User Modeling and User-Adapted Interaction (2021) 31:183–223
https://doi.org/10.1007/s11257-020-09278-0

Using scaffolding to formalize digital coach support for low-literate learners

Dylan G. M. Schouten[1] · Pim Massink[2] · Stella F. Donker[2] · Mark A. Neerincx[1] · Anita H. M. Cremers[3]

Received: 5 December 2019 / Accepted in revised form: 12 September 2020 /
Published online: 14 October 2020

Abstract

In this study, we attempt to specify the cognitive support behavior of a previously designed embodied conversational agent coach that provides learning support to low-literates. Three knowledge gaps are identified in the existing work: an incomplete specification of the behaviors that make up 'support,' an incomplete specification of how this support can be personalized, and unclear speech recognition rules. We use the socio-cognitive engineering method to update our foundation of knowledge with new online banking exercises, low-level scaffolding and user modeling theory, and speech recognition. We then refine the design of our coach agent by creating comprehensive cognitive support rules that adapt support based on learner needs (the 'Generalized' approach) and attune the coach's support delay to user performance in previous exercises (the 'Individualized' approach). A prototype is evaluated in a 3-week within- and between-subjects experiment. Results show that the specified cognitive support is effective: Learners complete all exercises, interact meaningfully with the coach, and improve their online banking self-efficacy. Counter to hypotheses, the Individualized approach does not improve on the Generalized approach. Whether this indicates suboptimal operationalization or a deeper problem with the Individualized approach remains as future work.

Keywords Virtual learning environment · Embodied conversational agent · Scaffolding · User modeling · Design research · Requirements engineering

✉ Dylan G. M. Schouten
dylan.schouten@gmail.com

[1] Delft University of Technology, Delft, The Netherlands

[2] Utrecht University, Utrecht, The Netherlands

[3] TNO Soesterberg, Soesterberg, The Netherlands

1 Introduction

People of low literacy struggle to independently participate in information societies (Buisman and Houtkoop 2014). Limited information (reading and writing) and communication (speaking and understanding) skills lead to participation issues, which can be cognitive, affective, or social in nature (Schouten et al. 2016). Cognitive issues relate to applying information and communication skills and possessing general knowledge about society. Affective issues relate to fear, shame, and low self-efficacy. Social issues relate to lack of motivation and trust in others. These issues can be addressed by providing societal participation learning that is grounded in *crucial practical situations* (real-life participation scenarios that involve the skills and knowledge needed to participate in society independently, such as online banking, grocery shopping, or engaging with local government; cf. Kurvers and van de Craats (2007); van de Craats (2007)), which allows low-literate learners to practice skills and gain knowledge and experience in a practical context of use. For this learning to be *effective*, especially for learners with limited information and communication skills, such as low-literate learners, the learning must be accessible (barriers to entry are lowered or removed), the learning experience must be positive (learners can and want to engage with the learning), and learners must reach desired learning outcomes (Schouten et al. 2017a). We aim to provide effective learning with VESSEL: a *Virtual Environment to Support the Societal participation Education of Low-literates* (Schouten et al. 2016, 2017a, 2020). VESSEL consists of situated, interactive exercises in the societal participation domain, and an autonomous, rules-driven Embodied Conversational Agent (ECA) coach that supports low-literate learners before, during, and after these exercises with cognitive, affective, and social learning support (see Fig. 1).

We use the socio-cognitive engineering method (SCE, cf. Neerincx et al. 2019; Neerincx 2011; Neerincx and Lindenberg 2008) in the development of VESSEL.

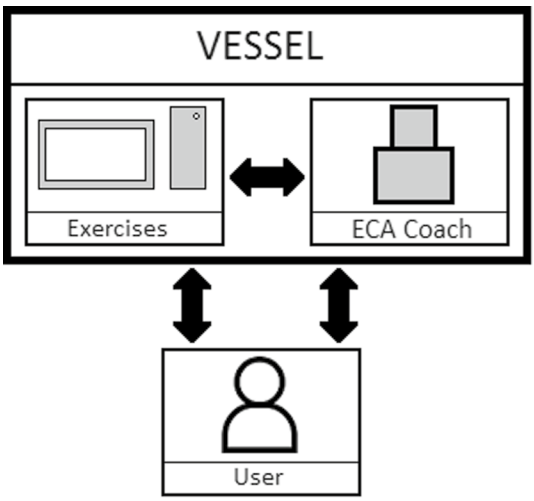

Fig. 1 Envisioned VESSEL design. Arrows indicate system interactions: the user performs exercises, the ECA coach monitors exercise state and user-system interaction, and the coach supports the user as appropriate. Image from Schouten et al. (2020)

The SCE method is an iterative software design and development method that moves (nonlinearly) through three phases, shown in Fig. 2. In the *foundation* phase, relevant operational demands (the software system's context of use), human factors data (theory relevant to user–system interactions), and technology (both technology currently in the system and envisioned technology) are combined into a foundation of data. In the *specification* phase, a requirements baseline is created containing requirements, claims, system objectives, and use cases. This is then used for the *evaluation* phase, where the validity of the specification is empirically tested. Evaluation results are used to iteratively update the foundation and refine the specification.

Previous work used a high-level requirements baseline (see Table 2) to develop a first VESSEL prototype, consisting of an ECA coach that offered three kinds of learning support for four exercises (easy and hard 'online banking' and 'service desk conversation' exercises, cf. Schouten et al. 2020). Cognitive support based on scaffolding, a teaching method that provides the right level of support at the right time (van de Pol and Elbers 2013), was offered during the exercises. Affective support based on motivational interviewing, a counseling technique that focuses on behavioral change (Miller and Rollnick 2009), was given after the exercises. Social support based on small talk, a form of social interaction important for building trust (Cassell and Bickmore 2003), was used before the exercises. All support was provided in the form of prerecorded spoken utterances and controlled by an operator, using the Wizard-of-Oz method to act as an ECA behind the scenes (cf. Maulsby et al. 1993). Notably, support was both created and provided in an *informal* manner. Support utterances were created based on an expert walkthrough of the system: researchers determined areas where low-literates would likely struggle and wrote utterances to address the predicted issues. And during the exercises, the Wizard-of-Oz operator interpreted user actions and speech and selected the utterance(s) considered best in this situation. Evaluation showed that the ECA coach resulted in a more positive cognitive, affective, and

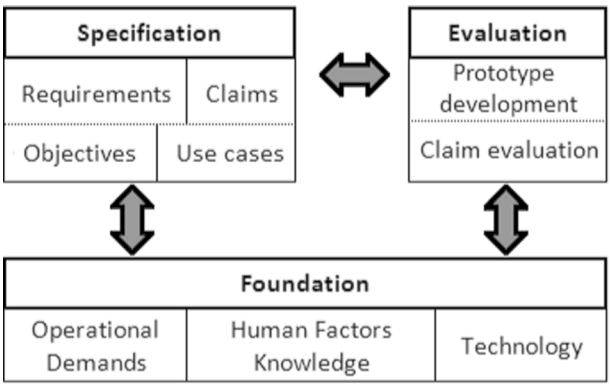

Fig. 2 Socio-cognitive engineering method used in this study. Double-sided arrows between the foundation, specification, and evaluation boxes indicate that development can move to any phase at any time (Neerincx et al. 2019; Neerincx 2011; Neerincx and Lindenberg 2008)

social learning experience, and higher self-efficacy about difficult online banking scenarios. As proof of concept, this shows that VESSEL can improve learning effectiveness for low-literate learners.

As the results from Schouten et al. (2020) were promising, the next development step is to create a *formal* design specification that accurately describes VESSEL's envisioned functionality as automated learning support. This involves two things: first, writing a comprehensive set of dialogue rules for the ECA coach's cognitive, affective, and social support behavior, which can be applied by automated computer support without requiring human interpretation, and second, incorporating new functionality as needed to improve support provision and learning effectiveness. Each of the three support types needs a separate refinement step. We focus on the coach's *cognitive* support in the present study, as effective cognitive support is necessary to ensure learners can understand the system and complete exercises. Affective support and social support are left to later work.

Our current implementation of cognitive support has three relevant knowledge gaps which the formalization process must address. First, because the existing set of coach support utterances is based on a noncomprehensive expert walkthrough, the utterances do not yet structurally and comprehensively cover the exercises. Not all challenging exercise elements have associated support utterances, and the existing utterances contain different levels of information and direct guidance, with no clear underlying logic. Formalized support will require a comprehensive set of support utterances for each exercise, in which the utterances cover every relevant aspect of the exercise and in which they are comparable in terms of information provided. Second, the coach's speech recognition functionality requires further operationalization. As the current speech recognition is left up to the Wizard operator's interpretation of user utterances and context, there are no formal rules in place to specify what learner utterances the coach should react to, and how. Formalized support will require a clear, unambiguous speech recognition ruleset. Third, we expect that *personalizing* cognitive learning support will substantially improve learning outcomes. But our current implementation of cognitive support does not have a coherent and unequivocal specification of *how* this support can be personalized. We hypothesize that (in concert with the above) VESSEL's learning effectiveness could be improved by incorporating *user modeling* (the process by an intelligent system infers user traits from user–system interaction, cf. Fischer 2001; Stephanidis 2001; Shute and Zapata-Rivera 2012; Horvitz et al. 2013) to better adapt the offered support to individual learners' circumstances and needs. To achieve this, formalized support will require a clear user model of support need, including an unambiguous list of user actions relevant to this model and a description of changes to the coach's support provision over time that can be made on the basis of this.

In this work, we aim to design and evaluate a VESSEL prototype that offers formalized cognitive learning support. Four steps are needed. First, we update the VESSEL foundation in three ways. We update operational demands by designing exercises based on crucial practical situations that demand cognitive support. We update human factors knowledge by incorporating more detailed scaffolding theory, as well as theory concerning user modeling. And we update technology by describing the envisioned role of speech recognition. Second, we refine the VESSEL specification: we operationalize the

foundation theory into a comprehensive set of coach dialogue rules, update the requirements baseline, and write a use case to illustrate expected findings. Using the refined specification, we define in what ways the coach can provide cognitive support based on the learner's progress in the current exercise. We call this approach to support provision the 'Generalized' approach. We also describe how the coach models the learner's skill level based on their performance, and how it can use this model to attune its support provision in later exercises. We call this the 'Individualized' approach. Third, we design and develop a VESSEL prototype, consisting of an ECA coach that can offer cognitive learning support along both the Generalized and Individualized approaches, and three online banking exercises. This prototype will be designed for use in a Wizard-of-Oz experimental setup, in which an operator applies the coach's support behavior and speech recognition behind the scenes by selecting prescribed outputs for the computer-sensed inputs (Maulsby et al. 1993). Fourth, we experimentally evaluate the prototype with low-literate learners. We investigate how the new prototype affects the cognitive, affective, and social learning experience and learning outcomes, compared to our previous work, and we investigate whether using both the Generalized and Individualized approaches leads to higher learning effectiveness than only using the Generalized approach. This leads to the following research questions:

- *Q1 Design*. How can we create a formal design specification for VESSEL that incorporates rules for cognitive learning support provided by an ECA coach?
  - *Q1a* Which operational demands, human factors knowledge, and technologies are needed to write these rules?
  - *Q1b* Which functionalities, interaction methods, and appearances should the ECA coach have to reflect this specification?
- *Q2 Evaluation*. What is the learning effectiveness impact of a VESSEL prototype that offers cognitive learning support according to the formal specification?
  - *Q2a* Are the learning effectiveness results of this prototype comparable to the VESSEL prototype that offered informal cognitive, affective, and social learning support?
  - *Q2b* Does using both the Generalized and Individualized approaches to learning support result in higher learning effectiveness than using only the Generalized approach?

The structure of this paper is as follows. Section 2 provides the refinement of the sCE foundation, necessary for deriving the concrete design specification in Sect. 3. Section 4 describes the resulting new VESSEL prototype. Sections 5 and 6 describe, respectively, the experiment that evaluates the prototype and the evaluation results. Section 7 presents conclusions and directions for future work.

