used to improve social skills [41]. In this context, authorable virtual humans are those requiring external input to interact with the user, such as a researcher observing the interaction and controlling the virtual human like a puppet, while autonomous virtual humans are fully self-contained and require no such external input. Here we combine these ideas and investigate the use of autonomous virtual humans for teaching basic social skills in the areas of greeting, conversation skills, and listening and turn taking.

The Social Tutor software developed for this purpose features three virtual human 'characters' namely a teacher, a peer with strong social skills, and a peer with developing social skills, who guide the learner through tasks and model social scenarios. The system also includes a mechanism to automatically assess users' task performance and present a personalised, dynamic lesson sequence tailored to their current needs. Additionally, the software features a multi-tier feedback, rewards, and reinforcement system in order to maximise retention of knowledge and generalization of skills to real-world contexts, a known difficulty for individuals with autism [29]. It is these assessment, feedback, and reinforcement systems that we focus on within the current paper.

2 Related work

To develop a successful virtual agent-based conversational social skills tutor, it is important to develop an understanding of what makes human tutoring so effective, which aspects can be implemented in a software context, and the specific requirements of individuals on the autism spectrum. It is also helpful to understand the strengths and challenges uncovered by existing technology-based interventions.

2.1 Human tutor behaviour

Peer tutoring is a well-established evidence-based practice that has been successfully used across a range of subjects, settings, and ages, and is shown to be effective for learners with and without disabilities [7].

There are two main schools of thought about what makes tutoring effective, the first being the tutor's ability to individualise instruction, and the second being the capacity for immediate feedback and prompting. In the first case, it has been hypothesised that the benefits are due to human tutors' ability to personalise their responses, provide in-depth individualisation of content, and determine the underlying reason for any difficulties their tutees are experiencing. In reality, it has been found that human tutors rarely perform these behaviours [32,42].

An alternative explanation along the same line of reasoning suggests that the benefits may stem from tutees' ability

to control the dialogue and ask questions to help them clarify and consolidate their understanding, however again it has been shown that tutees rarely take this initiative beyond simply confirming if a step or answer is correct [42].

The second hypothesis focuses on the benefits of immediate feedback and prompting, where students can work at their own pace until they make a mistake or reach a step they are unsure of, at which point the tutor can intervene to resolve the issue with a minimal loss of momentum. This helps reduce learner frustration, confusion, and the need for backtracking or re-doing work. It is this hypothesis that appears to account for much of the effectiveness of human tutoring [7,10].

This is excellent news for virtual tutoring systems such as the Social Tutor developed here, as they can easily be structured to support this process, particularly through the use of scaffolding techniques that ensure feedback is provided promptly and targeted to the specific skill being taught at the time.

2.2 Computer-assisted assessment

Research suggests that in order to be most effective, assessment should be integrated into the overall learning sequence rather than viewed as a separate activity, and should then be used to continually inform and adjust the activities presented to learners [4]. Computer assisted assessment within the Social Tutor occurs primarily at two different points; dynamic determination of the overall lesson sequence to be offered to learners, and immediate feedback within lesson tasks. To facilitate these processes, inspiration has been drawn from Shute and Towle's 'Learning Objects' approach [36] whereby all content in the intelligent tutoring system is broken down into small, self-contained, and reusable tasks that can be flexibly combined into meaningful and adaptive lesson sequences.

Shute and Towle argue that common methods of evaluating student mastery are insufficient, for example simply getting a set percentage or a certain number of consecutive assessment tasks correct [36]. Instead, they suggest the use of Bayesian inference networks (BINs) to provide probabilistic values which can be used to determine gaps or misunderstandings in the learners' knowledge. While this is a promising technique, unfortunately to successfully implement this requires tasks with open-ended or flexible answers, which in a software environment typically translates to writing paragraph-style answers. Producing such blocks of text is highly challenging for individuals with autism who experience both language and communication difficulties.

Following on from this, recent work has investigated the use of multimedia performance-based assessment, where users interact with a virtual scenario and the interaction data is fed into a BIN for assessment [21]. This research emerged after production of the Social Tutor software had concluded,

but appears promising for incorporation in the future, being both engaging for the learner and providing rich data that can be assessed using probabilistic techniques.

A related and common approach to assessing students' knowledge in existing autonomous tutoring systems is to use latent semantic analysis (LSA) to judge the semantic similarity of student responses to a provided 'ideal' response [17,18]. This has been found to provide grades on-par with those provided by human teachers. Unfortunately LSA relies on comparing paragraphs of text, again making it an unsuitable technique for use with individuals on the autism spectrum given their challenges with language and communication.

An alternative and more appropriate approach for this learner group and the current Social Tutor software is to use concept maps [26]. Concept maps are particularly applicable to autonomous tutoring software as they can be automatically assessed, and existing work has shown that paper-based graphic organisers can lead to strong learning gains in children with high functioning autism [13,34]. Given this, several concept map activities have been incorporated into the Social Tutor.

2.3 Existing technology interventions

A wide range of technology-based interventions exist to assist individuals with autism in various ways. Here a small sample of those that are either specifically focussed on teaching social skills or include a virtual character are included.

Use of virtual reality and augmented reality appears promising for assisting individuals with autism to learn new skills in a way that supports generalization to real-world environments, since learners are either entirely immersed in a simulated environment or interacting in the real environment with additional information overlayed on a display. Some success has been achieved using these approaches, however there are a number of barriers such as the cost of equipment and sensory issues involved in wearing goggles or helments [9], or the need for adult mediators to be involved in the activity [20].

There are numerous software programs in existence aimed at teaching social skills, however these often focus on the narrow domain of emotion recognition and, unlike the Social Tutor discussed here, are not designed to adapt to their users' needs nor do they employ autonomous virtual tutors [1,16,37,40]. Of the existing software that does incorporate autonomous virtual tutors, very little is designed for individuals with autism specifically, and the majority focuses on teaching skills with unambiguous answers that can be easily and automatically assessed by the software, for example tutors for reading [19,28], science [5], and mathematics [43].

