




Investigating changes in self-evaluation of technical competences in the serious game *Serena Supergreen*: Findings, challenges and lessons learned

Felix Kapp¹  · Pia Spangenberg² · Linda Kruse³ · Susanne Narciss⁴

Received: 1 May 2018 / Accepted: 20 September 2019 / Published online: 22 October 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Self-evaluation of one's competences is considered a core factor in various domains of human functioning, including learning and instruction, as well as academic and vocational choices. Researchers from the fields of metacognition and learning, as well as motivation and learning have thus intensively investigated issues related to the self-evaluation of competences. Insights from both lines of research have been used in the serious game project SERENA to inform the selection and design of technical tasks and tutorial feedback strategies. The main goal of the SERENA project was to develop a serious game for adolescent females that fosters their self-evaluation of competence regarding technical tasks. This paper describes how insights from metacognition, motivation and feedback research were integrated to inform the game design. Furthermore, it reports two evaluation studies conducted with 93 students in real school settings. The findings reveal that girls' self-evaluation of competences assessed in terms of perceived technical competences and self-concept of technical abilities, as well as intrinsic motivation regarding technical tasks can be strengthened with the serious game *Serena Supergreen*. The log-file analyses indicate that seeking feedback and help within the game is associated with an increase in perceived competences. The challenges encountered within this applied research field are discussed.

Keywords Self-evaluation of technical competences · Serious games · Perceived competence for technical tasks · Self-concept of technical abilities

Introduction

Gender segregation of adolescent career plans in the fields of science, technology, engineering and mathematics (STEM) is an important issue in many countries (e.g., Liu 2018; Sikora and

✉ Felix Kapp
felix.kapp@tu-berlin.de

Pokropek 2012; Stoet and Geary 2018; Wang and Degol 2013, 2017), since women are still underrepresented in science, and technology professions. For instance, in Germany, the number of young women in technical training occupations remains low at approximately 10% (Federal Institute of Vocational Education and Training 2017). Established perceptions of gender roles within the workforce can carry negative connotations and disadvantage females, resulting in less vocational choices and an unequal financial situation over the life-span, when compared to males (Kanji and Hupka-Brunner 2015). Therefore, there is a strong societal interest to raise women's quotas in technology-related professions.

Eccles and Wigfield (Eccles 1989; Eccles 1994; Eccles and Wigfield 2002) have described career choices in terms of an expectancy value model. They consider the career choice of young adults as an achievement-related choice, depending on the subjective task value and the expectations of success. The subjective task value consists of: (a) incentive and attainment values, (b) a utility value, and (c) the costs connected to the decision. The expectations of success are strongly dependent on one's self-concept of abilities, and on perceptions of the task's demands. Women's assessment of their abilities is therefore a crucial variable when choosing careers in the technology sector. In line with this assumption, studies on the gender gap in STEM have identified women's low confidence in their technical abilities as one of the main reasons for the small proportion of females employed in STEM fields (Eccles 1994; Lent and Brown 1996; Liu 2018; Sagebiel 2005; Tellhed et al. 2017; Tellhed and Adolfsson 2018). Thus, a goal for instructional research is to target this factor by developing interventions that increase women's confidence in their technical abilities. Metacognitive and motivational research with a focus on the self-evaluation of one's competences should inform the design of these interventions.

Self-evaluation of competences: Insights from metacognitive and motivational research

Self-evaluation of competences is an important factor in many domains of human behavior, including learning and instruction, and academic and vocational choices. Researchers have thus examined the role of self-evaluation of competences from various theoretical perspectives (e.g., metacognition and learning, motivation and learning, well-being and mental health, social psychology, personality, educational and clinical psychology). Consequently, a variety of terms have been used to refer to the phenomenon of self-evaluation (e.g., calibration, monitoring accuracy, meta-comprehension, perceived competence, self-concept of ability). In the present study, we focus on two research domains to inform task and feedback design within instructional interventions: metacognition and motivation.

Metacognition research on self-evaluation of competences has mainly investigated the conditions and outcomes of accurate and inaccurate self-evaluations. According to discrepancy reduction models of self-regulated learning (e.g., Winne and Hadwin 1998) and theories of metacognition (e.g., Efklides 2011; Nelson and Narens 1990), accurate self-evaluation of competences is crucial for monitoring and control processes in learning. Studies within this line of research have been conducted in a variety of laboratory settings including nonexperimental (e.g., calibration paradigm; Kleitman and Stankov 2007), experimental designs (e.g., meta-comprehension studies; Dunlosky and Lipko 2007), and quasi-experimental (e.g., monitoring accuracy studies; Pieschl 2009; Thiede et al. 2003).

Moreover, a variety of measures have been used to assess students' self-evaluations and their accuracy in relation to learning outcomes (e.g., prospective and retrospective judgements

of learning; ease-of-comprehension; ease-of-learning; absolute and relative bias; see Hadwin and Webster 2013; Nelson and Narens 1990; Schraw 2009, for reviews). Findings from this body of research reveal that people are often inaccurate when evaluating their competences or performance. In research on the gender gap in STEM-career choices it has been found that girls tend to underestimate their competences, while boys belief in their competences (e.g., Liu 2018; Tellhed and Adolfsson 2018; Woodcock and Bairaktarova 2015). Yet, findings also show that students' self-evaluations of performance are malleable (Bernacki et al. 2015), and can be influenced by various instructional means, including opportunities for practice testing, delayed summarization and key-wording (see Dunlosky et al. 2013, for a review), as well as generative learning activities (e.g., constructing a concept map; Redford et al. 2012). Furthermore, feedback has been found to be an important source for adjusting self-evaluations (e.g., Callender et al. 2016; van Loon and Roebbers 2017). Finally, metacognitive research reveals that different measures of self-evaluation capture different aspects of metacognitive processes and products (see the Special Issue on calibration in Learning and Instruction, 2013; e.g., Alexander 2013). Thus, Hadwin and Webster (2013) recommended that researchers use several measures depending on the research goals of their study.