2 Foundation

2.1 Operational Demands: Exercises

To accurately evaluate the effectiveness of cognitive learning support, exercises are needed that pose a significant cognitive challenge and demand coach support, but that can be completed with this support. If the exercise is too easy, learners will not require support; if the exercise is too difficult, no level of support will be effective. The first VESSEL prototype (Schouten et al. 2020) contained four exercises: an easy exercise and a hard exercise about online banking, and an easy exercise and a hard exercise about visiting a government service desk. Of these, only the hard online banking exercise meets our needs: the exercise was challenging and demanded significant coach support, but participants often completed it. For this prototype, three new challenging online banking exercises were created, using the 'Hard Online Banking' Web site from Schouten et al. (2020) as a task environment. In Exercise 1, the user must transfer money from their checking account to a webshop. In Exercise 2, the user must report a change of address to their bank. In Exercise 3, the user must transfer money from their savings account to their checking account. All exercises are intended to be equivalently challenging. To achieve this, we ensured that each exercise had the same number of *critical waypoints*, which we defined as those exercise steps that a learner *must* take to successfully complete it. In the context of online banking, critical waypoints can either be *navigation waypoints* (getting to the right part of the online banking Web site at the right time) or *data entry waypoints* (entering the right information in the right place). Each exercise was designed with exactly four navigation and four data entry waypoints, presented in the same order: three navigation waypoints, then four data entry waypoints, then one last navigation waypoint. All exercises come with written summary instructions showing the goal and necessary information, such as bank account number and money amount to transfer, or street name and postal code of a new address.

2.2 Human Factors Knowledge: Scaffolding

Three core elements of scaffolding are contingency, fading, and transfer of responsibility (van de Pol et al. 2010). *Contingency* refers to matching support to the learner's current ability. Three types of contingency are identified: domain contingency, instructional contingency, and temporal contingency. *Domain contingency* means ensuring that the exercise or (sub)task has the right level of challenge for the learner. Exercise challenge level should fall in the Zone of Proximal Development (Vygotsky 1980; Wood and Wood 1996). Mislevy et al. claim that: '*... the most accurate information about a test taker is obtained when the level of difficulty is close to the test taker's level of performance. However, there is also an important experiential aspect (...) Items that are too hard demoralize the test taker, while items that are too easy bore her.*' (Mislevy et al. 2014, p. 112). In VESSEL, we use exercise design to aim for domain contingency, as shown in Sect. 2.1.

Instructional contingency refers to tailoring the amount of support to the learner's skill level. This is derived from constructivist views of learning, which claim that learners actively construct knowledge and meaning by interacting with their environment (Berger and Luckmann 1966; Jonassen 1991). Learners should complete as much learning by themselves as possible for optimal outcomes (Johnson 2005; van de Pol and Elbers 2013), and they should attribute success to themselves instead of external sources, as this raises self-efficacy (Bandura 1997). Support should not take over too much responsibility too quickly. In VESSEL, we reach instructional contingency by categorizing the coach's support utterances into two categories: *Proactive* and *reactive* utterances. The coach can use *proactive utterances* when it detects that the learner needs support (e.g. by observing that learners have not made progress for some time). This is necessary because learners in tutoring sessions often do not actively ask for help (Graesser et al. 2011; Graesser and Person 1994). We use van de Pol et al. (2010)'s overview of scaffolding tools to define five proactive utterance subcategories: a proactive utterance can be a *prompt* (a simple question to gauge the learner's knowledge level), an *explanation* (an answer to either an earlier prompt or a learner question), a *hint* (an implicit suggestion of what the learner should do next that references the correct next step), an *instruction* (an explicit description of what the learner should do next), or *modeling* (an offer to demonstrate what the learner should do next, followed by the coach actually demonstrating it). Each of these utterance types provides support at a different level of directness. We define *support level* as a measure of the amount of direct guidance in a support category; support levels go from 1 (prompt) to 5 (modeling) as shown in Table 1. The coach can use *reactive utterances* to respond to learner speech or actions (described in detail in Sect. 2.4). Finally, the coach can give feedback based on learner progress. If the learner attempts to move to the next exercise waypoint and has taken all necessary steps *correctly*, the coach uses praising feedback; if the learner has taken any steps *incorrectly*, the coach uses corrective feedback to indicate that something went wrong. See Table 1.

Temporal contingency describes that support should be given at the right time, when the learner is confused or questioning (Wood 2001; Wood and Wood 1996). If support is provided too late, learners are frustrated by a lack of progress; if it comes too quickly, learning is impaired (Johnson 2005) and learners might resent the support for giving an answer they could have found themselves (D'Mello and Graesser 2012). In VESSEL, we reach temporal contingency by defining when the coach should use support utterances. For proactive utterances, we define that the coach should wait a certain amount of time between utterances (to avoid information overload and give learners a chance to parse and react to the utterance): we call this amount of time the *support delay*. We set a support delay of 20 s based on timing analysis of our previous work (Schouten et al. 2020). Reactive utterances should be used as soon as the appropriate conditions are met, in order to be useful (Gibbs et al. 2004).

Fading refers to gradually lowering the amount of offered support over time, as the learner's skill improves. Traditionally, human tutors use scaffolding by setting difficult exercises and immediately providing 'heavy' scaffolding (quick proactive guidance with a high support level, cf. Lepper and Woolverton 2002), and

Table 1 VESSEL ECA coach cognitive support categories

Support category	Description	Example
Proactive support		
Support level 1: Prompt	This utterance asks the user either whether they know the meaning of a particular keyword or whether they understand the next exercise step	"Do you know what 'online banking' means?"
Support level 2: Explanation	This utterance either answers a preceding prompt on the same topic or answers a direct user question about a particular keyword or exercise step.	"Online banking' means: doing banking, on your computer."
Support level 3: Hint	This utterance tells the user that their current action or position in the exercise is not correct and provides oblique direction: The utterance contains one explicit keyword that references the next step the user should take, but does not outright say that this is the case.	"You cannot change your address on this page. Can you see where you can change your personal information?"
Support level 4: Instruction	This utterance directly tells the user what action they should take, as an imperative statement. It uses the same keyword as the preceding hint.	"Click on the word: 'personal information'."
Support level 5: Modeling	This utterance offers to demonstrate the right action to the user.	"Shall I show you where you should go?"
Reactive support		
User utterance: Recognized keyword	The user asks a question that uses a keyword the coach recognizes. The coach provides an 'explanation' support utterance for that keyword.	"Coach, where do I go to do online banking?" "'Online banking' means: doing banking, on your computer."
User utterance: Unrecognized	The user asks a question that does not use any recognized keywords. The coach uses a general reaction utterance to indicate they do not understand.	"Coach, how do I make an account on this website?" "I'm sorry, I cannot help you with this."
User action: Correct	The user moves to the next exercise waypoint correctly. The coach tells the user they have done this.	(*if the user moves to the 'Personal Information' page*) 'Well done! The right page for you is "Personal Information"!'
User action: Incorrect	The user attempts to move to the next exercise when not all correct steps have been taken. The coach tells the user they have made a mistake	(*if the user fills out the wrong address and then tries to submit their address change*) 'Sorry, you have not yet filled out all information correctly.'

Describes exact rules for creating utterances to match each proactive and reactive support level and includes example utterances used to explain the phrase 'online banking' and the exercise step 'find the page where you change your personal information'

then lowering that heavy scaffolding as learners start performing better. However, previous work has shown that low-literate learners have strong negative emotional reactions to unexpected challenge and to exercises that exceed their self-confidence and self-efficacy (Schouten et al. 2017a). A system that starts out with heavy challenge and heavy scaffolding may lead to learners 'giving up,' and either quitting the exercise or relying on the coach to model everything. In VESSEL, we structure our support the other way around: support starts as low as possible and builds up to the level that learners need to proceed. To define when each type of support is given, we must first determine the likely moments and locations in the exercise that learners will need support for. We have used Bloom (1956)'s taxonomy of keywords and Bayles (2004)' overview of online banking critical factors to find all potentially *difficult elements* of the Web site: all pages and links that a learner can potentially click on, and all complex words and terms on pages that the learner must navigate through to complete the exercise. One proactive support utterance of each support level must exist for each difficult element. One utterance of each level is also needed for each critical waypoint of each exercise. We can then define our fading: for every difficult element, the coach must always start proactive support at support level 1 and increase that level every time the learner needs support again for that same element. Support levels are tracked per difficult element, meaning that a higher support level for one element does not impact other elements. Support levels can only go up, never down.

Transfer of responsibility means that learners must take their own responsibility for the success of the learning process. In VESSEL, this follows automatically from all other scaffolding steps. As learners move through an exercise, proactive support always starts at a low support level and gradually increases, encouraging learners to overcome challenges by themselves instead of waiting for help. Reactive support triggers on learner questions, encouraging learners to actively seek help when needed. And the coach's support delay ensures the gradual lessening of proactive support as learners become more capable of doing everything alone.

2.3 Human Factors Knowledge: User Modeling

User modeling refers to the notion of intelligent systems inferring user traits from observable user–system interaction. Fischer (2001) defines a user model as '*models that systems have of users that reside inside a computational environment*' (p. 70). User models can enable and support advanced user–system interaction by (i.a.) providing user-specific accessibility options (Stephanidis 2001), limiting the functionality a program provides to match inferred user needs without overloading them (Fischer 2001; Horvitz et al. 2013), and informing users of interaction possibilities and functions that they were not aware of (Fischer 2001; Stephanidis 2001; Bhowmick et al. 2010). In the specific context of education and learner support, user models are used to (i.a.) enable adaptive educational and e-learning systems (Ciloglugil and Inceoglu 2012; Tadlaoui et al. 2016), personalize online learning environments (Kaya and Altun 2011), and support learners with particular information access and modality needs (Benmarrakchi et al. 2017). Note that not all instances of system

adaptation to user behavior count as or involve user modeling. For instance, VESSEL's cognitive support model (Sect. 2.2) already uses user actions to drive its decision making. However, this is more accurately *task* modeling, not user modeling: the system in this instance is only interested in supporting the user with a specific task in a specific moment, not in building a long-term model of that user.

We aim to employ user modeling in VESSEL to improve learning effectiveness. Specifically, we are interested in adapting the aforementioned support delay to the user's overall performance with the exercises. Lehman et al. (2008) suggest that struggling learners must be helped along quickly and decisively, which we hypothesize we can do by lowering the delay. Conversely, we hypothesize that increasing the delay for successful learners gives them more time to complete exercises themselves, which will lead to optimal self-efficacy gains by encouraging transfer of responsibility. In both cases, this adaptation should be automatic, or driven by the system, rather than human-invoked (Stephanidis 2001).

We create a small, simple user model for VESSEL that encompasses the entire possibility space of all exercises. This is possible because VESSEL forms a relatively compact 'closed-world' system (cf. Fischer 2001), and we can clearly define an optimal path through and an optimal outcome for each exercise. The user model consists of two elements: the user's overall support delay value and the user's performance in previous exercises. Whenever the user completes a new exercise, the model evaluates their performance in this exercise, and the learner's need for support, by looking at the types and amount of support they needed to pass each critical waypoint in the exercise. If the user passed most waypoints with no support at all, or with prompt or explanation support, their performance in the exercise is rated 'good,' and the model increases their support delay by a certain amount. If the user mostly needed instruction and modeling support, their performance is 'bad,' and the model decreases their support delay. If the user passed most waypoints with hint support, their performance is 'medium': the balance between challenge and support is right for this user, so their support delay is not changed.

The user model thus outlined serves several purposes. First, using this model, VESSEL can quickly and unobtrusively adapt itself to individual learners. This allows us to present a simple unified VESSEL design at design time, but easily adapt to the needs of users at use time (Fischer 2001; Stephanidis 2001). Second, the model allows VESSEL to reach each user's optimal support delay over time, defined as the support delay in which the user consistently falls in the 'medium' category. As user skill levels improve over time, VESSEL will automatically follow suit. Finally, over longer periods of use, the model would allow us to track users' support delay progress and exercise performance over time, enabling more accurate learning assessment. However, this level of application lies outside the scope of the current work.