There are a small number of autonomous virtual tutors that have demonstrated successful outcomes from evalua-

tions with individuals on the autism spectrum. These include Baldi and Timo [6] and the Sight Word Pedagogical Agent [35], both designed for improving vocabulary. For social skills teaching there are currently few virtual tutors available. One example is Andy the autonomous virtual peer in the ECHOES program [3], and another is Sam the authorable virtual peer [41]. Evaluations of Andy have shown promising anecdotal support, however no significant conclusions could be drawn as to its effectiveness in the initial study. Evaluations of Sam have shown improved social behaviours, however being authorable means that an adult mediator is again required to control Sam during interactions with learners. In both cases specialised equipment is required to run and display the programs, and thus both Andy and Sam are more suited to use in a classroom or clinic setting rather than independent use on a home computer.

3 Thinking Head Whiteboard

The Social Tutor is composed of two standalone programs, the virtual human software 'Head X' [23] and the lesson interaction and display software 'Thinking Head Whiteboard'. Three instances of Head X are used, one to display a teacher character, one to display a peer with strong social skills, and one to display a peer with developing social skills. The interface was designed to be simple and intuitive so that study participants could use it without outside assistance. A screenshot of the topic selection screen can be seen in Fig. 1.

3.1 System overview

The Thinking Head Whiteboard was developed specifically for the current research and controls the associated Head X instances via means of the Synapse interface that accompanies Head X. The Whiteboard software also reads in and interprets XML curriculum files, XML lesson files, and external media files in order to display the interactive content, and then records learner interactions with the system in XML log files so that user progress can be automatically assessed and an appropriate lesson sequence dynamically presented.

The Whiteboard contains support for displaying a wide range of activity types including the previously mentioned concept maps, along with drag and drop sorting activities, interactive role-plays performed by the virtual characters, support for Social Story^{TM} [14] style activities, and display of multimedia such as videos. Utilising a wide range of activity types is intended to increase both user engagement and the chances of users generalizing the skills learned in the Social Tutor to novel contexts and real-world situations.

The Whiteboard also performs several other functions including managing access to lesson content, presentation of the content quiz used for evaluation of the software, and
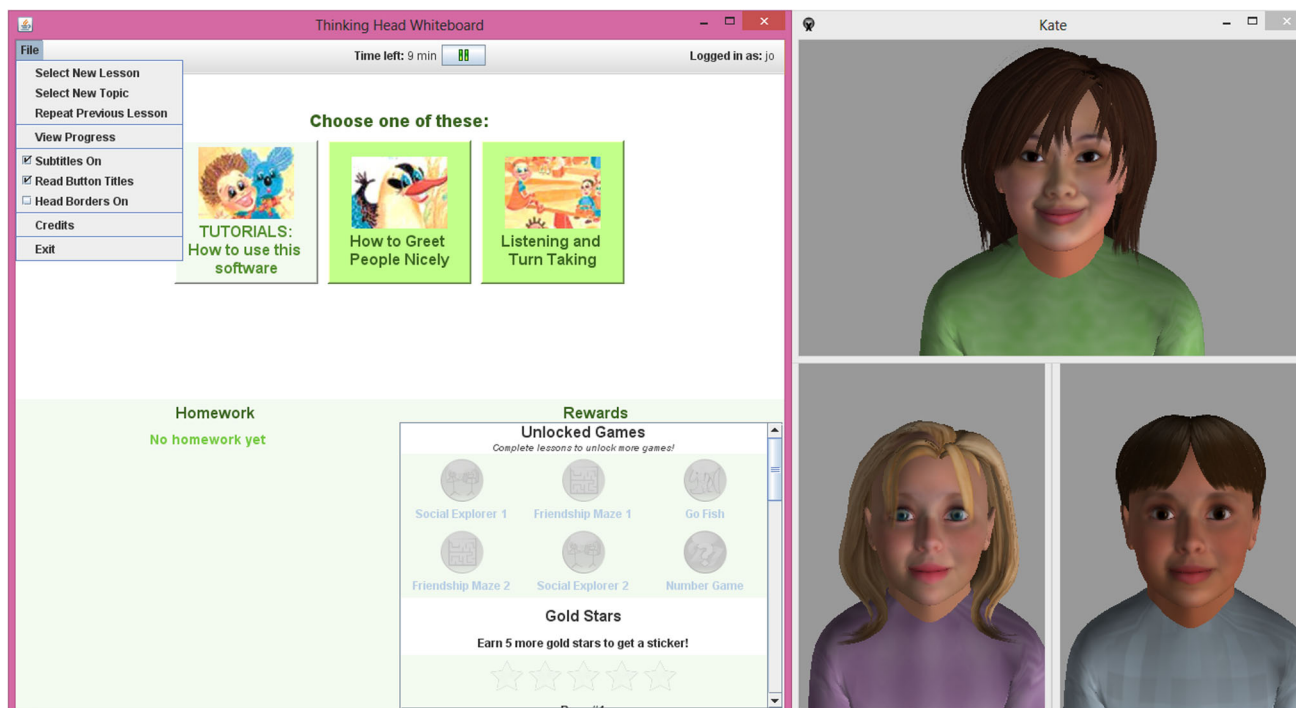
Fig. 1 Screenshot of the social tutor software showing the topic selection screen of the 'Thinking Head Whiteboard' on the left and the three 'Head X' instances on the right

supports lesson authoring by teachers and caregivers with non-technical backgrounds (see [27] for detail), however discussion of these functions is beyond the scope of this paper.

3.2 Automated assessment and lesson sequencing

The automated assessment functionality of the Thinking Head Whiteboard is a core feature, however the current implementation is very basic. As discussed in Sect. 2.2, BIN and LSA techniques have been found to be effective in existing tutoring systems, however their reliance on written answers and the incompatibility of that with the language and communication difficulties that accompany autism means that a simple heuristic-based algorithm has been implemented for the current iteration of the Social Tutor.