While the focus of metacognition research has been more on the accuracy and/or bias of self-evaluation of competences, motivation research has mostly focused on the level of self-evaluation of competences, how it relates to students' learning and performance, and the factors that contribute to changes in the level of self-evaluation (for recent reviews see Huang 2011; Muenks et al. 2018). Some studies have also addressed the issue of self-evaluation biases, including their benefits and costs (e.g., Bouffard and Narciss 2011; Leduc and Bouffard 2017). Within motivation research, the variety of self-evaluation measures is even broader than in metacognition research, ranging from more general and distal measures such as academic self-concept measures (e.g., mathematical or verbal self-concept; Marsh 1990), to measures within a concrete domain (e.g., perceived competence; Harter 1985; see also Gresham et al. 2000), and to specific measures regarding a specific task or set of tasks (e.g., initial confidence, Feather 1969; self-efficacy, Bandura 1977, 1997). The differences, commonalities, and predictive values of various self-evaluation measures have been thoroughly examined (e.g., Bong and Skaalvik 2003; Bong et al. 2012; Jansen et al. 2015; Marsh et al. 2019). A core implication is that researchers should select the level of specificity of their measures of self-evaluations, taking into account: (a) the level of specificity of the outcome variables (i.e., matching issue; Pajares 1996), and (b) the theoretical rationale for the study (Muenks et al. 2018).

Based on Bandura's social cognitive theory of human functioning (Bandura 1986, 1997), motivational research considers a high positive level of self-evaluation (despite being somewhat biased) to be adaptive because it acts as an inner resource that fuels the motivation to learn, promotes persistence in the face of difficulties, and protects against negative emotions (e.g., Bouffard and Narciss 2011; Butler 2011; Leduc and Bouffard 2017; Pajares 2001; Taylor and Brown 1988, 1994; Taylor et al. 2000). Empirically, this position has been partially supported by findings showing positive correlations among students' positive self-evaluations and their motivation, persistence, self-regulation and achievement (e.g., Multon et al. 1991).

In contrast, negatively biased or low self-evaluation of competences are considered and have been found to be detrimental for students' motivation and emotions (e.g., Narciss et al. 2011), performance after negative feedback (e.g., Eckert et al. 2006), willingness to engage in learning tasks (e.g., Narciss 2004), and persistence (e.g., Ferla et al. 2010). Hence, also this line of research provides support for the assumption that female's low self-evaluations of their STEM competences may be a core factor for their low rate of choosing professions in technical fields.

Bandura has investigated the role of self-evaluations in human functioning using the construct of self-efficacy, which refers to one's confidence in one's ability to accomplish the courses of action necessary to complete a task successfully (Bandura 1977, 1997). Self-efficacy is conceptualized as a task-specific kind of self-evaluation that can be changed through four sources of self-efficacy (Bandura 1977, 1997) there are four sources of self-efficacy: (1) enactive mastery or non-mastery experiences, (2) vicarious mastery or non-mastery experiences, (3) verbal persuasion and encouragement or discouragement, and (4) physiological experiences accompanying enactment. Bandura assumes that enactive or vicarious mastery experiences, as well as encouraging verbal persuasions positively influences self-efficacy. Bandura's social-cognitive theory of self-efficacy has led to the design of many powerful interventions in various fields, including education and therapy (Usher and Pajares 2008). It also offers a promising framework for informing the design of instructional interventions aimed at promoting positive self-evaluations of competence within the domain of technical tasks. Based on the self-efficacy framework, promoting positive self-evaluations requires that instructional design considers: (a) the selection and/or design of tasks that allow for mastery experiences, (b) the design of feedback strategies that help students to solve tasks individually to gain a sense of mastery. The feedback should furthermore provide (c) verbal persuasion to help students, in particular those who underestimate their competences, adjust their self-evaluation of competences. Serious games represent one recent innovative instructional format in which these design principles can be applied.

Serious game design informed by metacognition, motivation and feedback research

Serious games are digital games with a clear educational goal to provide the player with knowledge, skills or opinions (Marr 2010). At the same time, they contain core elements of entertainment games. They are considered to be open learning environments in which players can explore problems in an interactive manner, try out challenges, and receive and use various types of feedback reflecting the level of mastery achieved. Several authors point out that serious games provide the opportunity to experience task involvement, enjoyment and a sense of agency (i.e., self-efficacy), since actions within the game have an immediate and clear impact on the game world (e.g., Garris et al. 2002; Hacker 2017; Klimmt 2004). Thus, besides achieving learning objectives such as knowledge and skill acquisition, serious games provide the opportunity to engage in metacognitive processes, such as self-evaluation of one's competences. They, therefore, have the potential to help people modify the accuracy of their self-evaluations (e.g., their perceived competence). Meluso et al. (2012) found positive effects of a serious game on the science self-efficacy of 5th graders. Based on their model of Game-Based Learning, Plass et al. (2015) argued that players are engaged on cognitive, emotional, social and behavioral levels while playing the game. This engagement contributes to the achievement of learning goals, even in domains where perceptions of competences are initially low. Thus, serious games appear to be a promising approach to address girls' low perceptions of their technical competences. Games can provide girls with technical situations and tasks in which they can experience mastery, and in doing so, adjust their perceptions of technical competence. Yet, serious game designers are faced with the challenge to simultaneously provide players with an open environment that fosters agency *and* offers sufficient guidance to develop the cognitive and metacognitive skills necessary to become successful learners. To tackle this

challenge, the design of (a) the tasks or quests implemented in the game, and (b) the feedback strategies is crucial.

Design of quests Quests in serious games can be described as complex tasks consisting of various steps. They provide learners with occasions for enactive and/or vicarious agency opportunities. To offer opportunities for mastery experiences to a specific target group, quests need to be designed based on psychological task and competence analyses. This includes specifying the types of knowledge (in terms of task-related declarative or procedural knowledge) and cognitive operations that are relevant to the quest(s). Furthermore, it includes determining the levels of competences the target players would need to play the game, and what kind of typical errors they might commit. Thus, quests should be designed along the following four dimensions: (a) task-related knowledge and content, (b) cognitive operations necessary to solve the quests, (c) the format in which the quests are presented, and (d) the level of interactivity such as the scaffolds and feedback strategies provided for the task (Proske et al. 2012).