2.4 Technology: Speech Recognition Rules

In VESSEL, speech recognition is necessary to enable reactive coach support to learner questions (see Table 1). The coach can answer questions about the current

exercise by recognizing particular *keywords*. We create a dictionary of *known keywords*, which consists of the critical waypoints and difficult elements of each exercise. If the learner says something out loud, the coach checks whether any words in the learner's utterance match one of its keywords. If a known keyword is detected, the coach gives explanation-level support about that keyword. If the learner's utterance does not contain any known keywords, it is classified as *unrecognized*. In this case, the coach uses a general reaction utterance to indicate lack of understanding, using phrases such as '*I do not understand what you said*.' Additionally, the coach can understand the learner utterances '*yes*' and '*no*,' allowing it to parse learner answers to questions (see Table 1). It can also understand the category of all learner utterances that indicate lack of understanding, such as '*I did not understand that*' and '*Could you repeat what you said*,' which ensures that the system is accessible to learners who struggle with quickly interpreting spoken utterances (which includes low-literate second-language learners, cf. Schouten et al. 2017a).

3 Specification

3.1 Operationalization

In two steps, we translate the updated foundation into comprehensive rules for our ECA coach. First, we formally operationalize the coach's support behavior during exercises to create the Generalized approach. While the learner works through an exercise, the coach starts a timer that tracks the amount of time that has passed since its last support action. This timer runs continuously regardless of what the learner does, with one exception: the timer is paused whenever learner and coach engage in *learner–coach interaction*, which we define as any dialogue in which both the coach and the learner speak at least once, and the learner's utterances are in reaction to the coach's. Any dialogue that meets these criteria is defined as one occurrence of learner–coach interaction, regardless of length or number of exchanges, with the interaction ending if the learner and the coach do not say anything for 5 s. The timer is temporarily paused while the interaction is ongoing, and resumes when the interaction ends. When the timer exceeds the coach's support delay value, it checks what difficult element the learner is currently interacting with and which critical waypoint the learner should be trying to reach. The coach then gives the proactive support utterance at the support level of that critical element and resets the timer. If the learner interacts with a difficult element in any way before the support delay value is reached, the coach also resets the timer. If the learner triggers a reactive support utterance (by saying something out loud, or interacting with a waypoint correctly or incorrectly), the coach gives the appropriate utterance and resets the timer. The coach moves through this loop until the exercise is completed. Figure 3 shows the Generalized approach as a decision tree.

Second, we operationalize the Individualized approach, which uses the user model to attune the value of the support delay to learner performance in between exercises. In this study, we define that the support delay will always be increased or decreased by exactly 5 s. The support delay starts at 20 s for every learner; it can be

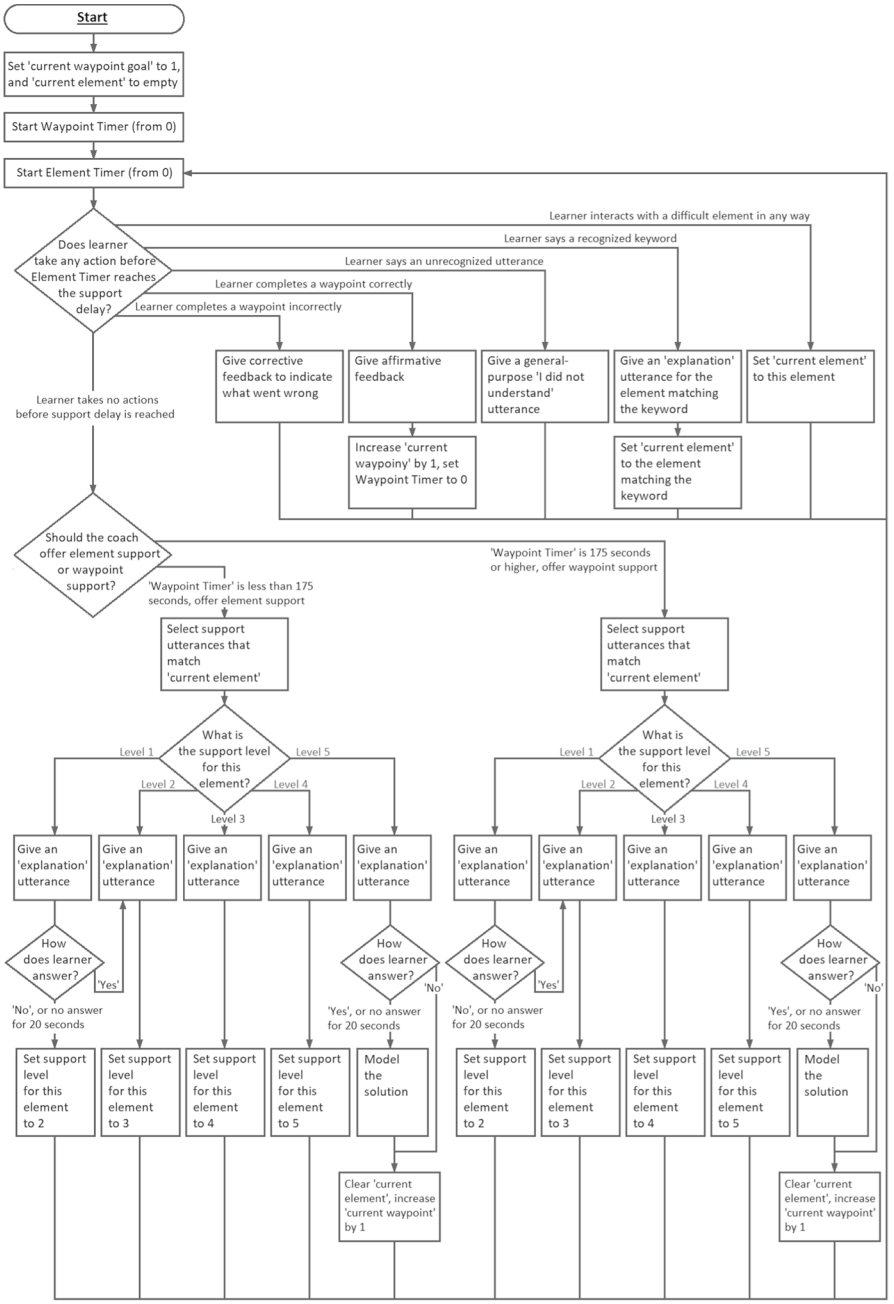

Fig. 3 Generalized approach rules decision tree. The value of '20 s' used here represents the standard support delay

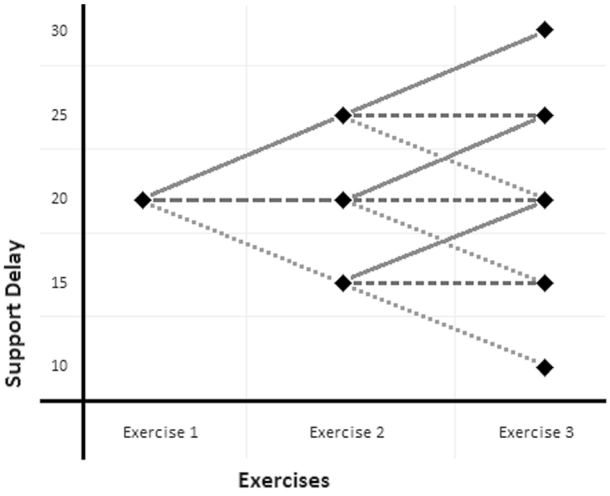

Fig. 4 Timing schema for the Individualized approach over three exercises. Filled lines represent a learner with 'good' performance, resulting in the support delay being raised, dotted lines represent a learner with 'bad' performance, resulting in the support delay being lowered, and dashed lines represent a learner with 'medium' performance, resulting in the support delay not changing

raised to a maximum of 30 or lowered to a minimum of 10. See Fig. 4 for a visualization of the Individualized approach.

3.2 Requirements Baseline

Here, we *refine* the existing VESSEL requirements baseline to reflect the updated support behavior rules; this means we update (expand/rewrite) the text of the existing requirements to better reflect our new understanding of the design of VESSEL and that we write new subrequirements where necessary. We refine only those requirements that change on the basis of these rules, for the *coach* aspect of VESSEL, the *exercises* aspect, or both. Requirements that are not described in this section stay unchanged. Table 2 presents the refined requirements baseline.

Requirement **R1. Adaptability** is refined for both the coach and the exercises. The coach should ensure that the support delay best matches the needs of individual learners, using the Individualized approach to attune the delay according to the rules in Sect. 3.1 and Fig. 4 (**R1.1-C**). And the exercises should be sufficiently challenging to learners. Exercises should exist for different skill and difficulty levels, but these should be neither too easy nor too hard (**R1.1-E**). This can only be evaluated after exercises have been put into practice: an exercise is too easy if learners need little or no coach support to complete it (support on average not exceeding level 1), and it is too hard if learners need strong coach support to complete every step (support on average exceeding level 4). When designing difficulty, it should be kept in mind that the coach's support can lower the difficulty of a too challenging exercise, but not raise the difficulty of a too easy one.

Table 2 Refined VESSEL requirements baseline based on contingency rules

Exercise requirements	Coach requirements	General requirements
R1. Adaptability. VESSEL should offer and/or support different learning styles and preferences. The focus of adaptability should be on providing the right level of difficulty (as perceived by the learner). Exercises should be difficult enough to be useful, but not so difficult that they scare low-literate learners off	**R1.1-C.** The coach should adapt its interaction style to individual user needs, wishes, and learning goals. Coach support should ensure that exercises fall inside the Zone of Proximal Development: exercises should neither be too easy nor too difficult. *Cognitive support should be offered following the support rules model of prompt, explanation, hint, instruction, and modeling. And support should be offered at a learner-appropriate delay*	**R1.1-E.** The exercises should each have a specific difficulty level, tailored to particular skill training and learning goals. The total corpus of exercises should span a range of difficulty levels. *Exercises should always be challenging and built on the assumption of coach support*
R2. Sensitivity. VESSEL should use non-confrontational language and content, demonstrate cultural awareness, and take existing emotional issues with regard to reading and writing and societal participation into account. The principal emotional barriers to address with sensitivity are fear, shame and anger. Low-literate learners should feel emotionally comfortable and experience being taken seriously	**R2.1-C.** The coach should always address learners calmly and kindly and avoid using phrases and broaching topics that upset low-literate learners	**R2.1-E.** The exercises should be as sensitive as needed to reach the intended learning goals and difficulty level
R3. Situatedness. VESSEL should use learning materials and contents that are closely related to the learner's physical environment and real-life experiences. Correctness of experience is the most important part of situatedness: the experience of training must be as close as possible to the real-life situation being trained. Learning exercises must teach low-literate learners to deal with cognitively, affectively, and socially challenging situations		**R3.1-E.** The exercises should use content drawn from crucial practical situations, tailored to and situated in the specific day-to-day experiences of low-literate learners

Table 2 (continued)

Exercise requirements	Coach requirements	General requirements
R4. Collaboration. VESSEL should have systems in place that enable, support, and foster social interaction and collaboration in learning. For low-literates, it is preferable to have collaboration come from nondigital sources. If collaboration is built into the software, it must emphasize the availability of teachers and low-literate peers		
R5. Multimodality. VESSEL should employ multimodality, offering content in multiple concurrent ways. Modality use must be adapted to individual preferences and to particular exercises. Using more modalities is better than using just one	**R5.1-C.** The coach should combine audio 'speech' with visual and textual supporting material	**R5.1-E.** The exercises should be as multimodal as needed to reach the intended learning goals and difficulty level
R6. Support. VESSEL should possess built-in support options. It is important to invoke the feeling of being supported. The right individual level of support must be found: too little support drives low-literate learners off, but too much support hampers learning and trades progress for comfort	**R6.1-C. R6.1-C.** The coach should use *dialogue rules based* on verbal scaffolding to offer cognitive learning support **R6.2-C.** The coach should use motivational interviewing techniques to offer affective learning support **R6.3-C.** The coach should use small talk to offer social learning support	
R7. Interactivity. VESSEL should employ real interactivity in offering content. Interactive exercises should be used to help low-literate learners practice their worst-case-scenario fears, and to learn applicable skills and gain experience	**R7.1-C.** The coach should interact with users proactively by starting conversations and offering help, *following cognitive support rules, and according to a predefined timing scheme* **R7-2-C.** The coach should interact with learners reactively by answering questions and demands for help. *The coach should only recognize and react to a particular set of predefined participant utterances, based on keywords*	**R7.1-E.** The exercises should be interactive, requiring learners to use input mechanics to engage with the virtual environment in order to complete them

Table 2 (continued)

Exercise requirements	Coach requirements	General requirements
R8. Gaming principles. VESSEL should use elements and principles of interactive gaming. Gaming principles should be used carefully, as they can be seen as childish. If gaming principles are used in the software, they should focus on evoking pride and a sense of achievement	**8.1-C.** The coach should focus on praising the learner for success over emphasizing learner failures	

Unformatted text is the original description (cf. Schouten et al. 2020), and text in *italics* has been added

Requirement **R6. Support** is zoomed in to only coach-offered cognitive support. The coach should offer cognitive support according to the Generalized approach rules decision tree (Fig. 3) (**R6.1-C**).