The XML curriculum file provides a broad guide for lesson sequencing, detailing which lessons are 'core' and to be completed by all learners, which are 'extra' and can be used for those who need additional practice, and the minimum correctness and accuracy learners must achieve for each lesson to be considered complete. Each XML lesson file also specifies any prerequisite lessons that must be completed before the one in question becomes available to the learner. This structure ensures that appropriate scaffolding takes place, and that learners are not offered complex content before sufficient mastery of prerequisite skills has been reached.

As learners interact with the system their performance on lesson tasks is continually recorded into an XML summary file. In conjunction with the XML curriculum file, this is used to determine which lessons to present to the learner next. Learners are presented with up to three lessons to choose from at a time, with tasks designated as 'core' given preference over 'extra' tasks, and unseen tasks given preference over those the learner has attempted before. The system can also take a step back in the lesson sequence if the user appears to be struggling, reiterating previous content until sufficient mastery is achieved to progress. Allowing users a selection of tasks provides a sense of control over their learning, which aims to increase learner engagement while still ensuring they are guided towards choosing activities that are appropriate for their current needs.

3.3 Feedback, rewards and reinforcement

A review of existing work found that researchers often fail to explicitly program for generalization and maintenance, and those that do often employ a "train and hope" approach [29]. The Social Tutor incorporates a number of features aimed at motivating students to continue to use the software and to support generalization of skills to novel contexts, with the simplest being that the virtual characters themselves are encouraging, providing immediate feedback in the form of

simple praise or prompts as learners interact with their activities.

Research shows that providing praise, reward, or punishment alone only has a small influence on educational outcomes [15], however feedback with helpful suggestions can be very beneficial for learners. In the Social Tutor, if students make a mistake while working through a task, the teacher character provides them with immediate constructive feedback. Where possible, if the student continues to make mistakes the hints become more detailed. Finally, the virtual teacher typically provides a short 'recap' at the completion of each activity so that the key steps of the target skill are continually reinforced.

To provide extrinsic reinforcement the Social Tutor also incorporates a three tiered rewards system. In the first tier, students gain a gold star for each completed lesson. This step is primarily intended to provide learners with a sense of progress. In the second tier, students can trade five gold stars in for a virtual 'sticker' and work towards completing their sticker collection. This is intended to act as a simple reward, while not detracting too much time from core learning activities. The third stage involves reward activities, where games reinforcing the curriculum content are unlocked at 50% and 100% completion of each topic. This tier is intended to be a more substantial reward, reflective of the effort and perseverance learners will have put in to reach this point.

A homework system encouraging students to practice their skills outside of the software has also been implemented. Homework activities become unlocked once students complete prerequisite activities within the software that introduce the steps of the target skill. Existing research suggests that opportunities to practice new social skills with neurotypical peers outside of the intervention context can be very beneficial for facilitating generalization of theoretical knowledge to real-world scenarios [33].

4 Method

Recent reviews of the methodologies commonly used when evaluating social skills interventions for children with autism have led to the identification of some recurring issues in existing research [29,33]. Wherever possible, our study aimed to meet the recommendations resulting from the identified reviews.

4.1 Experimental design

Key issues identified in existing research include a lack of studies using a control group, few studies involving more than ten participants, and a lack of blinded observer ratings [33]. In response to these issues, our study included an experimental group where participants were explicitly taught social skills, and a control group where participants received no social content. We aimed to recruit sixteen participants to each of these groups for a total cohort of thirty two participants, and caregivers were not provided with details of the differences in the two groups or informed which group their child had been assigned to until after all data completion was complete.

Despite generalization and maintenance being two widely acknowledged issues for individuals with autism in relation to interventions of any kind, at the time of the reviews many existing studies were still failing to explicitly address either of these. It was found that few studies measured for maintenance effects at all post-intervention, and those that did only did so once, and then rarely beyond 3 months post-intervention [29,33]. To address this, our study explicitly measured generalization of skills to real-world contexts using the Vineland Adaptive Behaviour Scales (Vineland-II) [39], and did so not only before and immediately after the active software use period, but also at both 2 and 4 months post-intervention.

4.2 Inclusion criteria

All participants had an existing diagnosis of autism spectrum disorder, were aged 6–12 years old, and were attending a mainstream school at the time of the study. These inclusion criteria ensured that the participant group was sufficiently homogeneous for the data to be informative, that the content presented was both age and context appropriate for all participants, and allowed a minimum level of communication skill and general functioning to be assumed when selecting target topics and designing lesson activities. The software was aimed at individuals considered to have 'high functioning autism' or 'Asperger Syndrome' under the DSM-IV [2]. Families also had to have access to a suitable computer for the duration of the study, with all families opting to use their personal home computers.

4.3 Group allocation

A matched pairs approach was taken when allocating participants to the experimental or control group. Participants were matched on age according to three 'buckets'–6 to 8 years olds, 9 and 10 years olds, and 11 and 12 years olds. For consistency and to ensure unbiased assignment, group allocation was conducted when the participant's caregiver confirmed their appointment for the researcher to visit their home. The first participant was placed into the experimental group, then the next participant, if their age fell into the same bucket as the first participant, was matched with them and placed in the control group. If the second participant's age did not fall into the same bucket, they were instead allocated into the experimental group.

Siblings were kept in the same group as each other to avoid caregivers guessing which group their children had

been assigned to, and to avoid issues relating to fairness and jealousy between siblings. Three pairs of siblings were involved in the study, with two sets allocated to the experimental group and one to the control group.

4.4 Experimental and control software

The software used by participants in both the experimental and control groups was identical, and consisted of three instances of the virtual character software Head X and one instance of the Thinking Head Whiteboard software run simultaneously. Only the specific lesson content presented to the groups was different. This was to ensure that the experience of both groups was as similar as possible.

Participants allocated to the experimental group were presented with short lessons that explicitly taught conversational social skills. These skills included greeting others, listening and taking turns, and beginning, ending and maintaining conversations. In order to provide sufficient depth and breadth of content to learners it was determined that a combination of learning materials were required, and three curricula were selected for this purpose.