Design of interactive feedback strategies Since feedback has a strong influence on task involvement, enjoyment and self-efficacy, the design of feedback strategies within games has been identified as an important success factor (Merchant et al. 2014). Furthermore, metacognitive research reveals that feedback is an important source for adjusting self-evaluations (e.g., Callender et al. 2016; van Loon and Roebers 2017). Yet, designing feedback strategies for a serious game is challenging; there are many ways of providing feedback, and the feedback may consist of diverse informative components, including information about gains or losses, receiving points, or detectable changes in the game world as a consequence of one's actions (DeSmet et al. 2015; Wouters et al. 2013; Young et al. 2012). One approach to facilitate mastery experiences within a serious game is to accompany quests with interactive tutorial feedback strategies (Narciss 2008, 2013, 2017). According to Narciss (2013) designing a feedback strategy requires researchers to specify: (a) potential sources of feedback (e.g., the game environment itself, the avatar that is controlled by the player, and/or the non-player characters (i.e., any character in a game which is not controlled by a player; (Lim and Reeves 2010)), (b) regulation levels addressed by the feedback (e.g., task-related or related to the players' competences), and (c) content of the feedback messages within these levels. At the task level, feedback can inform about errors and/or successful steps. At the competence level, it can elicit progress with regard to the competences addressed by the quests (e.g., increases in the level of technical competences). Finally, the form and mode of presenting feedback content must be specified. Feedback strategies can be designed adaptively to the players' needs and provide tutorial components aimed at supporting students when they encounter difficulties. Non-player characters can serve to provide tutorial feedback components, as well as specific and global feedback about the development of the players' competences (i.e., elicit the level of achieved competence).

The present study

The contribution of the current study is to present findings from a design-based methodological approach in which the serious game *Serena Supergreen* was developed and its effects investigated empirically. The focus of the design and evaluation was to determine how players'

self-evaluation of competences and intrinsic motivation in technical tasks can be fostered throughout the game. *Serena Supergreen* contains 21 technical quests in the field of renewable energy, which have been designed based on competence analysis under the consideration of the level of expertise of 13 to 15 year-olds. The quests are accompanied by interactive feedback strategies: the game environment, non-player characters and the avatar provide the player with feedback within the game on both task and competence levels.

Research questions

The research questions of the two evaluation studies were threefold: Firstly, following the recommendation of using several self-evaluation measures (Hadwin and Webster 2013), we investigated the effects of the serious game on two measures of self-evaluation of competences—a more general measure addressing *self-concept of technical abilities*, and a more specific task-related measure addressing perceived competence in solving technical tasks (hereafter referred to as *perceived technical competence*). The latter measure was assessed during the game: in level one of the game, players were asked to apply for a job and respond to four items referring to concrete activities (e.g. “building and fixing technical devices”), which were related to the technical quests in the game. Using both general and specific measures of self-evaluation of competences allows us to compare their advantages and limitations in capturing changes of self-evaluations within a serious game. Secondly, we examined the effects of the game on intrinsic motivation for technical tasks. Thirdly, we explored how in-game behaviors relate to changes in the measures of self-evaluation. Given the gender gap in STEM-evaluations, we explored also gender differences for all the questions.

These research questions were addressed in two studies: In the first study, participants played a short version of the game in a classroom setting. The effects on motivation, self-concept of technical abilities and perceived technical competence were investigated with pre- and post-game questionnaires. In the second study, participants from two classes played the full game. Detailed log-file analysis allowed us to gain insights into the associations between in-game behavior and the effects of the game on self-evaluation of competences. The second study also explored whether there were differential effects of the game for boys and girls.

Design and evaluation of the serious game *Serena Supergreen*

The following section describes the serious game *Serena Supergreen*, and subsequently presents the two evaluation studies.

Design of the serious game *Serena Supergreen*

Genre and procedure *Serena Supergreen* is a point-and-click adventure in which the avatar *Serena* and her friends are confronted with various technical quests. In the introductory tutorial, the player learns the character’s basic controls, her interaction possibilities, and how to collect and combine items within the adventure. At the end of the introduction, the player can choose one of four different *Serena* avatars, at which point starts the game.

Storyline Serena Supergreen is a digital game, in which the player takes on the role of the avatar Serena—a girl wanting to go on vacation with her friends. In order to finance their vacation, she has to apply for various jobs in the local mall. Level one begins in Serena's room (see Fig. 1), where she prepares documents for a job application. After solving the first technical quest (changing a broken light bulb), players learn about Serena's quest - to go on holiday with her friends, she must earn money to cover the travel expenses.

Level two takes place at the mall and consists of several rooms, in which Serena is required to solve several quests to earn money for the vacation. She works in a pet shop (Fig. 2), a music store, a repair café, a quiz area, and an outdoor store. In each job, Serena is confronted with technical quests and accompanied by several non-player characters such as the pet shop owner, the music store owner or customers at the repair café.

The third level is situated on an island, where Serena and her two friends arrive accidentally while travelling. In order to save themselves they have to repair technical devices (e.g. panels of a solar-roof; rotor of a wind-mill, see Fig. 3).

The game ends with Serena and her friends rescuing themselves and leaving the island. During her adventure, Serena takes on different problem-solving roles and masters increasingly difficult technical quests.

Technical quest design Within the game, the player must solve 21 quests, which address various technical knowledge and content. These quests are embedded in the three levels of the storyline and their difficulties vary according to the three levels. For instance, on level one, the first quest is to change the light bulb in Serena's room. The quest is introduced while Serena is getting ready to work in the mall. Without the light, she is not able to choose her clothes, so she has to select an adequate new bulb, switch off the power and change the bulb in order to get to work. On level two, in the mall, technical quests are presented as part of her job (e.g., repairing the terrarium in the pet shop; repairing a solar cell of a mobile phone in the repair café). On level three, the technical quests are part of the challenges Serena and her friends have to master on the island (e.g. getting the solar panels to work for energy supply in the house; repairing a broken wind power system). 10 additional quests with no explicit technical background complement the technical quests (such as "find/talk to the best friend" or "agree on a travel destination"; see Table 1 for a summary).