Requirement **R7. Interactivity** is refined for only the coach. The coach can interact with learners either proactively or reactively. The coach's proactive interaction with the learner should be driven by the support rules decision tree (**R7.1-C**). And the coach's reactive interaction with the learner should be based on Sect. 2.4's speech recognition rules (**R7.2-C**).

3.3 Use Case: Formalized Cognitive Support for Online Banking

One use case is provided here: the coach giving formalized cognitive support to a learner doing an 'online banking' exercise about transferring money to a different account. Use cases consist of: *Preconditions* (conditions that are assumed true at the start of the use case), an *action sequence* (the steps taken by the user and the system over the course of the use case), and *post-conditions* (measurable desired outcomes that result from following the action sequence, i.e. the claims associated with the VESSEL requirements baseline). Two actors are used: '*Coach*' refers to the ECA coach providing formalized cognitive learning support, and '*user*' refers to the low-literate learner engaging with VESSEL. Particular action sequence steps reference Table 2's requirements to indicate that this step meets the requirement. Six claims are incorporated: cognitive/affective/social learning experience and cognitive/affective/social learning outcomes. Accessibility claims are not used because the user is presumed to already be working with VESSEL.

Preconditions:

1. The user is interacting with the coach-supported VESSEL system.
2. An online banking exercise has been selected.
3. The coach and the online banking Web site are both visible to the user.

Action sequence

1. The coach introduces the goal and the scope of the exercise to the user. (R.1.1-E, R2.1-C, R3.1-E, R5.1-C)
2. The user uses mouse and keyboard to interact with the online banking Web site and a microphone to talk to the coach. (R7.2-C, R7.1-E)
3. Since the coach is using the Individualized approach, it checks the user model for this particular user. Since the user has been successful at previous exercises, the coach sets this user's support delay to 25 s. This value will be used throughout the exercise. If the coach had not been using the Individualized approach, it would have set a support delay of 20 s without looking at the user model. (R1.1-C)
4. The user tries to navigate to the correct page on the online banking Web site, but takes a long time doing so. After 25 s of the user not making any progress, the coach offers the first level of cognitive support: a prompt. (R6.1-C, R7.1-C)

5. The user still cannot find the right page to navigate to. After another 25 s, the coach escalates the level of support to level 2: explanation. (R6.1-C, R7.1-C)
6. The user reaches the right page and starts filling out information. The user encounters a term they do not understand and ask the coach about it. The coach finds this keyword in its dictionary and offers explanation-level support about this keyword immediately. (R6.1-C, R7.2-C)
7. The user fills out some data incorrectly then tries to move on. The coach notices this and offers corrective feedback. (R2.1-C, R6.1-C, R7.2-C)
8. The user corrects the mistake and completes the exercise. The coach informs the user that the exercise is over. The coach updates the user model with the results from this exercise. Because the user has performed well, the coach increases the support delay to 30 s. In the following exercise, this delay will be used. (R1.1-C)

Post-conditions

1. The user has actively performed the exercise: the user has done at least one exercise step without the coach modeling the correct solution.
2. The user had a positive experience while doing the online banking exercise: the user's mood has either stayed at the same level of valence or has increased.
3. The user has interacted with the coach: the user has either asked the coach a question or answered one of the coach's questions.
4. The user has learned about the online banking steps and can recall this information later.
5. The user's self-efficacy with regard to online banking has increased.
6. The user considers the coach to be friendly and helpful.

4 Evaluation: Prototype Development

Functionality. The prototype consists of the three online banking exercises described in Sect. 2.1, and an ECA coach that offers cognitive learning support according to the Generalized and Individualized approaches described in Sects. 2.2 and 2.3. For the purpose of evaluation, the coach is designed to be controlled via the Wizard-of-Oz method (Maulsby et al. 1993).

Interaction methods. Learners interact with the online banking Web sites using mouse and keyboard. Learners can talk to the coach in natural language. The Wizard operator uses Fig. 3 decision tree to select what utterance the coach says at what moment, choosing prerecorded spoken utterances from a list. In the case of unexpected user actions or utterances, the Wizard can also use the set of general reaction utterances to get the exercise back on track without interruption.

Appearance The visual appearance of the ECA coach used in Schouten et al. (2020) is reused here. See Fig. 5. The coach ECA has one facial animation (opening and closing its mouth while sound is playing, to visually convey that it is 'speaking'), and no gestures or body language.

Fig. 5 VESSEL coach ECA (top right) and summary instructions (in Dutch) for online banking exercise 3

5 Evaluation: Methods

5.1 Experimental Design

An experiment was carried out to evaluate the learning effectiveness impact of our formalized-coach VESSEL prototype, as well as to compare the relative effectiveness of the Generalized and Individualized approaches. We therefore used the six learning effectiveness claims that were presented as use-case post-conditions: cognitive, affective, and social learning experience, and cognitive, affective, and social learning outcomes. Six high-level hypotheses were drafted corresponding to these six claims. Each hypothesis was then zoomed in on two predictions: one prediction about the overall system impact, and one prediction comparing the Generalized and Individualized approaches.

Learning Experience

- **H1 Cognitive Experience (Performance)**
  - **H1a** The learner takes active part in the exercise: The amount of instruction/modeling support needed to complete exercises is less than 100% of the possible maximum.
  - **H1b** Learners who receive support along the Generalized and Individualized approaches require less coach support to complete exercises than learners who receive only Generalized-approach support and expend less subjective mental effort.
- **H2 Affective Experience (Positive Affect)**
  - **H2a** The learner's affective state does not get more negative after completing an exercise with formalized coach support.

  - **H2b** The affective state of learners who receive Generalized and Individualized support changes more positively than learners who receive only Generalized support.

- **H3 Social Experience (Engagement)**
  - **H3a** The number of learner–coach interactions (*defined in Sect.* 5.4) is more than 0 during an exercise with formalized coach support.
  - **H3b** Learners who receive Generalized and Individualized support interact with the coach less often than learners who receive only Generalized support.

Learning Outcomes

- **H4 Cognitive Outcomes (Success)**
  - **H4a** The learner scores more than 0 points on the recall test after completing three exercises with formalized coach support.
  - **H4b** Learners who receive Generalized and Individualized support take less time to complete any exercise and score higher on the recall test after completing all three exercises, than learners who receive only Generalized support.
- **H5 Affective outcomes (self-efficacy)**
  - **H5a** The learner's self-efficacy about online banking increases after completing an exercise with formalized coach support.
  - **H5b** The self-efficacy increase of learners who receive Generalized and Individualized support is higher than learners who receive only Generalized support.
- **H6 Social Outcomes (Retention)**
  - **H6a** The learner judges the formalized coach as being helpful and friendly.
  - **H6b** Learners who receive Generalized and Individualized support judge the coach as more helpful and friendlier than learners who receive only Generalized support.

To test these hypotheses, a mixed-method repeated-measured experiment was designed, combining within-subjects and between-subjects measurements. The study's main independent variable was **Support Model**, with two levels: *Generalized Model* and *Individualized Model*. Participants were invited to complete the three online banking exercises in three experimental sessions, each one week apart: Participants did Exercise 1 in the first week, Exercise 2 in the second week, and Exercise 3 in the third week. Participants were randomly assigned one of two conditions at the start of the first week: 50% of participants worked in the Generalized Model condition throughout the entire experiment, wherein only the Generalized approach was used to provide support, and 50% of participants worked in the Individualized Model condition throughout the entire experiment, which used both Generalized and Individualized approaches.

5.2 Measures

Nineteen quantitative dependent variables were measured. Fifteen were self-report questions, measured using three questionnaires (Sect. 5.4), and four were objective performance metrics. Table 3 shows an overview of the variables.

5.3 Participants

Participants for the study were selected using Kurvers et al. (2013)'s language learner profiles, which subdivide first-language learners (L1) and second-language learners (L2) into five categories. Only learners that matched profiles 2 (fairly skilled L1 and L2 learners), 3 (L2 learners of average skill), and 4 (L1 learners of low skill) were invited to participate, as learners in profiles 1 (highly skilled L1

Table 3 Overview of measures

Variable	Description
Subjective measures: societal participation questionnaire (SPQ)	
SPQ.1. Self-efficacy (formal information skill)	"I can take out insurance"
SPQ.2. Self-efficacy (formal communication skill)	"I can ask for help at a service desk"
SPQ.3. Self-efficacy (informal information skill)	"I can read a map"
SPQ.4. Self-efficacy (informal communication skill)	"I can talk to my neighbors"
Subjective measures: self-assessment questionnaire (SAQ)	
SAQ.1. Self-efficacy (reading Dutch)	"I can read Dutch"
SAQ.2. Self-efficacy (online banking)	"I can do online banking"
SAQ.3. Self-efficacy (computer use)	"I can use a computer"
SAQ.4. Affect (valence)	"How good do you feel right now?"
SAQ.5. Affect (arousal)	"How active do you feel right now?"
SAQ.6. Affect (dominance)	"How strong do you feel right now?"
Subjective measures: exercise results questionnaire (ERQ)	
ERQ.1. Subjective mental effort	"How much effort did it take you to complete the exercise?"
ERQ.2. Coach affect (valence)	"The coach was happy"
ERQ.3. Coach affect (arousal)	"The coach was busy"
ERQ.4. Coach-affect (dominance)	"The coach took charge"
ERQ.5. Coach-affect (usefulness)	"The coach helped with the exercise"
Objective measures: direct measurement per exercise	
DM1. Completion time (s)	Time from start of exercise to completion
DM2. Level of coach support	Highest level of coach support needed to pass any waypoint
DM3. Learner–coach interaction	Amount of learner–coach interaction during the exercise
DM4. Recall test score	Score on end-of-experiment recall test

Includes measure source (societal participation questionnaire, self-assessment questionnaire, exercise results questionnaire, or direct measurement) and description

and L2 learners) and 5 (L1 and L2 learners with serious learning difficulties) are, respectively, too skilled to benefit from our level of support, and too low-skilled to engage with the prototype at all. Because the same selection procedure was used in our previous work (Schouten et al. 2020), we also assumed that these participants would have similar information and communication skill levels. Practically, this means we assumed that participant formal information skill levels (information skills in social settings characterized by rigid impersonal rules, such as online banking, cf. Schouten et al. 2016) were lower than their formal communication skill and informal information/communication skill levels (related to social settings characterized by flexible personalized rules). Participants were recruited from reading and writing classes throughout the Netherlands. Twenty-eight low-literate participants completed the entire experiment: Twenty-one men and seven women, with ages ranging from 24 to 73 ($M = 52.1$, $SD = 12.3$). Nineteen of the participants identified as natively fluent in Dutch; the other nine identified as 'somewhat fluent.' Other languages spoken by the participants (either natively or as a second language) included Arabic, Aramaic, Bosnian, Edo, English, French, Hindustani, Italian, Papiamentu, Russian, Somali, Spanish, and Turkish. Eight participants reported prior experience with online banking; of those, seven participants considered online banking easy to do. The 20 participants without online banking experience all found online banking hard.