The 'Playing and Learning to Socialise' (PALS) curriculum [11] is aimed at kindergarten aged children and provides fundamental skills, while the 'Skillstreaming' [25] and 'Social Decision Making/Social Problem Solving' (SDM/SPS) curricula [12] build on the foundation PALS provides and extend it with more advanced instruction. All three curricula have been empirically evaluated, including having demonstrated efficacy for children on the autism spectrum or with related special needs.

Participants allocated to the control group received no explicit social skills training and were instead presented with a series of mazes to complete which increased in length and difficulty as the participant progressed, ranging from easy single page mazes such as that shown in Fig. 2, through to medium, hard and extra hard mazes where the mazes have more obstacles, multiple pages, or require more than one item per obstacle to pass through.

In both the experimental and control conditions the lessons were grouped into themed topics which the participant could select from, and the rewards system functioned identically with a virtual gold star being collected for each lesson that the participant completed. Reward games in the experimental condition aimed to reinforce lesson content, including turn-taking games such as 'Go Fish' and 'Guess My Number' and games that require the user to interact with cartoon characters in pro-social ways, while reward games in the control condition contained no social content, instead offering retro-style games such as 'Tetris' and 'Snake'.

Similarly, both groups were offered homework activities to complete outside of their software use time, with the experimental group being asked to practice the skills they were

Fig. 2 Example of a control group maze activity

learning within the software such as greeting someone new or asking someone their name, while the control group were asked to complete tasks such as reading, drawing, or designing their own maze.

4.5 Data collection tools

Considerable effort was made to minimise the burden on families participating in the study, therefore the selection of measurement tools favoured those that could be completed electronically and independently by the participant and caregiver. To ensure the impact of the Social Tutor software was accurately reflected in the data collected, a multifaceted approach was taken.

As previously discussed, generalization to novel contexts is a known difficulty for individuals with autism, with children often learning how to 'do the intervention' without applying what they are learning to situations outside of the intervention context. Thus, assessing both theoretical level understanding and real-world skill performance is critical. Given this, theoretical knowledge was measured using the bespoke in-software content quiz, which consisted of activities similar in structure to those given throughout the 3 weeks of software use but with novel content, and application of skills to real-world situations was measured using the Vineland-II [39], which was presented as a Google Form.

The Vineland-II is flexible in delivery, with only the sections most relevant to the current study needing to be administered, which assisted in reducing unnecessary burden on caregivers. The standardisation process also specifically targeted several clinical populations, and thus it has been validated for use with children on the autism spectrum [8,30,38].

Table 1 Overview of evaluation schedule

	Pre-test	Software use	Immediate post-test	2 months follow up post-test	4 months follow up post-test
Timeline	Day 1	Ongoing for 3 weeks	The day after software use ends	2 months after post-test	2 months after second post-test
Researcher visit	Yes	No	On request	On request	On request
Caregiver actions	Vineland-II	Support child to use software if required	Vineland-II and questionnaire	Vineland-II	Vineland-II
Participant actions	Content quiz and questionnaire	Software use: 10–15 min a day, 3–5 times a week	Content quiz and questionnaire	Content quiz	Content quiz
Researcher actions	Install software and support participant and caregiver to complete their tasks		Remind caregiver to complete assessments	Remind caregiver to complete assessments	Remind caregiver to complete assessments.
					When all data received, provide unlock instructions and payment.

To provide better insight into how participants interacted with the software, how much content they were able to cover in their active software use period, and the profile of this interaction, the software itself also continually recorded log data reflecting participant interactions and performance.

In addition to these measures, participants completed a pre-test questionnaire to provide an indication of their expectations and computing expertise prior to using the software, and both participants and caregivers were asked to complete post-test questionnaires reflecting on their experience with the software and what they would like to see changed, added or removed in future iterations of the Social Tutor. The questionnaires were designed to provide insight into what aspects were best received and where difficulties arose, as well as providing direction for future development. All questionnaires were delivered using Google Forms.

4.6 Evaluation procedure

The evaluation was designed so that only a single initial visit by the researcher to the participants' home was necessary, with all data collection performed electronically either via the Social Tutor software itself or Google Forms. During the researcher's visit, a discussion of the study process took place and informed consent was obtained, then the researcher installed and tested the software on the family computer. The participant and caregiver completed their pre-test data collection tasks while the researcher was present to ensure that any difficulties or misunderstandings could be addressed immediately.

Following this visit, participants were asked to use the software for one 10–15 min session per day, 3–5 days per week, for 3 weeks. The Social Tutor software was designed to automatically control activities so that participants could access their lessons only for the designated 3 weeks period, and would be automatically presented with the content quiz at the appropriate intervals.

Further, the software incorporated a timer to alert participants when their minimum 10 min of software use for the day was up, and automatically exited once the maximum of 15 min was reached, allowing participants to self-manage their time with the software and reducing the burden on caregivers. It should be noted that the software would only auto-exit once the user moved away from their current lesson activity and back to the lesson selection screen, ensuring that participants were not abruptly cut off mid-task.

As discussed previously, generalization and maintenance of skills from the intervention to the real world must be explicitly addressed and measured in any intervention aimed at individuals with autism. To allow maintenance to be measured across time, data was collected immediately prior to software use, at the end of the 3 week software use period, and at both 2 and 4 months post-intervention. A summary of the evaluation schedule can be seen in Table 1.

To ensure that all data collection tasks were completed on time, the researcher contacted families with an email reminder a few days before each set was due, including links to the appropriate set of Google Forms along with reminders of participant IDs and other details necessary for performing the data collection. On completion of the final content quiz the software automatically unlocked the user's account and their lesson activities became accessible again. Once all data was complete and received by

the researcher, the family were given a small payment for their participation and were provided with the necessary instructions for them to access the version of the software that they were not initially assigned to. Families were able to keep the software at the conclusion of the evaluation period.

5 Results and discussion

Results from the software evaluation indicated that the Social Tutor was well received overall by both participants and caregivers, and was generally used in the manner intended, with software log data and participant and caregiver post-test questionnaire responses indicating that it was used productively by participants and that the intended amount of time was spent on both individual lessons and on the software as a whole, both per session and over the intervention period.