The 21 technical quests were designed in four steps: Firstly, the topics of the quests were chosen based on an analysis of different curricula of vocational training in the field of

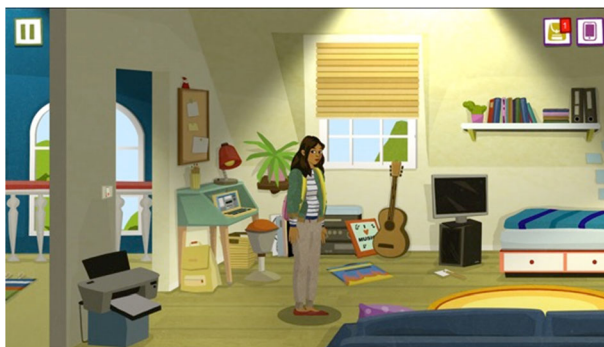


Fig. 1 Level one – Serena's room: players have to change the light bulb in order to get dressed and prepare for job applications at the mall



Fig. 2 An excerpt of level two - the mall: one part of the mall is the pet shop in which players face four technical quests

renewable energy (energy-saving sources should be present such as solar energy and wind power). In order to foster master experiences multiple opportunities for success were created by designing various quests for each topic. The number of quests varied for each topic according to the importance of the topic within the area of renewable energy. Thus, energy-saving and wind power are represented by four technical quests each within the game because of their societal importance and numbers of open positions in vocational training in Germany. E-mobility and solar power are still relevant but offer slightly less jobs, and were thus incorporated with two quests each. Secondly, we gathered information on the prior knowledge of the target group with regard to the selected topics by asking several experienced vocational teachers and the target group in two workshops. We used this information when deciding on the difficulty of the quests. Thirdly, we defined the learning objectives for the game in terms of competences (e.g., “being able to describe different characteristics of lamps in terms of lumen and watt”). Finally, the learning objectives were transformed into quests, which were implemented in the game story. Quests in level three (the island) were built on the learning objectives from quests in level two (the mall). For example, the “solar plant” quest from the island was based on concepts from the “solar charger” quest in the mall. The “fix the gondola” quest contains competences which were addressed in the quest “pet shop: fix water pump” and



Fig. 3 Level three - the island: the player has to repair a wind power system and a solar energy plant in order to save her friends

Table 1 Quests of Serena Supergreen

Level	Quest	Technical Task Description
Introduction	Release scissors and talk to monkey → Serena wakes up	None (tutorial on how to use the game)
Level one – Serena’s Room	Collect items for job application Fix the bulb	None Lamp and its characteristics (To get dressed, Serena must change a broken bulb; she has to decide which bulb is correct, and then change it)
Level two – the mall	Meet her friend Kiki Get mall ID Pet shop: fix freshwater Pet shop: fix seawater Pet shop: fix terrarium Pet shop: fix water pump Meet her friend Myra Charge e-bike battery Do e-mobility quiz in the mall Meet owner of repair café Build solar-charger Meet Tony (works at outdoor store) Install lamp Meet Myra Disassemble ventilation Print new gear in 3D printer Reassemble ventilation Meet Kiki and Myra	None (Story relevant) None Lamp and its characteristics Lamp and its characteristics Lamp and its characteristics Fluid mechanics None (Story relevant) E-mobility E-mobility None (Story relevant) Solar energy None (Story relevant) Electricity None (Story relevant) Mechanics of ventilation systems Mechanics of ventilation systems, 3D printer Mechanics of ventilation systems None (Story relevant)
Level three – the island	Arrive at island Evaluate house Look at flood gate Fix solar roof Experiment usage Fix gondola Find hole Make glue Fix rotor Open flood gate Happy end	None (story relevant) Electricity Mechanics Solar energy Electricity Wind power Wind power Wind power Wind power Mechanics Story relevant

“Disassemble ventilation”. This way, the player can experience mastery in more complex tasks while acquiring the necessary prior knowledge in the initial levels of the game.

All quests were analyzed and described with regard to the instruments or objects necessary to solve them and the minimum amount of steps needed. This analysis included the identification of possible problems or mistakes that could occur during the quest. Based on this error analysis and the modeling of quests, feedback strategies were designed.

Interactive feedback strategies within Serena Supergreen Interactive feedback strategies were designed based on the Interactive Tutoring Feedback Model (Narciss 2008; 2013; 2017). To promote positive development of the players’ self-evaluations of their technical competences, feedback was specifically aimed at addressing two core sources of self-efficacy (Bandura 1977): (a) mastery experiences – through providing evaluative and tutorial feedback components to enable successful accomplishments with the quests (tutorial feedback at task level), and (b) verbal persuasion – through the explicit communication of

positive and competence-related evaluative feedback components at the end of a set of quests (summative progress feedback at competence level).

The tutorial feedback strategies at the task level implemented for each quest provides players with mastery experiences (Narciss 2008, 2013, 2017) from three different sources. Firstly, as shown in the example from the pet shop (Fig. 4), the *game environment* offers immediate feedback on the outcome of a player's behaviors during a quest (e.g., if the player installs a lamp that produces excessive heat, the temperature of the aquarium rises and the fish begin to show unease). Secondly, the *avatar* monitors their actions in the quest by self-talk (e.g., "That does not look like a good idea. It is dangerous to change the bulb without switching the electricity off."). The avatar is immediately commenting when the player attempts to combine two objects which cannot be combined or tries to carry out actions which would harm the main character. If no action is carried out for over 30 s, the avatar is openly reflecting on the next steps, and by doing so offers hints on what do to next. Thirdly, the *non-player characters* give task-related advice and provide tutorial feedback, especially if players encounter difficulties in solving the quest themselves. The feedback sources are introduced to the players in the tutorial at the beginning of the game. During the first quest they can experience how the game environment offers feedback, how the avatar comments on actions, and how to approach a non-player character. The tutorial feedback components are designed to help players overcome obstacles and/or mistakes when solving the quest. They do not provide the correct answer or solution but instead, offer hints in finding necessary equipment and actions. To receive task-related advice, the player has to approach the non-player character where a menu with up to five possible questions appears. These are questions Serena can ask



Fig. 4 Task-related tutorial feedback strategy in a quest in the pet shop (quest “aquarium – level two, the mall”). The players receive information from the game environment, the avatar and the non-player character

the non-player character. They relate to the current task and contain typical errors. Once the question is asked the tutorial dialogue begins.