5.4 Materials

The experimental setup consisted of two laptops, each connected to one external monitor (Fig. 6), which were used by the experimenters to run the experiment. The external monitors were used by the participants to see and interact with the exercises. The left laptop and monitor were used for the online banking exercises, and

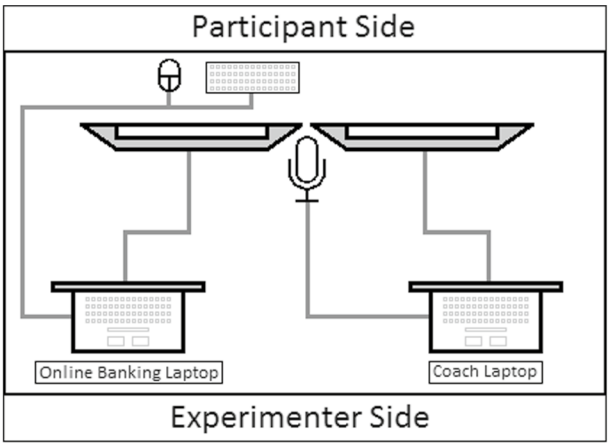

Fig. 6 Schematic overview of experimental setup. Two monitors (upper figures) are connected to two laptops (lower figures). Keyboard and mouse on participant side are connected to Online Banking Laptop; microphone placed between monitors is connected to coach Laptop

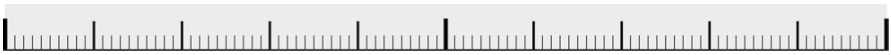

Fig. 7 Visual analogue scale used to measure self-efficacy, subjective mental effort, and coach affect

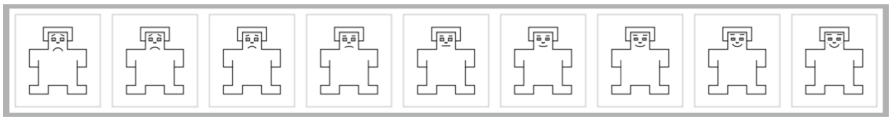

Fig. 8 Self-Assessment Manikin used to measure participant pleasure/valence

the right laptop and monitor were used for the coach. On the participant side, a mouse, keyboard, and microphone were provided as well; the microphone was used to 'explain' how participants were able to talk to the coach, as well as to record audio of the proceedings (with consent).

Four questionnaires were used. Three questionnaires measured the 15 self-report variables (see Table 3). First, the 'societal participation questionnaire' (SPQ) measured participant self-efficacy about four example crucial practical situations: taking out insurance (a representative example of an information skill used in a formal social context, cf. Schouten et al. 2017a), talking at a service desk (communication skill in a formal context), reading a map (information skill in an informal context), and talking to neighbors (communication skill in an informal context). Second, the 'self-assessment questionnaire' (SAQ) measured participant self-efficacy regarding the exercise, and participant affective state. Third, the 'exercise results questionnaire' (ERQ) measured subjective mental effort, and participant affect towards the coach. Two answer methods were used: a visual analogue scale (Fig. 7), and the Self-Assessment Manikin (SAM) Fig. 8. Answers to self-efficacy, mental effort, and coach affect questions were given using the visual analogue scale, as this method does not require reading and writing skills and allows participants to rate concepts that are otherwise hard to describe or categorize (Huskisson 1983). Answers to self-affect questions were given using the SAM, which measures three affective dimensions: pleasure/valence, arousal, and dominance (Bradley and Lang 1994). Questions were always read aloud to participants, who would then mark their answer on the matching bar or figure. The fourth 'demographic' questionnaire measured participant age, sex, schooling history, time spent in the Netherlands, languages known, and prior experience with online banking. These questions were read out loud as well; the researchers wrote down the answers.

In addition to the questionnaires, four objective measures were taken. First, participant completion time was measured with a stopwatch. Second, exercise support level was calculated by tabulating the number of times each coach utterance type (Table 1) was used in an exercise and dividing the sum of the resulting support levels (1 for prompts, 2 for explanations, etc.) by the number of critical waypoints. Third, learner–coach interaction was recorded with the microphone. Lastly, a 'recall test' was created to measure participants' learning success. The test consisted of six

A4-printed screenshots of the online banking Web site. For each of the six pictures, participants were given 60 s to answer one question, referencing an activity from one of three exercises. Answers were scored as either fully correct (1 point), partially correct (.5 points), or incorrect/out of time (0 points).

5.5 Procedure

The three experimental sessions were held over the course of three weeks, each one week apart. Two researchers were present: one researcher acted as the dedicated Wizard-of-Oz controller for the coach, while the other managed all participant interaction and controlled the online banking task environment. The first session started with general introduction, informed consent forms, and the demographic questionnaire. The first SPQ was administered, followed by the first SAQ. The managing researcher explained the general experiment flow and activated the coach, which was controlled by the second experimenter. The coach introduced itself to the user, explained the first exercise, and showed the instruction material. Participants were told to complete the first exercise with the help of the coach. No time limit was set. As soon as participants were finished, researchers administered an ERQ and a second SAQ. Participants were then debriefed, ending the first session. In between the first and second sessions, all participants' performances were rated, using the 'good/medium/bad' categorization described in Sect. 2.3. For participants in the Individualized condition, the user model was updated and support delays were changed where necessary (as shown in Fig. 4).

In the second session, researchers started by administering an SAQ. After that, flow proceeded as per the first session, with participants completing the second exercise before filling out an ERQ and an SAQ. In between the second and third sessions, participant performances were again rated, and support delays were again updated for participants in the Individualized condition. The third session (with the third exercise) was similar to the previous two, except for additions at the end: after the final exercise results and SAQ, researchers administered a second SPQ. After this, the recall test was explained and administered. Finally, participants were fully debriefed (including a 'look behind the scenes' for the Wizard-of-Oz method, and a short qualitative interview to see how they experienced working with the prototype and the coach) and rewarded for participation.

6 Evaluation: Results

Three analysis steps were done. First, the data were characterized and starting assumptions were checked, by looking at participant descriptives, exercise difficulty levels, and the effectiveness of the different support levels. Second, quantitative analyses are conducted on the Table 3 measures in order to verify the hypotheses. And third, two post hoc analyses were carried out: the predictive value of several variables on recall test score was tested, and groups of participants were evaluated based on initial performance. Finally, qualitative observations were made by

the researchers, during the experiment and by listening to the audio recordings afterwards.

Before analysis, data validity was checked in four ways, following Nimon (2012)'s outline of statistical assumptions in General Linear Model (GLM) analyses. First, P–P and Q–Q plots were used to assess multivariate normality. Results showed that multivariate normality was upheld for all measures except three: measures SPQ.3 and SPQ.4 show mild and medium abnormality, respectively. And while measure DM2 shows a good normal distribution, dividing this variable into DM2a and DM2b (see also Table 5) shows that while DM2a is normally distributed, DM2b is mildly abnormal. Second, Mauchly's test of sphericity was used to assess data variance. Results showed that the assumption of equal pair variance was upheld for all measures except measure SAQ.1. Third, questionnaire reliability was assessed. Cronbach's α was .730 for the SPQ, .872 for the SAQ, and .734 for the ERQ. No data reduction measures were used. Fourth, the dataset was checked for overall correctness. Logging issues were discovered in the support level data for three participants; these participants were excluded from further support level analyses (pertaining to DM.2 and DM.3), but otherwise included. Given these results, we were confident to proceed with the planned analyses.

6.1 Assumptions

Four assumptions were checked: the assumption of participant starting skill, the assumption of equal exercise difficulty, the assumption of support model effectiveness, and the assumption of temporal contingency. The *assumption of participant starting skill* was that the formal information skill level for low-literate participants would be low when compared to their formal communication skill and informal information/communication skills. The *assumption of equal exercise difficulty* was that all three exercises would require similar amounts of time and support to complete. The *assumption of support model effectiveness* was that, from prompt to modeling, the five utterances in the support model would be more effective at helping learners complete exercise steps. The *assumption of temporal contingency* was that a coach with a lower support delay (with 10 s being the lowest possible delay and 30 s the highest) would result in a higher average support level and a lower average exercise completion time.

To check the assumption of participant starting skill, SPQ means were compared with a paired-samples T test (Table 4). Analysis shows that before the start of the experiment, participants rated their formal information skill (SPQ.1) as significantly lower than their formal communication skill (SPQ.2, $t(27) = -4.313$, $p = .000$), informal information skill (SPQ.3, $t(27) = -2.657$, $p = .013$), and informal communication skill (SPQ.4, $t(27) = -5.413$, $p = .000$). Informal information skill was also rated as lower than informal communication skill ($t(27) = -3.049$, $p = .005$). After experiment, the exactly same pattern was seen (respectively ($t(27) = -5.396$, $p = .000$), ($t(27) = -2.918$, $p = .007$), ($t(27) = -5.670$, $p = .000$), and ($t(27) = -3.228$, $p = .003$)). As such, this assumption was upheld.

Table 4 Societal participation questionnaire means and standard deviations

	Pre-experiment	Post-experiment
SPQ.1. 'I can take out insurance.'	49.86 (SD = 36.98)	44.82 (SD = 32.18)
SPQ.2. 'I can get help at a service desk.' (*formal communication skill*)	80.61 (SD = 23.39)	78.00 (SD = 22.15)
SPQ.3. 'I can read a map.' (*informal information skill*)	69.43 (SD = 33.50)	64.39 (SD = 32.38)
SPQ.4. 'I can talk to my neighbors.' (*informal communication skill*)	86.86 (SD = 21.79)	81.50 (SD = 25.47)

To check the assumption of equal exercise difficulty, a repeated-measures GLM analysis compared exercise completion time and average support level for the full exercise, as well as support level for only the navigation steps and support level for only the data entry steps. Table 5 shows the results of the analysis. Significant differences were found: the second exercise required a lower overall support level to be completed, the third exercise required a lower navigation support level, and all three exercises required different amounts of data entry support. As such, the assumption of equal difficulty was not upheld. In light of these findings, we chose not to alter our a priori planned hypotheses evaluations, but to incorporate these findings into a post hoc analysis (Sect. 6.3).

To check the assumption of support model effectiveness, we tabulated the total number of support utterances given for each level. We also counted how many utterances in each level successfully helped a participant get to the next critical waypoint; i.e. if the instruction 'click on the word Online Banking' got a participant to navigate to the online banking page, then that utterance was successful. Table 6 shows the number of utterances for each category, as well as the success rate. The numbers show that in the order of prompt, explanation, hint, instruction, and modeling, the success rate of each utterance goes up. As such, this assumption was upheld.

Finally, to check the assumption of temporal contingency, one-way ANOVA analyses were done on the average support level and average completion times of exercises 2 and 3, using coach support delay for that exercise as an input. Exercise 1 was not used, as all participants had a support delay of 20 s in that exercise. Table 7 shows that as the coach's support delay went down, the average support level increased (exercise 2: $F(2, 23) = 5.755$, $p = .010$; exercise 3: $F(3, 22) = 4.555$, $p = .013$), but average completion time did not decrease as expected. We chose to continue with our envisioned hypothesis evaluations, and to keep these findings on hand when interpreting the results of any analysis that leans on the assumption of temporal contingency.