Post-test questionnaire responses also indicated that participants felt the virtual characters were friendly, the software was beneficial to their learning, and that it was overall easy to use, meeting the goal of providing a non-judgemental learning environment, although more game-like elements and personalisation capabilities were requested for future iterations of the software.

Recruited participants Only participants who completed at a minimum both the pre-test and post-test content quiz or both the pre-test and post-test Vineland-II were included in the final cohort. This resulted in thirty one children being included, with sixteen in the experimental group ($M = 8.81$ years, $SD = 1.83$) and fifteen in the control group ($M = 9.20$ years, $SD = 2.08$). An independent samples t-test found no significant difference in mean age between the two groups ($p = 0.497$, $r = 0.10$). Participant gender was not specifically controlled for, however a 2-sample t-test for equality of proportions was carried out and found no significant difference in ratio between the two groups, $\chi^2(1, N = 31) = 0.51$, $p = 0.47$. This confirms that the experimental group and control group were sufficiently similar for meaningful comparisons to be made.

5.1 Whole-system outcomes

The total percentage correctness was taken from the pre-test and post-test content quiz data and the difference between them calculated. The results for the control group showed only a very small mean change from pre-test to post-test ($M = 1.14\%$, $SD = 7.16$, 95% CI $[-3.41, 5.69]$), with paired t-test showing this to be statistically insignificant; $t(11) = 0.56$, $p = 0.587$, $d = 0.16$. In contrast, analysis of data from the experimental group resulted in a larger mean

change from pre-test to post-test ($M = 7.36\%$, $SD = 9.05$, 95% CI $[2.54, 12.19]$), with paired t-test indicating statistical significance; $t(15) = 3.25$, $p = 0.005$, $d = 0.81$. This suggests that use of the Social Tutor software directly led to gains in social skills knowledge, a very encouraging finding.

It should be noted that due to the small sample sizes, t-tests were determined to be the appropriate analysis tool, however for completeness a repeated measures two-way ANOVA was also conducted. This indicated a significant interaction in correctness scores between group and test period (i.e. pre-test, post-test, 2 months post-test or 4 months post-test) $F(3, 30) = 3.421$, $p = 0.03$, however simple main effects analysis found no significant differences between the control and experimental groups or between test periods. This is likely due to the small sample sizes and supports the use of t-tests for analysis of this data.

While not initially anticipated, post-hoc analysis of content quiz data also led to the discovery of response subgroups, where it was found that a quarter of experimental group participants ($n = 4$) responded particularly well to the Social Tutor software, markedly improving their correctness scores from pre-test to post-test while on average completing fewer lessons and spending less time in total using the software. In contrast, half of the experimental group ($n = 8$) made modest improvements in their correctness scores while on average completing more lessons, and the remainder ($n = 4$) fell into a low responding subgroup who made negligible improvements in correctness scores despite spending the most time on the software of all subgroups.

5.2 Assessment and reinforcement systems

The assessment and reinforcement systems were designed to increase participant motivation and engagement, support learners to improve their theoretical social understanding at their own pace, and to facilitate participants to convert this new knowledge into real-world behavioural changes. As discussed in Sect. 5.1 above, analysis of the content quiz data indicates that improvements in theoretical understanding were successfully achieved.

Analysis of the post-test questionnaire data indicated that overall participants and caregivers responded positively to the Social Tutor, with the average response for most Likert-style questionnaire items falling between the neutral to strongly agree range, as can be seen in Fig. 3. Interestingly, for items concerning enjoyment and ease of use, control group participants and caregivers rated the Social Tutor more positively than experimental group participants. For enjoyment items this is to be expected given that control group participants received game-like maze activities instead of targeted social skills lessons, and for usability questions the responses are likely a reflection of the more complex nature of the experimental group lesson content, which would make
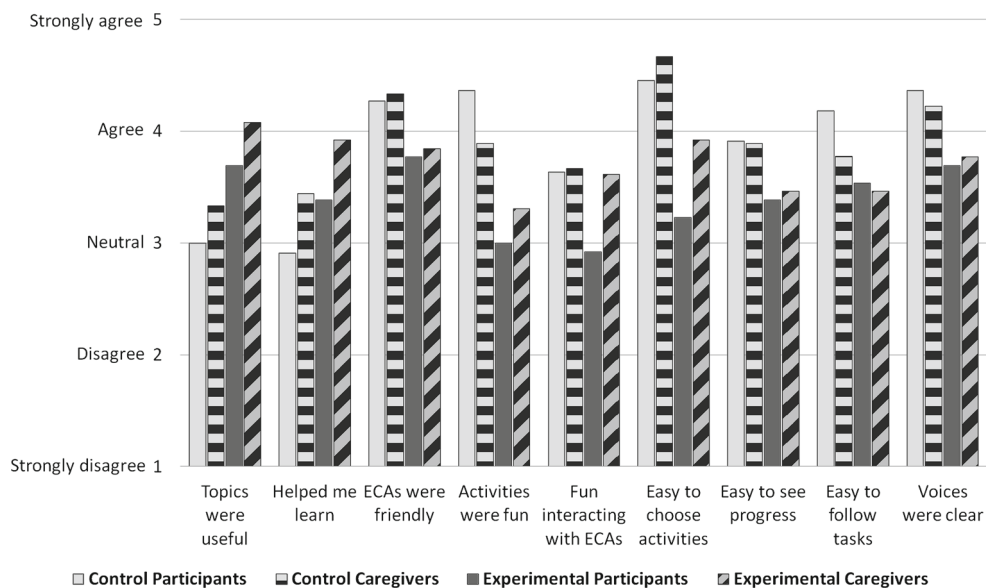
Fig. 3 Mean responses to Likert-style post-test questionnaire items

any shortcomings of the software more apparent. For example, due to the nature of the activities being explained, the spoken dialogue of the virtual humans was shorter and more repetitive in the control group than the experimental group. If participants found the synthetic voices difficult to understand or unpleasant to listen to, this would have impacted more on the experimental group than the control group, despite the fact that the same voices were used in both cases.