The summative progress feedback strategy at the competence level consists of the following events. First, at the beginning of the game, players are required to assess their technical competences (i.e., generate internal feedback) in order to submit a job application. They are asked to respond to five statements relating to how they perceive their core competences needed for the jobs in the mall on a scale from one to ten (see Fig. 5).

Second, after finishing the various jobs in the mall, players receive feedback from the non-player character regarding their competence development. This feedback refers to the perceived technical competence statements rated at the beginning of the game (Fig. 5). The initial self-reported responses serve as the baseline for computing the current state of the five perceived competence items. Since we aimed to foster positive self-perceptions of competence (especially for girls), a decrease in the scores is unlikely to occur. In the case of a player scoring a high initial value, but subsequently has difficulties with a quest, the value is expected to stay the same because of affirmation from the non-player character. In the case of a player scoring a low initial value, but then they successfully solve the quests (in terms of time and actions) the values would rise up to three points (e.g., one point for successfully solving the quest, two points for solving the quest without any mistakes and an initial value above five, and three points for solving the quests without any mistakes and an initial value below five). The adjusted values are communicated to the players via the mall-app and the corresponding non-player character. Figure 6 shows the feedback dialogue at the end of the pet shop sequence. The player has just completed four technical quests, is paid for the job, and a competence-related summative feedback from the pet shop owner appears. The increase on the five perceived competence statements is visualized on Serena's smartphone.

Evaluation study 1: Effects on self-evaluation of technical competences

Research questions In the first evaluation study we addressed four research objectives. We investigated the effects of the serious game on self-evaluation of competences, and on intrinsic motivation, and if these effects differ for boys and girls. Furthermore, we compared changes in

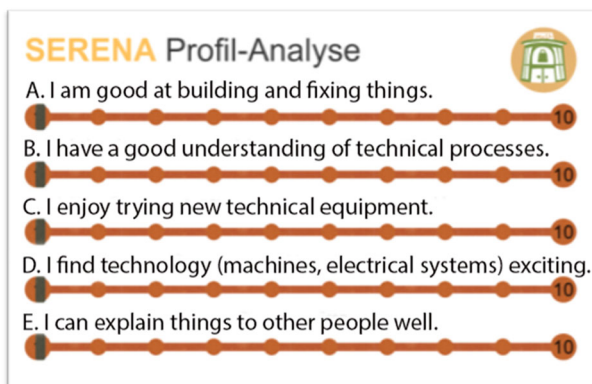


Fig. 5 The competence related self-assessment statements capturing perceived technical competence as part of the job application quest



Fig. 6 Summative feedback by the Non-Player Character (NPC) in the pet-shop. The increase on the five job-related competences is shown on Serena's smartphone. Dark green units indicate initial levels of perceived competence, light green units were added by the NPC

self-evaluation of competences when captured with a more general versus a more specific measure. More specifically, we aimed at contributing empirical findings to the following research questions: (1) To what extent are players' self-evaluation of competences more positive after playing the game than before? (2) To what extent is players' intrinsic motivation more positive after the game than before? (3) Do the effects of the game differ for boys and girls? (4) How sensitive are more general versus specific self-evaluation measures with regard to changes in self-evaluation of competences?

Method: Participants, design and procedure Fifty-four students (38 girls, age 13 to 15) from two German secondary schools participated in the study. A pre- and post-game design was used to investigate the effects of the game on self-evaluation of competences. All participants played the serious game *Serena Supergreen* for approximately two hours. The playing time was restricted to two hours as the study was conducted in the normal schedule of a real school setting. That is, a short version of the game consisting of the introduction, Serena's room and the mall was used (levels 1 and 2, respectively). Participants played the game either on an iPad mini ($n = 20$) or at a PC station ($n = 34$). All participants played individually and with earphones to listen to the sound effects. The study took place in the computer science room of the participating schools. Students firstly responded to the pre-game questionnaires, and then played *Serena Supergreen* for 60 min. After the first playing period all participants had a 10-min break, and then played the game for another 60 min. The study ended with the post-game questionnaires. In case of technical problems while playing, participants could ask for help from one of the present research assistants.

Measures In the pre-game questionnaires, demographic variables and experience with computer games were assessed. The effects of the game on the self-evaluation of competences were investigated with two measures of self-evaluation. The more general *self-concept of technical abilities* was assessed before and after playing the game, using five items (e.g. "In the domain of technology I am...", from "not talented" (1) to "very talented" (6), Cronbach's $\alpha = .83_{pre}/.82_{post}$). This scale has been adapted from the "self-concept of school-relevant-ability scale" (Schöne et al. 2012). The more specific *perceived technical competence* measure was assessed within and after the game. In level one of the game, players were asked to self-assess five job-related competences in order to apply for a job in the mall. Four items address aspects

of perceived competence with regard to technical abilities (e.g., “I am good at building and fixing things.”, from 1 to 10, Cronbach’s $\alpha = .85_{pre}/.84_{post}$; see Fig. 5). The fifth item refers to players’ ability to explain things in general and was not included. A printed version of these questionnaires was given to the participants post-game.

Participants’ *intrinsic motivation related to technical tasks* was measured before and after playing the game with two scales adopted from the Expectancy-Value Form of Learning Motivation (Kisielski and Narciss 2018). The scales address (a) task interest (three items, e.g., “I do not find the solving of technical tasks interesting at all.” (reverse-scored), Cronbach’s $\alpha = .68_{pre}/.59_{post}$), and (b) task enjoyment (three items, e.g. “I really like to work on technical tasks”, Cronbach’s $\alpha = .92_{pre}/.94_{post}$). Response options ranged from 1 – “totally disagree” to 6 – “totally agree”. The distinction between the two scales is based on the assumption that content-related value and activity-related value are related but separate components of intrinsic motivation (e.g., Rheinberg and Engeser 2018).

Statistical analysis The changes in self-concept of technical abilities and perceived technical competence were analysed firstly using a MANOVA with repeated measures (the two measures correlated at $r = .56$, $p < .001$) and subsequent ANOVAs with repeated measures. Changes with regard to intrinsic motivation (task interest and task-enjoyment; $r = .61$, $p < .001$) were analysed using a MANOVA with repeated measures. Due to the unequal distribution of boys and girls within the sample non-parametric tests were used to analyse possible gender effects. The Wilcoxon signed-rank test analysed if the self-concept and the intrinsic motivation changed from pre to post for boys and for girls. The Kruskal-Wallis test compared boys and girls pre and post with regard to the dependent variables.