6.2 Hypotheses Evaluation

To evaluate hypotheses H1 through H6, the data from the SAQ, ERQ, and the direct measurements (see Table 3) were systematically analyzed. Table 8 shows a schematic overview of all data measurements, ordered per hypothesis. Included in the table are means and standard deviations per measurement moment (before/after

Table 5 Exercise descriptives

	1st Exercise	2nd Exercise	3rd Exercise	Test statistic
DM1. Average completion time (in s)	691 (SD = 302)	568 (SD = 232)	704 (SD = 315)	
DM2. Average support level (all waypoints)	2.02 (SD = 1.06)	1.58 (SD = 1.03)	2.03 (SD = 1.25)	$F(2, 23) = 5.183$ $p = .014$ $\beta = .774$
DM2a. Average support level (navigation waypoints)	2.74 (SD = 1.04)	2.71 (SD = 1.30)	1.79 (SD = 1.37)	$F(2, 23) = 9.117$ $p = .001$ $\beta = .956$
DM2b. Average support level (data entry waypoints)	1.29 (SD = 1.33)	0.44 (SD = 1.07)	2.27 (SD = 1.36)	$F(2, 23) = 26.245$ $p = .000$ $\beta = 1.000$

Completion time is measured in seconds. 'Average support level' means: the average highest level of support needed to pass critical waypoints. F value (F), significance (p), and observed power (β) are given if $p < .05$

Table 6 Number of utterances given for each support type, and success rate for each, over the entire experiment

	Prompt	Explanation	Hint	Instruction	Modeling
Number given	329	290	253	166	85
Number successful	38	38	87	81	85
Success rate	11.6%	13.1%	34.4%	48.8%	100%

1124 support utterances were recorded in total

Table 7 Average support level and completion times for exercises 2 and 3, per coach support delay category

	10s	15s	20s	25s	30s
Exercise 2					
Number	X	4	14	7	X
Average support	X	2.65	1.65	.82	X
Level*		(SD = 1.24)	(SD = .91)	(SD = .47)	
Average	X	654	566	396	X
Completion time (s)		(SD = 215)	(SD = 241)	(SD = 204)	
Exercise 3					
Number	2	0	13	4	6
Average support	3.81	–	2.24	2.16	.90
Level*	(SD = .97)		(SD = 1.19)	(SD = .53)	(SD = .86)
Average	961	–	732	838	490
Completion time (s)	(SD = 238)		(SD = 295)	(SD = 235)	(SD = 217)

Rows marked with * show significant ANOVA differences at $p<.05$. Columns marked 'X' are not relevant: in exercise 2, 10s and 30 timings were impossible to reach by design

exercise 1/2/3), which statistical test was used to analyze the measure, and the relevant test statistic, if significant. Three types of tests were used: repeated-measures General Linear Model analysis, one-sided *T* tests, and one-way ANOVAs. All GLM analyses were done using all data points as one factor (meaning they all contained either one factor with three levels, or one factor with six levels); additionally, all GLM analyses were conducted either without any between-subjects factors (for hypotheses H1a to H6a) or using participant experimental condition as a between-subjects factor (for hypotheses H1b to H6b). One-way ANOVA analyses were conducted on participant experimental condition. One-sided T tests were conducted on select values, as shown in Table 8. Note that Table 8 only shows hypothesis evaluations for H1a to H6a; evaluation of hypotheses H1b to H6b showed no significant results and as such was not included in the table.

The following results were found. For all exercises, the average support level was lower than 4, indicating no exercise required instruction and/or modeling support for every critical waypoint. This supports H1a. For measures SAQ.4, SAQ.5, and SAQ.6, repeated-measures GLM shows no significant differences across exercises.

Table 8 Hypothesis evaluation table

	Mean (standard deviation)						Hypothesis Hxa	
	B1	A1	B2	A2	B3	A3	Test	Outcome
H1: cognitive experience (performance)								
ERQ.1		62.04 (24.17)		62.93 (19.47)		56.67 (27.44)		
DM.2		2.02(1.06)		1.58 (1.03)		2.03 (1.25)	1s-T (4)	$A1:t(25)=-9.329\ p=.000\ A2:t(25)=-11.792\ p=.000\ A3:t(25)=-7.899\ p=.000$
H2: affective experience (positive affect)								
SAQ.4	6.93 (2.05)	6.64 (2.31)	6.96 (2.00)	6.75 (1.78)	7.04 (1.75)	6.25 (2.22)	Rep. GLM	
SAQ.5	4.54 (2.44)	4.43 (2.56)	4.79 (2.25)	5.18 (2.13)	5.11 (2.38)	5.14 (2.31)	Rep. GLM	
SAQ.6	6.36 (1.78)	6.71 (1.82)	6.82 (1.70)	6.29 (1.72)	6.39 (1.75)	6.21 (2.10)	Rep. GLM	
H3: social experience (engagement)								
DM.3		4.36 (2.00)		3.20 (1.87)		4.04 (2.23)	1s-T(0)	$A1:t(25)=10.914\ p=.000\ A2:t(25)=8.552\ p=.000\ A3:t(25)=9.073\ p=.000$
H4: cognitive outcomes (success)								
DM.1		691 (302)		568 (232)		704 (315)		
DM.4						3.25(1.46)	1s-T (0)	$t(27)=11.810p=.000$
H5: affective outcomes (self-efficacy)								
SAQ.1	78.11 (18.15)	79.75 (18.18)	80.21 (21.51)	79.46 (21.48)	77.75 (19.96)	76.50 (20.32)	Rep. GLM	
SAQ.2	31.43 (35.50)	45.57 (33.08)	39.57 (31.11)	42.21(30.70)	48.39(32.25)	46.07(32.15)	Rep. GLM	$F(1,27)=4.591p=.041$
SAQ.3	54.68 (27.76)	60.68 (27.81)	58.07 (28.58)	58.57 (26.81)	62.36 (26.83)	60.50 (27.93)	Rep. GLM	
H6: social outcomes (coach opinion)								
ERQ.2		73.03 (22.67)		73.65 (21.26)		75.21 (21.05)	1s-T(50)	$A1:t(27)=5.746\ p=.000\ A2:t(27)=6.192\ p=.000\ A3:t(27)=6.337\ p=.000$
ERQ.3		30.00 (23.41)		37.65 (28.92)		32.46 (27.96)	1s-T(50)	$A1:t(27)=-4.128\ p=.000\ A2:t(27)=-2.379\ p=.024\ A3:t(27)=-3.318\ p=.003$
ERQ.4		65.88 (28.37)		66.19 (28.78)		67.04 (30.82)	1s-T (50)	$A1:t(27)=3.165\ p=.003\ A2:t(27)=3.134\ p=.004\ A3:t(27)=2.925\ p=.007$

Table 8 (continued)

	Mean (standard deviation)						Hypothesis Hxa	
	B1	A1	B2	A2	B3	A3	Test	Outcome
ERQ.5		79.53(18.97)		80.90(23.63)		81.21(21.92)	1s-T(50)	$A1: t(27) = 8.805\ p = .000\ A2: t(27) = 7.283\ p = .000\ A3: t(27) = 7.534\ p = .000$

Legend: B1/A1 means 'Before exercise 1'/'After exercise 1,' etc. Column 'Hypothesis Hxa' contains test statistics that evaluate hypotheses H1a to H6a. 'Rep. GLM' means repeated-measures GLM. '1s-T (X)' means one-sample T test, testing against H0 = X. 'One-way ANOVA' means one-way ANOVA on experimental condition. Test results are given for $p < .05$. Gray boxes mean no value was measured or no test was conducted, and blank boxes mean no significant result was found

This indicates participant affective state did not get significantly more negative as a result of working with the coach, supporting H2a. On average, participants interacted with the coach more than 0 times in each exercise, supporting H3a. On average, all participants scored higher than 0 on the recall test, supporting H4a. A closer look at the data shows that no single participant scored 0 on the test. Measure SAQ.2 (self-efficacy—online banking) was significantly different across exercises. Follow-up analysis shows that value B1 ('before exercise 1') was significantly lower than the other five, indicating that online banking self-efficacy has increased after completing exercise 1. Figure 9 shows this result. This partially supports H5a, as self-efficacy does not increase after every exercise. Finally, one-sided T tests show that the averages of ERQ.2, ERQ.4, and ERQ.5 are significantly higher than the scale midpoint and that ERQ.3 is significantly lower. This suggests that participants judged the coach as affectively positive, calm, dominant, and helpful, weakly supporting H6a.

Finally, tests for between-subjects effects showed no significant results for age, sex, schooling history, time spent in the Netherlands, languages known, and prior experience with online banking.

6.3 Post Hoc Analyses

6.3.1 Recall Test Analysis

Regression analyses were carried out to test whether the following variables could predict recall test scores: average support level throughout all exercises, completion time per exercise, average completion time across all exercises, participant age, participant sex, participant experience with online banking, and number of weeks spent living in the Netherlands. Prior to this, a bivariate correlation analysis was carried out to see which variables should be included in a single regression test. This analysis showed that several variables were significantly correlated (at $p < .05$), limiting their applicability for regression analysis. The following variables were selected

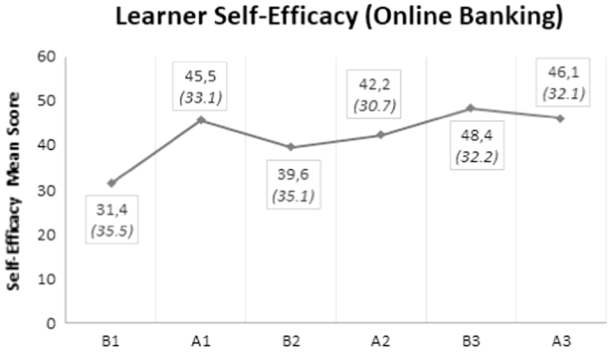

Fig. 9 Mean of 'online banking self-efficacy' for the six measurement moments. Boxes indicate mean (standard deviation). Horizontal axis shows the six measurement moments. Vertical axis shows score on SAQ.2, in range [0–100], measured using the visual analogue scale (Fig. 7). Columns 'B1' to 'A3' refer to measurement moments 'Before exercise 1' to 'After exercise 3'

for a stepwise linear regression for knowledge test results: average support level, participant sex, time spent in the Netherlands, and experience with online banking. One significant result was found: average support level negatively predicts knowledge test results ($t = -3.806$, $p = .001$). A curve estimation analysis was done to confirm this. Linear, quadratic, and logarithmic models were tested. Both a linear model ($F(1, 23) = 14.483$, $p = .001$) and a logarithmic model ($F(1, 23) = 19.708$, $p = .001$) confirmed that a higher average support level corresponded to a lower recall test score. See Fig. 10.

6.3.2 Performance Group Analysis

One interpretation of the preceding hypothesis and recall test analyses is that participant online banking skill levels did not change significantly over the course of three exercises. In this case, 'participant online banking skill level' should be treated as a set trait instead of a dependent variable. If all three exercises were equal in challenge, exploratory techniques (e.g. cluster analysis) could reveal this. However, Table 5 shows that the exercises are not equal in terms of the level of support needed to complete them: Web site- and exercise-specific learning effects in the second and third exercises conflate the grouping. This strongly implies that some exercises were more challenging or difficult than others. Taking this into account, we clustered

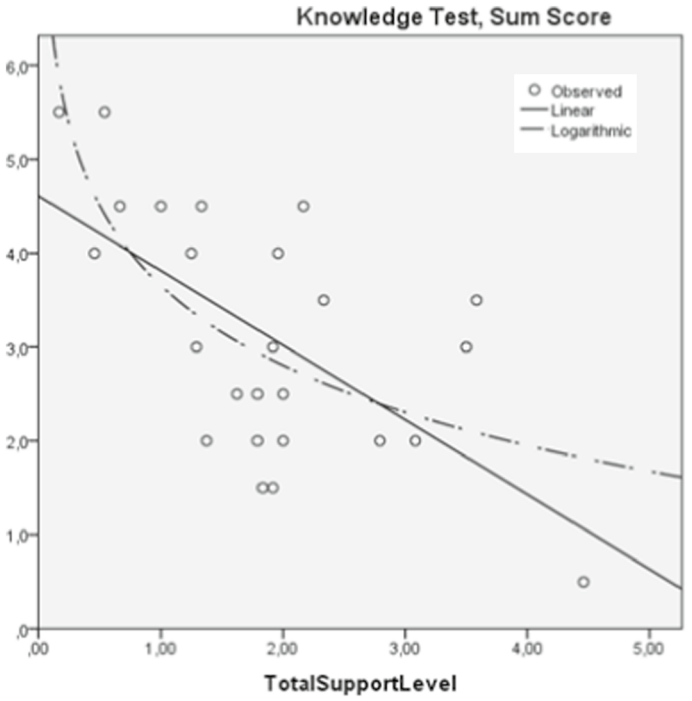

Fig. 10 Curve estimation result for recall test score as a function of average support level throughout exercise

participants into three 'performance groups' based on their performance in the first exercise; we made the assumption here that their performance in this exercise was the most accurate reflection of their 'actual' online banking skill level, before any potential learning effects from the experiment and the effects of Individualized support came into play. Six people were assigned to the 'Bad' group, ten people to the 'Medium' group, and nine people to the 'Good' group, based on their performance in the first exercise, using the established user model categorization (Sect. 2.3). The repeated-measures GLM analyses in Sect. 6.2 were then run again with this variable as a between-subjects factor with three levels: *Bad*, *Medium*, and *Good*. Two effects were found: compared to Medium and Good, participants in the Bad group had significantly lower computer-use self-efficacy overall (main effect of between-subjects factor, $F(2, 23) = 5.402$, $p = .012$, Fig. 11), and (on average) dropped in positive affect after completing any exercise (interaction effect, $F(2, 23) = 3.525$, $p = .047$, Fig. 12). As a result of this last finding, hypothesis H2a is no longer fully supported, but partially supported: the affective state of participants in the Good and Medium groups did not get worse as a result of working with the coach, but the affective state of participants in the Bad group did.