Data from the post-test questionnaire indicated that some participants found certain learning tasks to be too hard or confusing, with a few participants experiencing this with the concept map style activities. This may be because tutorial lessons explaining how to do concept maps and other activity types were optional, and while explanations were available on the page of the activity itself, participants may not have realised how to access them or may have found the instructions insufficient. This suggests that further refinement of activities and more supportive hints when users are experiencing difficulties are needed. Related to this, a caregiver also suggested more feedback be included. Several mechanisms are in place with the aim of providing this kind of information to participants, however there is scope for more in-depth and targeted feedback to be incorporated. For example, the virtual teacher currently praises the participant for correct answers and provides a short recap at the end of each activity, but the interaction could be made more meaningful by having the virtual tutor detect when the user is getting frustrated or displaying misunderstandings, and stepping in with additional hints or information at critical points.

In contrast to reports of tasks being too difficult in the open ended survey responses, in the related rating scale item participants and caregivers in the experimental group reported activities to be between 'just right' (a rating of 3) and 'a little too easy' (a rating of 2) ($M = 2.38$, $SD = 1.12$ and $M = 2.85$, $SD = 0.90$ respectively). These seemingly contradictory responses suggest that both of these states may have been true at various times throughout the evaluation, or may simply be reflective of the heterogeneity of individuals on the autism spectrum.

From the log data that was automatically collected by the Social Tutor software, it was observed that the third core topic of 'Beginning, Ending and Maintaining Conversations' was not attempted by any of the participants. This topic required completion of the 'Listening and Turn Taking' topic before it was unlocked and made available, and from inspection of the data it appears that this did not occur for any participant. This provides further evidence that the 3 week period of active software use was too short, and that a richer analysis may result from a longer evaluation.

As discussed in Sect. 5.1, response subgroups were identified within the experimental group data. Particularly relevant here is the low responding subgroup. These participants started with higher pre-test scores than other participants and displayed negligible changes to their results at posttest. From observation of the log data, it was found that these participants also displayed more erratic patterns of topic and lesson selection. One possible explanation for these patterns is that the Social Tutor may not have been offering these participants suitably advanced or challenging content.

Pairing the observation of low responding subgroup behaviour with the fact that during the 3 weeks intervention period no participant managed to complete enough prerequisite activities to unlock the third and final topic of 'Beginning, Ending and Maintaining Conversations' indicates that the current implementation of the automated assessment and dynamic lesson sequencing system is insufficient to meet user needs in all cases. Improving the system to ensure that participants with higher pre-existing knowledge can move more quickly through the content and skip activities that are too basic for them has the potential to improve educational outcomes and engagement levels.

At the post-test survey, six participants and one caregiver noted that they felt that the sticker and rewards system was a particular highlight of the software, making this the most highly repeated positive comment across all categories. This supports the inclusion of an extrinsic rewards system to encourage users to persevere with their lessons. While the themes of the virtual sticker sets on offer were chosen to reflect common interests of children within the target age range, one caregiver suggested making the sticker sets customisable. This could also be done with the 'gold star', and in both cases would help to ensure the items on offer are as motivating as possible for the learner.

Related to this, the user log data that was automatically collected by the Social Tutor revealed that only a few participants managed to unlock any of the reward games during the 3 week active software use period. These games currently unlock at 50% and 100% topic completion, however adding additional reward games and lowering the amount of completed content required to access them, for example unlocking games at approximately 30% intervals, is likely to be more motivating for learners.

As previously discussed, optional homework lessons were incorporated to reinforce lesson content and encourage participants to generalize their knowledge to real-world contexts. Questionnaire responses indicated that these lessons were viewed neutrally overall by the experimental group, with one caregiver citing them as a strength of the system, and one caregiver and one participant citing them as a challenge. From anecdotal conversations and observations it appears that this feature has the potential to become a powerful tool for connection between the software and real-world experience, but requires further refinement to achieve this potential.

Particularly noteworthy was the participant who successfully performed their homework task once, but when prompted by their caregiver to repeat the skill on a different occasion, did not understand why they should since their homework was already done. This highlights that homework tasks as standalone, one-off activities are not sufficient, but by instead approaching homework as an ongoing activity, possibly earning a reward each time a target behaviour is enacted,

it may have the potential to turn newly learned behaviours into beneficial, ongoing habits.

Together the Likert-style and open ended questionnaire results indicate that, while the Social Tutor was well received overall, the user experience for experimental group participants could still be improved. Along with suggestions for more gamification of lesson content, the automated assessment and lesson sequencing system appears to be a prime candidate for further refinement.

6 Future development

Following the results of the software evaluation, it is recommended that the lesson sequencing algorithm be updated to be more adaptive and to draw from a student model, rather than simply being based on a set of heuristics and static rules as it currently is. This could enable it to adapt to individual learning styles, do a better job at determining when a student is ready to progress to the next complexity level, and potentially be more targeted when a student is struggling, for example by giving them content that addresses their specific difficulty rather than just generically back-tracking them and repeating content they have already progressed through. Related to this, the virtual teacher behaviour could also be modified to have it provide more targeted feedback during activities, which may help avoid the need for back-tracking altogether in some cases.

In the current implementation of the Social Tutor all learners begin at the same point, and this results in higher achieving learners having to complete a substantial amount of basic content before they can unlock the advanced lessons that are more likely to suit their needs. An approach that may assist students to reach the appropriate point within the Social Tutor content more quickly may be to feed the results of the initial content quiz directly into the assessment system.

Each question within the initial quiz has been designed to directly assess a particular skill or concept, and each aligns with a specific topic within the curriculum. Therefore, if a student achieves 100% correctness on a quiz question this could be used to mark the associated lessons as complete, effectively fast-tracking the student through the Social Tutor content. In addition to this, short 'pop quizzes' could also be used for the same purpose, possibly offered when the system detects that a student is performing particularly well or set at specific intervals.