Results

An analysis of the control variables revealed that neither demographic variables nor experience with computer games had an influence on the change of the dependent variables. The MANOVA with repeated measures for the self-evaluation of competence measures revealed significant changes from pre- to post-game (*Wilks Lambda* = .78, $F(2,32) = 4.55$, $p < .05$, $\eta^2_p = .22$). The follow-up ANOVAs revealed that students’ judgements with regard to their general self-concept of technical abilities were not significantly different when comparing pre- and post-game responses ($F(1, 51) < 1$, $M_{pre} = 4.09$, $SD_{pre} = .79$; $M_{post} = 4.10$, $SD_{post} = .80$). However, a significant increase in scores for the specific measure of perceived technical competence was detected. Due to missing log-files the data of only 35 participants were used for the statistical analysis ($F(1, 34) = 9.66$, $p < .01$, $\eta^2_p = .22$, $M_{pre} = 6.36$, $SD_{pre} = .1.80$; $M_{post} = 6.99$, $SD_{post} = 1.64$). Descriptive statistics are shown in Table 2.

The MANOVA with repeated measures for the intrinsic motivation measures revealed no significant overall change from pre- to post-assessment (*Wilks Lambda* = .90; $F(2, 48) = 2.76$, $p = .07$, $\eta^2_p = .10$). Yet, descriptive statistics indicate an increase of task interest (see Table 2).

With regard to differences between boys and girls an analysis of the specific evaluation measure (i.e., perceived technical competence) was not possible due to missing log files and the resulting small sample size (27 girls, 8 boys). The Kruskal-Wallis test for the general self-concept of technical abilities revealed no differences between boys and girls before ($H(1) = 2.421$, $p = .120$) and after ($H(1) = 1.221$, $p = .269$) playing the game. The Wilcoxon signed-rank tests did not indicate differences for boys pre-post ($n = 15$, $z = -.728$, $p = .467$) nor girls

Table 2 Descriptive statistics for the self-concept of technical abilities, perceived technical competence and intrinsic motivation related to technical tasks measures, pre- and post-game

	Pre		Post		n
	M	SD	M	SD	
Self-concept of technical abilities ¹	4.09	.79	4.10	.80	52
Perceived technical competence ²	6.36	1.80	6.99	1.64	35
Task enjoyment ³	3.95	1.15	3.95	1.30	51
Task interest ³	4.32	.93	4.60	.94	52

¹ range from 1 (e.g., “not talented”) to 6 (e.g., “very talented”); ² range from 1 (“low”) to 10 (“high”); ³ range from 1 (“totally disagree”) to 6 (“totally agree”)

pre-post ($n = 37$, $z = -.667$, $p = .505$) for the general self-concept of technical abilities. The reported intrinsic motivation did not change for girls pre-post with regard to task enjoyment ($n = 37$, $z = -.885$, $p = .376$) and task interest ($n = 37$, $z = -1.044$, $p = .297$). Boys reported a higher task interest ($n = 15$, $z = -2.436$, $p < .05$) after they had played the game. The task enjoyment did not change over time ($n = 15$, $z = -.946$, $p = .344$). The Kruskal-Wallis tests indicate no differences between boys and girls in the pre-assessment ($H_{\text{task enjoyment pre}}(1) = .907$, $p = .341$; $H_{\text{task interest pre}}(1) = 3.749$, $p = .053$). However, boys reported more interest in the post-assessment than girls ($H_{\text{task interest post}}(1) = 8.568$, $p < .01$). No differences with regard to task enjoyment were found in the post-assessment ($H_{\text{task enjoyment post}}(1) = 2.202$, $p = .138$). Descriptive data for the general self-concept of technical abilities and motivation are displayed for boys and girls in Table 3.

Brief discussion

The present research design appears to be a promising approach to capture changes in self-evaluation measures within a serious game. The results of the first evaluation study shows that playing the Serena Supergreen game in a controlled classroom setting for approximately two hours had positive effects on players’ self-evaluations of competences, but only when assessed via the specific perceived technical competence measure. There were no changes from pre- to post-game on the more general measure of self-concept of technical abilities. The perceived

Table 3 Descriptive statistics for the self-concept of technical abilities and intrinsic motivation measures pre- and post-game for girls and boys

	Girls				Boys			
	Pre		Post		Pre		Post	
	M	SD	M	SD	M	SD	M	SD
Self-concept of technical abilities ¹ (37 girls, 15 boys)	3.96	.75	4.01	.74	4.40	.84	4.32	.92
Task enjoyment ³ (37 girls, 15 boys)	3.89	1.03	3.74	1.28	4.20	1.47	4.38	1.28
Task interest ³ (37 girls, 15 boys)	4.17	.94	4.35	.87	4.69	.84	5.19	.83

¹ range from 1 (e.g., “not talented”) to 6 (e.g., “very talented”); ² range from 1 (“low”) to 10 (“high”); ³ range from 1 (“totally disagree”) to 6 (“totally agree”)

competence measure was embedded at the beginning of the game, and answered again post-game. The items addressed specific perceptions of competences necessary for solving quests in the game. Playing the short version of the serious game did not significantly affect the intrinsic motivation for technical tasks, but data indicate positive changes of task interest. The lack of increase in intrinsic motivation and in the more general self-evaluation measure suggests limitations related to the length and duration of playing the serious game. Meluso et al. (2012) could show an improvement in science self-efficacy for students who played a serious game for 150 min over a period of four days. The unequal distribution of boys and girls within the sample and the small amount of collected log files made it difficult to investigate gender differences. Another limitation is related to the sensitivity of the global self-evaluation measure. These limitations were addressed in the second evaluation study.