6.4 Observations and Interviews

Experimenters observed that participants managed to work with the coach as intended. The provided support was sufficient for the exercises: almost all participants took active part in the exercises, even when these were obviously difficult, and managed to complete them fully. Only three times did participants 'give up' and wait for the coach to model every remaining step. While doing the exercises, participants listened to the coach's support and generally followed direct instructions if they understood them. Participants successfully interacted with the coach within the constraints of our speech recognition and support behavior rules. The experimenters felt that the 20-s support delays were very long and that for particular participants (e.g. participants who would switch their attention around very quickly) the support

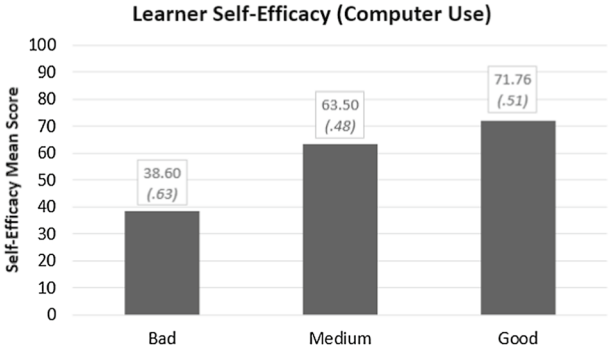

Fig. 11 Performance group analysis showing main between-subjects effect on SAQ.3 (computer-use self-efficacy). Boxes indicate mean (standard deviation). Horizontal axis shows the three performance groups. Vertical axis shows score on SAQ.3, in range [0–100], measured using the visual analogue scale (Fig. 7)

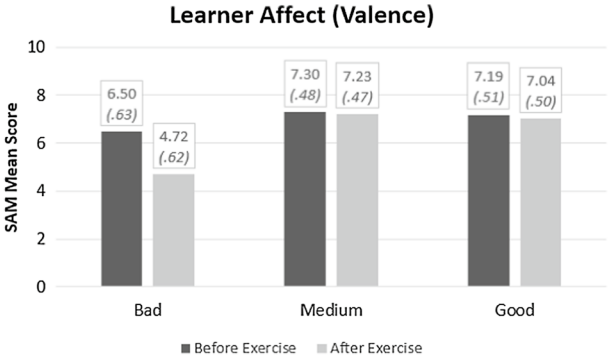

Fig. 12 Performance group analysis showing interaction effect on SAQ.4 (valence). Boxes indicate mean (standard deviation). Horizontal axis shows the three performance groups; two bars per group indicate measures taken before and after any exercise. Vertical axis shows score on SAM.1, in range [1–9], measured using the Self-Assessment Manikin (Fig. 8)

utterances did not arrive 'at the right time.' But on the participant side, this was not experienced. In interviews, participants simply accepted that the coach was slow sometimes and that 'she took some time to give a good answer.'

Different progress results were seen in the different support delay timing conditions. In the 20-s condition, most participants were able to complete all critical waypoints without requiring instruction or modeling support. And support was often given while participants were actively engaged with the exercise. Similar patterns were seen in the 15- and 25-s conditions (but this is limited by the low number of observations in these conditions). Different patterns were seen in the 10- and 30-s conditions. In the 10-s condition, participants received support so quickly that they often had no time to process it before another utterance was due. Many more instances of instruction and modeling support were seen here than in other conditions (also shown in Table 7). In the 30-s condition, although many participants in this condition hardly needed help, it was observed that when participants did need help to proceed, they had to sit through long and noticeable waiting times. More so than in other conditions, participants seemed annoyed that the coach would not immediately answer their questions.

Two additional observations stand out. First, while participants did often interact with the coach, experimenters felt as though the total amount of human–coach interaction in this study was lower than in the previous one (Schouten et al. 2020). Participants that spoke to the coach talked as they would to a human conversation partner, using complete sentences and sometimes even gesturing at the screen. But not all participants spoke to the coach often, or at all. Particularly, while participants often reacted to coach questions and prompting, very little proactive interaction was seen. In Schouten et al. (2020), participants very often talked to the coach extensively and in great detail, including asking it highly complex questions and even telling it stories about their own lives. This rarely happened in our current study (although it did happen, with one participant even joking he'd '*like to take [the coach] on a date sometimes*'). And second, while all participants completed

all exercises with the coach's help, ending interviews revealed that many reflected on this negatively. Participants did not see the situation as them working together with the coach for the goal of learning, but as them failing to complete a challenge and the coach needing to 'rescue' them. One participant, who completed exercise 3 in good time but needed modeling to find the very last navigation waypoint, complained that ... *I wouldn't have been able to do it without the coach*".

7 Conclusions and Discussion

7.1 Conclusions

This study intended to answer two research questions. Question Q1 was: '*How can we create a design specification for VESSEL that incorporates rules for cognitive learning support provided by an ECA coach?*' Sub-question Q1a, '*Which operational demands, human factors knowledge, and technologies are needed to write these rules?*,' was answered in Sects. 2 and 3. In Sect. 2, we showed how hard online banking exercises provide a task environment for cognitive learning support, how the scaffolding concepts of contingency, fading, and transfer of responsibility inform the coach's cognitive support behavior, how user modeling can be employed to adapt offered support to individual performance and circumstances, and how we envision the role of speech recognition. By incorporating this into the foundation, we resolved our knowledge gaps. In Sect. 3, we created dialogue rules to specify the ECA coach's cognitive support behavior, refined the requirements baseline to incorporate these rules, and wrote a new use case to illustrate the envisioned user–system interaction. Sub-question Q1b, '*Which functionalities, interaction methods, and appearances should the ECA coach have to reflect this specification?*,' was answered in Sects. 3 and 4: a new VESSEL prototype was created on the basis of our specification, including formalized cognitive learning support rules and user modeling functionality.

Question Q2 was: *"What is the learning effectiveness impact of a VESSEL prototype that offers cognitive learning support according to the formal specification?*" Sub-questions Q2a, '*Are the learning effectiveness results of this prototype comparable to the VESSEL prototype that offered informal cognitive, affective, and social learning support?*,' and Q2b, '*Does using both the Generalized and Individualized approaches to learning support result in higher learning effectiveness than using only the Generalized approach?*,' were answered by experimentally evaluating the prototype in Sects. 5 and 6. We tested six hypotheses for each sub-question (12 in total). For question Q2a, hypotheses H1a, H3a, and H4a were fully supported, and H2a, H5a and H6a were partially supported, showing that the ECA coach's formalized cognitive support resulted in high learning effectiveness for low-literate learners. Cognitively, learners used the coach for guidance, but did not rely on it for everything. Affectively, the coach had no negative influence on the user's mood for users in the Good- and Medium-performance groups. Self-efficacy regarding online banking increased after doing the first exercise and stayed at the new high level afterwards. And socially, learners interacted with the coach as if it was human, and

judged 'her' as a friendly, useful helper. These results suggest that the formalized coach meets our design goals. Learners can use the coach to successfully complete challenging exercises, resulting in nonzero recall and a significant increase in self-efficacy. The lowered affective state of users in the Bad-performance group is unexpected, however, and this should be further investigated in future work.

Comparing these results to Schouten et al. (2020) shows interesting similarities. Both studies show a significant increase in self-efficacy after completing one coach-supported hard online banking exercise. In both cases, the actual reported values for self-efficacy are just below or around the scale midpoint (0 in Schouten et al. (2020), 50 in this study), suggesting that while online banking self-efficacy does *increase*, it is still not very *high*. Other value similarities include positive affect when the coach was present (halfway between scale midpoint and maximum value), difficulty and required effort of the exercise (idem), and the degree to which the coach was seen as a supportive agent (close to scale maximum). These similarities suggest that the learning effectiveness results of this prototype and the Schouten et al. (2020) prototype are comparable, answering question Q2a. Importantly, these results seem to indicate that moving the coach to keyword-based speech recognition was not a significant problem for low-literates. Experimenter observations corroborate that low-literate participants had little problems using the coach. Almost all participants asked their questions slowly and clearly, using the exact terms from the Web site even without being instructed to do so. When problems did occur, it was often because participants used unanticipated question phrasings and keywords, or because they assumed too much real human conversation ability on the coach's behalf: for instance, certain participants expected the coach to be able to use past conversation context, attempting to reference things that happened earlier in the experiment or even in earlier experimental sessions. Further development, including expanded keyword lists and more dialogue rules, could alleviate these problems. Experimenters did feel that there was less learner-started social interaction in this study than in the Schouten et al.'s (2020) study, which both experimenters were also part of. We suspect that this happened because the previous study's coach used small talk for social support at the start of exercises. By asking the learner questions about their life and talking about 'her own experiences,' this coach afforded being spoken to like a human partner, acclimatizing low-literate learners to the idea they could actively ask questions. Since the current coach did not do this, participants may not have considered to try. Learners still reactively answered the coach's questions, but would only sometimes proactively ask questions. Future work should study whether small talk influences learner–coach interactions in this way.

For question Q2b, hypotheses H1b through H6b were all not supported. This shows that including the Individualized support model did not significantly improve on the Generalized model. Observed qualitative differences were not reflected in quantitative measures. Two possible explanations can be offered. One (unlikely) option is that support delay does not have a significant influence on the learning experience of low-literates at all. We instead suspect that our manipulations did not actually match learning support to user skills, meaning we did not achieve fading the way we envisioned. Qualitative and quantitative data support this explanation: lower support delays caused information overload, while higher support delays caused

a need for waiting. Future work should verify this: perhaps smaller delay changes with less extreme end points, over a longer period of time, would result in fading as expected.

7.2 Limitations

Three limitations are identified in this study. First, the number of participants recruited for this study is relatively low for the purposes of quantitative statistics. This problem is difficult to avoid when doing experimental research with low-literates, as the available pool of potential participants is low: finding and recruiting low-literates is a non-trivial issue (cf. Schouten et al. 2017a), and we further limited this pool by using Kurvers et al. (2013)'s language learner profiles as a selection criterion (see Sect. 5.3). This calls the power of our results into question. While analysis has shown that our data uphold multivariate normality and equal pair variance, and observed power was generally satisfactory, a larger sample size would solidify our findings. A standout point is the fact that eight of our 28 participants reported prior experience with online banking, which seems like a strong potential confound. As the online banking environment used in our work was created for this experiment (meaning no participants could have direct experience with it), and as between-subjects analysis showed no significant effect for 'prior online banking experience,' we are confident about the accuracy of our findings; nevertheless, future work should give this factor strong consideration.

Second, the experimenters ran into some implementation issues with the prototype. The Wizard operator could not correctly control the coach 100% of the time: technical difficulties in the coach's control program caused unavoidable time delays of up to 12 s between selecting a coach utterance and that utterance actually playing. This problem was first encountered during pilot testing, but was not resolved before the actual experiments took place. As a result, the Wizard operator had to train to factor them in. This formed a source of noise in the support provision. Additionally, in situations where the participants performed actions the coach was not built to expect, the Wizard had to append new rules on the fly. For instance, at the start of the first experiment, there was no rule for what to do if the participant returned to an earlier-completed waypoint. This situation was encountered during the first experiment, at which point a rule was created to handle it. Afterwards, this rule was incorporated into the coach and executed consistently. However, initial occurrences of situations like this still introduced noise.