To support inclusion of a rich student model, continuous assessment would be required. This is already built into the Social Tutor to an extent, with log data being recorded for every interaction a student has with the software. In addition to this log data, both reflective and affective information could be gathered and incorporated into the student model. One simple way to gather reflective data could be to periodi-

cally prompt users to complete quick self-assessments rating how they feel they are progressing. While this is not a reliable measure of social skill competency, self-reflection is a valuable skill in itself, and has been shown to support deep learning and long-term retention of skills [15].

From the post-test questionnaire, it was identified that the existing dynamic lesson sequencing algorithm can actually be a cause of user frustration. The algorithm prioritises new, previously unseen activities and, if available, offers them before re-offering activities the user has tried before. For some participants this is frustrating, as they may feel ready to retry a particular activity they have seen before, but which is not currently offered to them due to this lesson sequencing algorithm, for example in the case where the software has three or more 'new' activities to present to the user instead. Modifying the way this works to have two lists of activities, one being the three 'suggested' activities the current algorithm identifies, and a second easily accessible but perhaps less prominent list of 'previously attempted' activities may be a possible solution.

The behaviour of this 'previously attempted' list would need to be managed strategically to ensure it did not simply contain lessons that have been accessed recently, and thus result in learners continually repeating their favourite tasks and not progressing through the broader content. For example, lessons already completed to a sufficient level of mastery could be displayed in a less prominent manner, such as being grayed out or shown further down the list, and mechanisms could be put in place so that learners are only able to access each lesson once or twice a session, or are blocked from repeating the same lesson twice in a row, forcing them to alternate their favourite activities with other tasks.

Also from the post-test questionnaire, caregivers have indicated that they would appreciate more targeted feedback, specifically suggesting that a summary of the skill steps at the end of each lesson would be beneficial. While this is already provided in many instances, it clearly needs to be implemented more widely and explicitly throughout the system. In addition to this worthwhile suggestion, ensuring that feedback is provided not only at task-level but also process-level is recommended to give users a chance to consolidate their learning more effectively, and combining this with self-assessment and reflection as discussed above may assist to further develop learners' reasoning and problem solving skills.

7 Conclusion

A major motivator for the use of autonomous tutoring systems is the capacity to present content and lesson sequences tailored to individual learners' needs. However, care must be taken to ensure that the assessment techniques used in this process are appropriate to the target population and the skills being taught. Common approaches such as Bayesian inference networks and latent semantic analysis are generally not suitable for use with learners on the autism spectrum as they rely on the generation of large blocks of text, and individuals with autism commonly experience severe difficulties with communication and language skills. In developing the Social Tutor described here, it therefore became necessary to implement alternative approaches, thus automatically assessed concepts maps and simple heuristics were employed.

While the dynamic lesson sequencing system implemented in the Social Tutor functioned sufficiently for many participants as reflected in positive post-test questionnaire results, log data and user feedback indicate that there is still room for improvement. In particular, a mechanism should be implemented to ensure participants are quickly placed at a difficulty level appropriate to their needs. This is especially important for users with stronger pre-existing skills, who in the current Social Tutor implementation must still complete a substantial amount of 'basic' content before they can proceed to the more 'advanced' content. Possibilities here include feeding the results of the initial content quiz directly into the system rather than having all users begin at the same point, and periodically offering short 'pop quizzes' to fast-track learners if the system detects they are performing well.

In developing any autonomous tutoring system it is also necessary to understand the broader needs and challenges of the target population. For individuals with autism, generalization to novel contexts is a core challenge that has been observed across numerous intervention types. The Social Tutor implemented a number of techniques to facilitate this process, however more still needs to be done to ensure that changes in social skill knowledge gained from the Social Tutor translate into observable changes in their social behaviour. One feature of the Social Tutor that appears very promising but was under utilised in the current implementation is that of homework. Providing more extrinsic rewards to encourage users to engage with homework tasks, and having these as repeated activities rather than one-off tasks may encourage learners to apply their newly learned theoretical knowledge to real-world situations more reliably.

Overall the software developed here was successful in increasing users' theoretical knowledge of social skills and was perceived positively by participants and caregivers alike. With further refinement to the dynamic lesson sequencing system and implementation of additional support for generalization, it is hoped that the Social Tutor can offer an engaging and powerful means for improving the social skills of children on the autism spectrum.

References

1. Abirached B, Zhang Y, Aggarwal JK, Tamersoy B, Fernandes T, Carlos J (2011) Improving communication skills of children with ASDS through interaction with virtual characters. In: 2011 IEEE 1st international conference on serious games and applications for health (SeGAH), pp 1–4

2. Association AP (2000) Diagnostic and statistical manual of mental disorders: DSM-IV-TR. 4th ed., text revision edn. American Psychiatric Association, Washington, DC

3. Bernardini S, Porayska-Pomsta K, Smith TJ (2014) Echoes: an intelligent serious game for fostering social communication in children with autism. Inf Sci 264:41–60

4. Black P (2015) Assessment: friend or foe of pedagogy and learning. Past as Prologue, p 235

5. Blair K, Schwartz D, Biswas G, Leelawong K (2007) Pedagogical agents for learning by teaching: teachable agents. Educ Technol 47(1):56

6. Bosseler A, Massaro D (2003) Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. J Autism Dev Disord 33(6):653–672

7. Bowman-Perrott L, Davis H, Vannest K, Williams L, Greenwood C, Parker R (2013) Academic benefits of peer tutoring: a meta-analytic review of single-case research. School Psychol Rev 42(1):39–55

8. Carter AS, Volkmar FR, Sparrow SS, Wang JJ, Lord C, Dawson G, Fombonne E, Loveland K, Mesibov G, Schopler E (1998) The vineland adaptive behavior scales: supplementary norms for individuals with autism. J Autism Dev Disord 28(4):287–302

9. Cheng Y, Huang CL, Yang CS (2015) Using a 3d immersive virtual environment system to enhance social understanding and social skills for children with autism spectrum disorders. Focus Autism Other Dev Disabil 30(4):222–236