Evaluation study 2: Gender specific effects on self-evaluation and intrinsic motivation

Research questions In the second evaluation study, the objectives were threefold. Firstly, we investigated changes in (a) self-evaluations of competences, and (b) intrinsic motivation when students played the complete version of the game. Secondly, we examined whether there were gender-specific effects on these variables. In general, girls tend to self-evaluate their STEM competences less favourably than boys (e.g., Liu 2018; Tellhed and Adolffsson 2018; Watt et al. 2012). Hence, the question arises as to whether changes in self-evaluation and intrinsic motivation related to technical tasks occur differentially or similarly for boys and girls. Thirdly, we used log-files to explore how a player's behaviors (such as seeking feedback) relate to changes in self-evaluations of competences. Specific hypotheses and questions were derived to address these objectives: (1) Based on the findings of the first study, we expected that playing the full version of Serena Supergreen would lead to (a) more positive self-evaluations of technical competences, and (b) increases in intrinsic motivation (for both task enjoyment and task interest). (2) Regarding gender-specific effects, we expected that girls would benefit more than boys from the game, because if they initially have lower self-evaluation and intrinsic motivation scores, there might be more room for improvement for them. (3) How does in-game behavior, namely seeking feedback and help from a non-player character, relate to improvements in the self-evaluation of competences? Based on Bandura's (1997) theory on the role of verbal persuasion, we expect that interactions with non-player characters will lead to an increase in the self-evaluations of technical competences.

Method: Participants, design and procedure Thirty-nine students (19 girls, 20 boys, $M_{age} = 15.13$, $SD = 0.47$) of a German secondary school participated in the second study. All participants played the full version of the serious game Serena Supergreen on an iPad. After an introduction and completion of pre-game questionnaires, students played the game in the classroom over two regular lesson blocks of 90 min each. After each lesson block, there was a 10-min break. Excluding the breaks, students spent 180 min playing the game. Afterwards they completed the post-game questionnaires.

Measures In the pre-game questionnaires, demographic variables and prior experience with computer games were assessed. The measures of self-evaluation were the same as in study 1 (*self-concept of technical abilities* Cronbach's $\alpha = .92_{pre}/.91_{post}$; *perceived technical*

competence Cronbach's $\alpha = .85_{pre}/.93_{post}$). Both measures were assessed pre- and post-game. Participants' *intrinsic motivation related to technical tasks* was measured pre- and post-game with the same scales as in the first study (task interest Cronbach's $\alpha = .82_{pre}/.88_{post}$; task enjoyment Cronbach's $\alpha = .96_{pre}/.94_{post}$). The log-files of each player were analyzed using the following behavioral metrics: (a) the number of solved quests, (b) the number of interactions with non-player characters, (c) the duration of interactions with non-player characters, and (d) the number of actions used to solve the first five technical quests. A *performance index* was also calculated for each player (Quest Performance = minimum number of actions needed / number of actions used), with a higher value indicating better performance. The first five technical quests were used to calculate the performance index because 90% of the participants completed them. Finally, (e) the total number of actions carried out during the game was extracted for each player from the log-files.

Statistical analysis The changes in self-concept of technical abilities and the perceived technical competence were analysed for boys and girls using a MANOVA with repeated measures (the two measures correlated at $r = .75, p < .001$) and subsequent ANOVAs. Changes in intrinsic motivation (task interest and task enjoyment) were analysed by means of a MANOVA (the two measures correlated at $r = .81, p < .001$) with repeated measures. The relation between in-game behavior and changes in the self-evaluation of technical competences were analysed by linear regressions. ANOVAs were used for analysing the differences between girls and boys with regard to in-game behavioural data.

Results

The MANOVA with repeated measures for the global and specific self-evaluation measures revealed a significant main effect (*Wilks Lambda* = .83, $F(2,36) = 3.65, p < .05, \eta^2_p = .17$), indicating there were significant changes in the measures of self-evaluations of competences from pre- to post-game. Students reported higher self-concept of technical abilities after playing the serious game ($F(1,37) = 4.37, p < .05, \eta^2_p = .11, M_{pre} = 4.34, SD_{pre} = .91; M_{post} = 4.53, SD_{post} = .82$). For the more global measure, there was no interaction effect between gender and time, however there was a significant main effect of gender ($F(1,37) = 22.72, p < .001, \eta^2_p = .38$). Girls rated their technical abilities less favourably than boys, averaged over time (see Table 4). Similarly, perceived technical competence scores also increased post-game ($F(1,37) = 6.30, p < .05, \eta^2_p = .15, M_{pre} = 6.77, SD_{pre} = 2.02; M_{post} = 7.14, SD_{post} = 1.98$), and there was a significant main effect of gender ($F(1,37) = 22.12, p < .001, \eta^2_p = .37$), indicating that girls reported lower ratings. For the more specific measure, the interaction between time and gender was statistically significant ($F(1,37) = 8.03, p < .01, \eta^2_p = .18$). Perceived competence increased more for girls than for boys (see Table 4 for descriptive statistics).

Participants' intrinsic motivation related to technical tasks increased significantly from pre- to post-game (*Wilks Lambda* = .77, $F(2,36) = 5.51, p < .01, \eta^2_p = .23$). Subsequent ANOVAs were used to explore the two motivation components in further detail. Task enjoyment increased significantly from pre- to post-game ($F(1,37) = 9.08, p < .01, \eta^2_p = .20$). Task interest, however, did not significantly change over time ($F < 1$). Changes in both motivational components were similar for boys and girls (see Table 4).

Table 4 Descriptive statistics for the self-concept of technical abilities, perceived technical competence and intrinsic motivation related to technical tasks measures pre- and post-game

	Girls (<i>n</i> = 19)				Boys (<i>n</i> = 20)			
	Pre		Post		Pre		Post	
	M	SD	M	SD	M	SD	M	SD
Self-concept of technical abilities ¹	3.77	.55	4.08	.61	4.88	.86	4.96	.78
Perceived technical competence ²	5.34	1.26	6.17	1.55	8.11	1.64	8.06	1.94
Task enjoyment ³	3.19	1.17	3.86	0.80	4.37	1.36	4.63	1.28
Task interest ³	3.90	.80	4.14	1.06	4.83	1.13	4.88	1.23

¹ range from 1 (e.g., “not talented”) to 6 (e.g., “very talented”); ² range from 1 (“low”) to 10 (“high”); ³ range from 1 (“totally disagree”) to 6 (“totally agree”)

To gain a deeper insight into individual differences in in-game behaviour, as well as differences between girls and boys playing the game, the log-files for each of the thirty-nine players were analysed. Table 5 depicts the descriptive statistics for the number of solved quests, the number of interactions with non-player characters, the duration of these interactions, the total amount of actions and the performance index.