Finally, the post hoc analysis of learner skill and performance reveals a problem with the assumptions underlying our work. Section 6.3.1 shows that learners who required more support to complete exercises scored more poorly on the recall test. One interpretation is that, in our three weeks of testing, learners' actual skill in doing online banking has not changed. Rather, learners with initial high skill levels needed little support and scored well, while learners with low skill levels needed much support and scored poorly. This is supported by the performance group analysis, which shows that learners who performed poorly in the first exercise consistently had a worse mood after exercises and judged their computer use self-efficacy

to be low (Sect. 6.3.2). It looks as if the coach has only helped learners *complete* the exercise, not *master* it. In the interviews (Sect. 6.4), participants blamed themselves for failing to succeed alone and viewed the coach's instructions as an admittance of that failure. Two assumptions underlying requirements R1.1-C (coach adaptability) and R1.1-E (exercise adaptability) are that learning support can lower the experienced difficulty of challenging exercises and that low-literate learners should not be allowed to fail. But results suggest that participants still experienced the exercise as very challenging; the coach's help was not seen as lower difficulty, but as unfair help. Even though all exercises were completed, learners attributed failure on the level of separate critical waypoints to themselves. Future study should investigate whether this happens consistently. If the coach cannot actually lower experienced difficulty, and if low-literate learners weigh failure on any waypoint level more heavily than success on the overall exercise, the assumptions underlying our coach's behavior must be rethought.

7.3 Future Work

This study has demonstrated the value of using formalized cognitive learning support for low-literates. Learners successfully interacted with the coach to complete challenging exercises, which resulted in a positive learning experience and higher online banking self-efficacy. These findings indicate that our current VESSEL development direction has merit. We will build on this in future work: now that our cognitive support has been formalized and evaluated, we can try to do the same for affective and social support. In a next SCE iteration, we turn our attention towards building a prototype that provides support not only contingent on the learner's cognitive needs, but also their affective and social needs (cf. Schouten et al. 2017b).

Acknowledgements Informed consent was obtained for all work described in this publication involving human participants. This work was supported by the 'Interaction For Universal Access' project as part of the Dutch national program COMMIT.

References

Bandura, A.: Self-Efficacy: The Exercise of Control. W.H. Freeman, New York (1997)
Bayles, M.: Online banking: why people are branching out. Transfer **35**, 90 (2004)
Benmarrakchi, F.E., El Kafi, J., Elhore, A.: User modeling approach for dyslexic students in virtual learning environments. Int. J. Cloud Appl. Comput. (IJCAC) **7**(2), 1–9 (2017)

Berger, P.L., Luckmann, T.: The Social Construction of Reality: A Treatise in the Sociology of Knowledge (1966). Double and Company, New York (1984)
Bhowmick, P.K., Sarkar, S., Basu, A.: Ontology based user modeling for personalized information access. IJCSA **7**(1), 1–22 (2010)
Bloom, B.S.: Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain. Longman, New York (1956)
Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994)
Buisman, M., Houtkoop, W.: Laaggeletterdheid in Kaart. Technical report, Expertisecentrum Beroepsonderwijs & Stichting Lezen en Schrijven (2014)
Cassell, J., Bickmore, T.W.: Negotiated collusion: modeling social language and its relationship effects in intelligent agents. User Model. User Adapt. Interact. **13**, 89–132 (2003). https://doi.org/10.1023/A:1024026532471
Ciloglugil, B., Inceoglu, M.M.: User modeling for adaptive e-learning systems. In: International Conference on Computational Science and Its Applications, pp. 550–561. Springer (2012)
de Greef, M., Segers, M., Nijhuis, J.: Feiten & Cijfers geletterdheid. Technical report, Stiching Lezen & Schrijven (2014) http://www.lezenenschrijven.nl/assets/uploads/publicaties/LS_FeitenCijfers_2.0_web_3.pdf. Accessed 29 Jan 2015
D'Mello, S., Graesser, A.C.: AutoTutor and affective autotutor: learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans. Interact. Intell. Syst. (TiiS) **2**(4), 23 (2012)
Fischer, G.: User modeling in human–computer interaction. User Model. User Adapt. Interact. **11**(1–2), 65–86 (2001)
Gibbs, G., Simpson, C., James, D., Fleming, S.: Learning and teaching in higher education. **1** (2004)
Graesser, A.C., D'Mello, S., Cade, W.: Instruction based on tutoring. Handbook of Research on Learning and Instruction, pp. 408–426 (2011)
Graesser, A.C., Person, N.K.: Question asking during tutoring. Am. Educ. Res. J. **31**(1), 104–137 (1994)
Groot, A., Coppens, K., Lam, J.F.: Motiveren van laaggeletterden: Een literatuurstudie naar succesvolle interventies. ECBO (2019)
Horvitz, E.J., Breese, J.S., Heckerman, D., Hovel, D., Rommelse, K.: The lumiere project: Bayesian user modeling for inferring the goals and needs of software users. (2013) arXiv preprint arXiv:13017385
Huskisson, E.C.: Visual analogue scales. Pain Measurement and Assessment, pp. 33–37 (1983)
Johnson, G.M.: Instructionism and Constructivism: Reconciling Two Very Good Ideas. Online Submission (2005)
Jonassen, D.H.: Objectivism versus constructivism: do we need a new philosophical paradigm? Educ. Tech. Res. Dev. **39**(3), 5–14 (1991)
Kaya, G., Altun, A.: A learner model for learning object based personalized learning environments. In: Research Conference on Metadata and Semantic Research, pp. 349–355. Springer (2011)
Kurvers, J., Dalderop, K., Stockmann, W.: Cursistprofielen Laaggeletterdheid NT1 & NT2. Technical report, Steunpunt Taal en Rekenen VE, Tilburg (2013)
Kurvers, J., van de Craats, I.: Literacy and second language in the low countries. In: Young-Scholten M (ed) Low-Educated Second Language and Literacy Acquisition Proceedings of the Third Annual Forum, Durham, pp. 17–23 . Roundtuit Publishing, Newcastle upon Tyne (2007)
Lehman, B., Matthews, M., Person, N.: What are you feeling ? Investigating student affective states during expert human tutoring sessions, pp. 50–59 (2008)
Lepper, M.R., Woolverton, M.: The wisdom of practice: lessons learned from the study of highly effective tutors. In: Aronson, J. (ed.) Improving Academic Achievement, Chap 7, pp. 135–158. Academic Press, New York (2002)
Maulsby, D., Greenberg, S., Mander, R.: Prototyping an intelligent agent through Wizard of Oz. In: ACM SIGCHI Conference on Human Factors in Computing Systems, Amsterdam, pp. 277–284 (1993). https://doi.org/10.1145/169059.169215
Miller, W.R., Rollnick, S.: Ten things that motivational interviewing is not. Behav. Cognit. Psychother. **37**(2), 129–140 (2009). https://doi.org/10.1017/S1352465809005128
Mislevy, R.J., Oranje, A., Bauer, M.I., von Davier, A., Hao, J., Corrigan, S., Hoffman, E.: Psychometric Considerations in Game-based Assessment. GlassLab (2014)
Neerincx, M.A.: Situated cognitive engineering for crew support in space. In: Personal and Ubiquitous Computing, vol. 15, no. 5, pp. 445–456 (2011). https://doi.org/10.1007/s00779-010-0319-3/fulltext.html

Neerincx, M.A., Lindenberg, J.: Situated cognitive engineering for complex task environments. In: Schraagen, J.M., Militello, L.G., Ormerod, T., Lipshitz, R. (eds.) Naturalistic Decision Making and Macrocognition, pp. 373–390. Ashgate Publishing Limited, Aldershot (2008)
Neerincx, M., Vught, W., Blanson Henkemans, O., Oleari, E., Broekens, J., Peters, R., Kaptein, F., Demiris, Y., Kiefer, B., Fumagalli, D., et al.: Socio-cognitive engineering of a robotic partner for child's diabetes self-management. Front. Robot. AI **6**, 118 (2019)
Nimon, K.F.: Statistical assumptions of substantive analyses across the general linear model: a mini-review. Front. Psychol. **3**(AUG), 1–5 (2012). https://doi.org/10.3389/fpsyg.2012.0032210.3389/fpsyg.2012.0032210.3389/fpsyg.2012.00322
OECD: Literacy in the Information Age: Final Report of the International Adult Literacy Survey. Technical report Statistics, Canada (2000)
Schouten, D.G.M., Deneka, A.A., Theune, M., Neerincx, M.A., Cremers, A.H.M.: An embodied conversational agent coach to support low-literate societal participation learning: design, development, and evaluation. Under review (2020)
Schouten, D.G.M., Paulissen, R.T., Hanekamp, M., Groot, A., Neerincx, M.A., Cremers, A.H.M.: Low-literates' support needs for societal participation learning: empirical grounding of theory- and model-based design. Cogn. Syst. Res. **45**, 30–47 (2017a). https://doi.org/10.1016/j.cogsys.2017.04.007
Schouten, D.G.M., Smets, N.J.J.M., Driessen, M., Fuhri, K., Neerincx, M.A., Cremers, A.H.M.: Requirements for a virtual environment to support the social participation education of low-literates. Univers. Access Inf. Soc. **16**(3), 681–698 (2016). https://doi.org/10.1007/s10209-016-0502-z
Schouten, D.G.M., Venneker, F., Bosse, T., Neerincx, M.A., Cremers, A.H.M.: A digital coach that provides affective and social learning support to low-literate learners. IEEE Trans. Learn. Technol. **11**(1), 67–80 (2017b). https://doi.org/10.1109/TLT.2017.2698471
Shute, V.J., Zapata-Rivera, D.: Adaptive educational systems. Adapt. Technol. Train. Educ. **7**(27), 1–35 (2012)
Stephanidis, C.: Adaptive techniques for universal access. User Model. User Adapt. Interact. **11**(1–2), 159–179 (2001)
Tadlaoui, M.A., Aammou, S., Khaldi, M., Carvalho, R.N.: Learner modeling in adaptive educational systems: a comparative study. Int. J. Mod. Educ. Comput. Sci. **8**(3), 1 (2016)
van de Pol, J., Elbers, E.: Scaffolding student learning: a micro-analysis of teacher–student interaction. Learn. Cult. Soc. Interact. **2**(1), 32–41 (2013). https://doi.org/10.1016/j.lcsi.2012.12.001
van de Pol, J., Volman, M., Beishuizen, J.: Scaffolding in teacher–student interaction: a decade of research. Educ. Psychol. Rev. **22**(3), 271–296 (2010). https://doi.org/10.1007/s10648-010-9127-6
van de Craats, I.: Obstacles on highway L2. In: Low-Educated Second Language and Literacy Acquisition: Research, Policy and Practice, pp. 149–163. Radboud University Nijmegen (2007)
Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press, Cambridge (1980)
Wood, D.: Scaffolding, contingent tutoring, and computer-supported learning. Int. J. Artif. Intell. Educ. **12**(3), 280–293 (2001)
Wood, D., Wood, H.: Vygotsky, tutoring and learning. Oxf. Rev. Educ. **22**(1), 5–16 (1996)

Dylan G. M. Schouten holds an MSc in human–technology interaction (TU/e, the Netherlands), focused particularly on user–system interaction experience. He is currently finalizing his PhD thesis on the design and evaluation of VESSEL, a Virtual Environment to Support the Societal participation Education of Low-literates (TU Delft, the Netherlands), while working as a game systems designer.

Pim Massink holds an MSc in applied cognitive psychology (Utrecht University, the Netherlands). He is currently working as a software consultant at a start-up in Switzerland and provides advice on user-friendly software to companies in sub-Saharan Africa.

Stella F. Donker is an associate professor of experimental psychology at Utrecht University. She received her PhD in medical sciences from the University of Groningen (2002).

Mark A. Neerincx is full professor of human-centered computing at TU Delft and principal scientist of perceptual and cognitive systems at TNO. Recent research focuses on the situated cognitive engineering of electronic partners (ePartners) that support the social, cognitive and affective processes in human–automation collaboration to enhance performance, resilience, health and/or well-being. Examples are the Horizon 2020 'Personal Assistant for healthy Life-style' (PAL) project that develops a physical and virtual robot for children with diabetes, the 'IUALL' project on inclusive design and ePartners that enhance citizens' participation in the (local) society, and the 'SWELL' project on sensing, modeling and support techniques for stress self-management).

Anita H.M. Cremers is senior scientist at TNO as well as professor ('lector') at Utrecht University of Applied Sciences (the Netherlands). She holds an MSc in Computational Linguistics (Tilburg University, the Netherlands) and a PhD degree in natural language human–computer dialogue (Eindhoven University of Technology, the Netherlands). Her current main fields of expertise are human–computer interaction design and co-design. She focuses mainly on users with limited cognitive and ICT skills.