10. Chi M, Siler S, Jeong H, Yamauchi T, Hausmann R (2001) Learning from human tutoring. Cognit Sci 25(4):471–533

11. Cooper J, Goodfellow H, Muhlheim E, Paske K, Pearson L (2003) PALS social skills program: playing and learning to socialise: resource book. Inscript Publishing, Launceston

12. Elias MJ, Butler LB (2005) Social decision making/social problem solving for middle school students: skills and activities for academic, social, and emotional success. Research PressPub, Ottawa

13. Finnegan E, Mazin AL (2016) Strategies for increasing reading comprehension skills in students with autism spectrum disorder: a review of the literature. Educ Treat Child 39(2):187–219

14. Gray C (2001) How to respond to a bullying attempt: what to think, say and do. The Morning News 13(2)

15. Hattie J, Timperley H (2007) The power of feedback. Rev Educ Res 77(1):81

16. Hopkins IM, Gower MW, Perez TA, Smith DS, Amthor FR, Casey Wimsatt F, Biasini FJ (2011) Avatar assistant: improving social skills in students with an asd through a computer-based intervention. J Autism Dev Disord 41(11):1543–1555

17. Hu X, Xia H (2010) Automated assessment system for subjective questions based on LSI. In: IEEE, pp. 250–254

18. Jackson G, Guess R, McNamara D (2010) Assessing cognitively complex strategy use in an untrained domain. Top Cognit Sci 2(1):127–137

19. Jacovina ME, Jackson GT, Snow EL, McNamara DS (2016) Timing game-based practice in a reading comprehension strategy tutor. In: International conference on intelligent tutoring systems, Springer, Berlin, pp 59–68

20. Ke F, Im T (2013) Virtual-reality-based social interaction training for children with high-functioning autism. J Educ Res 106(6):441–461

21. de Klerk S, Eggen TJHM, Veldkamp BP (2016) A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in a bayesian network. Comput Hum Behav 60:264–279

22. Kroncke AP, Willard M, Huckabee H (2016) What is Autism? History and foundations. Springer International Publishing, Cham, pp 3–9

23. Luerssen M, Lewis T, Powers D (2011) Head X: customizable audiovisual synthesis for a multi-purpose virtual head, Lecture Notes in Computer Science, vol 6464, book section 49, pp. 486–495. Springer, Berlin

24. MacMullin JA, Lunsky Y, Weiss JA (2016) Plugged in: electronics use in youth and young adults with autism spectrum disorder. Autism 20(1):45–54

25. McGinnis E, Goldstein AP (2012) Skillstreaming the elementary school child: a guide for teaching prosocial skills, 3rd edn. Research Press, Champaign

26. Meyer J, Land R (2010) Threshold concepts and troublesome knowledge (5): dynamics of assessment. Threshold concepts and transformational learning. Routledge, London, pp 61–79

27. Milne M, Leibbrandt R, Raghavendra P, Luerssen M, Lewis T, Powers D (2013) Lesson authoring system for creating interactive activities involving virtual humans the thinking head whiteboard. In: 2013 IEEE symposium on intelligent agent (IA), pp 13–20

28. Mostow J (2005) Project listen: a reading tutor that listens. http://www.cs.cmu.edu/~listen/research.html

29. Neely LC, Ganz JB, Davis JL, Boles MB, Hong ER, Ninci J, Gilliland WD (2016) Generalization and maintenance of functional living skills for individuals with autism spectrum disorder: a review and meta-analysis. Rev J Autism Dev Disord 3(1):37–47

30. Perry A, Flanagan HE, Dunn Geier J, Freeman NL (2009) Brief report: the vineland adaptive behavior scales in young children with autism spectrum disorders at different cognitive levels. J Autism Dev Disord 39(7):1066–1078

31. Putnam C, Chong L (2008) Software and technologies designed for people with autism: what do users want? In: Proceedings of the 10th international ACM SIGACCESS conference on computers and accessibility, assets '08, ACM, New York, NY, pp 3–10

32. Putnam RT (1987) Structuring and adjusting content for students: a study of live and simulated tutoring of addition. Am Educ Res J 24(1):13–48

33. Rao PA, Beidel DC, Murray MJ (2008) Social skills interventions for children with aspergers syndrome or high-functioning autism: a review and recommendations. J Autism Dev Disord 38:353361

34. Roberts V, Joiner R (2007) Investigating the efficacy of concept mapping with pupils with autistic spectrum disorder. Br J Spec Educ 34(3):127–135

35. Saadatzi MN, Pennington RC, Welch KC, Graham JH, Scott RE (2017) The use of an autonomous pedagogical agent and automatic speech recognition for teaching sight words to students with autism spectrum disorder. J Spec Educ Technol 32:173–183

36. Shute V, Towle B (2003) Adaptive e-learning. Educ Psychol 38(2):105–114

37. Silver M, Oakes P (2001) Evaluation of a new computer intervention to teach people with autism or asperger syndrome to recognize and predict emotions in others. Autism 5(3):299–316

38. Sparrow SS (2011) Vineland adaptive behavior scales. Springer, New York, NY, pp 2618–2621
39. Sparrow SS, Cicchetti DV, Balla DA (2005) Vineland-II: vineland adaptive behaviour scales-survey forms manual. NCS Pearson, Bloomington
40. Sturm D, Peppe E, Ploog B (2016) emot-ican: Design of an assessment game for emotion recognition in players with autism. In: 2016 IEEE international conference on serious games and applications for health (SeGAH), IEEE, pp 1–7
41. Tartaro A (2007) Authorable virtual peers for children with autism. In: CHI '07 extended abstracts on human factors in computing systems, CHI EA '07, ACM, New York, NY, pp 1677–1680
42. VanLehn K (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychol 46(4):197–221
43. Woolf B, Arroyo I, Muldner K, Burleson W, Cooper D, Dolan R, Christopherson R (2010) The effect of motivational learning companions on low achieving students and students with disabilities. In: Intelligent Tutoring Systems, Springer, Berlin, pp 327–337