In-game behavior and its relations to self-evaluation measures Linear regressions were conducted to analyze the prediction of in-game behavior on the changes in perceived technical competence and self-concept of technical abilities. The number of interactions with non-player characters who provided feedback predicted the increase in both perceived technical competence ($F(1,37) = 5.90$, $p = .020$, $R^2 = .14$; $b^* = .37$) and self-concept of technical abilities ($F(1,37) = 8.40$, $p = .006$, $R^2 = .19$; $b^* = .43$). Girls and boys differed with respect to the duration of interactions ($F(1,33) = 16.69$, $p < .01$, $\eta^2_p = .34$), the total amount of actions within the game ($F(1,33) = 8.81$, $p < .01$, $\eta^2_p = .21$), and the number of solved quests ($F(1,33) = 5.00$, $p < .05$, $\eta^2_p = .13$). Girls spent more time interacting with non-player characters, carried out a significantly lower amount of actions during the game, and solved less quests (see Table 4 for descriptive statistics). Boys and girls started with the same level of technical competences as reflected in a similar performance index over the five first technical quests ($F < 1$).

Table 5 In-game behavior recorded via log-files

	Girls (<i>n</i> = 17)		Boys (<i>n</i> = 18)		All (<i>n</i> = 35)	
	M	SD	M	SD	M	SD
Number of solved quests	13.71	1.99	15.22	2.02	14.49	2.12
Number of interactions with NPC	28.82	6.46	27.67	4.52	28.23	5.50
Duration of interactions with NPC (sec)	1054.3	169.8	822.2	166.2	935.0	203.1
Total amount of actions within the game	1650	263	1904	244	1781	281
Performance index ^a	.47	.06	.45	.10	.46	.08

NPC non-player character

^a Performance index indicates the average performance for the five first technical quests of the game calculated by performance = minimum number of actions need to solve the quest / number of actions used. A higher number indicates better performance

Brief discussion

The results of the second evaluation study reveal that the game positively influences perceived technical competence, self-concept of technical abilities and intrinsic motivation for technical tasks. The positive effect for the perceived technical competence was stronger for girls than for boys. Boys and girls acted differently within the game. Girls spent more time interacting with non-player characters. They solved less quests than the boys and carried out a significantly smaller amount of actions during the game. Our findings support prior work on the role of feedback for adjusting self-evaluations (e.g., Callender et al. 2016; van Loon and Roebers 2017), and the theoretical assumption that feedback strategies delivered via non-player characters is a core element to help students adjust and strengthen their self-evaluations of competences. We found a strong relationship between the number of interactions with non-player characters and the increase in both perceived technical competence and self-concept of technical abilities. Students who frequently interacted with non-player characters within the game reported larger increases in both measures of self-evaluation. Furthermore, in line with prior findings on gender-related differences in self-evaluations in STEM domains (e.g., Watt et al. 2012), we found that girls evaluated their technical competences significantly lower than boys (in both measures of self-evaluation). However, there was no difference between boys and girls with regard to the performance index, which measured how many actions players needed to solve the first five technical quests within the game. The latter two findings are of particular interest, as they suggest that performance cannot account for the gender-specific self-evaluation differences, since there were no differences in performance.

General discussion

Serious games are promising environments for self-regulated learning in important domains such as STEM. The main contribution to the literature was developing a design-based research approach that evaluated the potential for the serious game *Serena Supergreen* to increase self-evaluations of competences, particularly in girls. The game-design process was informed by insights from metacognition, motivation and feedback research. The two evaluation studies investigated the extent to which the game had beneficial effects on players' perceived technical competence, self-concept of technical abilities, and intrinsic motivation related to technical tasks. The second study furthermore allowed comparing the effects of the game for boys and girls. Moreover, we analyzed in-game behavior to investigate if theoretically-derived sources of self-evaluations (namely, feedback in the form of verbal persuasion from non-player characters) relate to changes in self-evaluations.

In summary, the results indicate that the game fosters increases in perceived technical competence, self-concept of technical ability and intrinsic motivation related to technical tasks. Analysis of in-game behavior shows that feedback strategies within the game lead to positive effects on self-evaluations of competences. The two studies reveal that it is possible to increase in particular female players' self-evaluation of competences. Since, there were no differences in the starting level of performance between male and female players, however female players started with a significantly lower level of perceived competence, this finding might be interpreted as an indicator for a more accurate self-evaluation adjustment. Yet, the issue of how effectively a serious game may improve players' self-evaluation of competences deserves further studies analyzing in more detail the relations between the levels of self-evaluation and

performance. Furthermore, in order to make players aware of the accuracy of their self-evaluations, future studies might think of including feedback strategies that start with asking students to generate firstly internal feedback (e.g., by self-evaluation reflection prompts), and then offer external feedback as recommended by the ITFL-model (Narciss 2008, 2013, 2017). We refrained from doing so for the present study, since we did not want to interrupt the storyline too much during the game. Yet, such interactive, reflective feedback strategies would be worth to be investigated in future studies.

The two studies furthermore address several challenges when investigating the conditions and effects of using serious games as an intervention to strengthen girls' self-evaluations of their technical competences, and intrinsic motivation for technical tasks.

In the first study, we detected an increase in the level of self-evaluation for the specific measure of perceived technical competence, but not for the more general measure of self-concept of technical abilities. After playing the two first levels of the game over a time period of approximately two hours, players reported higher perceived competence (e.g., "building and fixing things" and "understanding how technical processes work") compared to at the beginning of the game.

In the second study, the results from the first evaluation study were replicated and extended. Students played the complete game and we found positive effects for both measures of self-evaluation (i.e., perceived technical competence, self-concept of technical abilities), and the intrinsic motivation measures, particularly for task enjoyment. In line with Meluso et al. (2012), who found an increase in science self-efficacy of 5th graders after playing a serious game, the game positively influenced self-evaluation. This effect was stronger for girls than for boys. Boys and girls also acted differently within the game, such that girls spent more time interacting with non-player characters and carried out less actions during the game. Our findings substantiate the theoretical assumption that feedback strategies delivered via non-player characters play a vital role in strengthening self-evaluations of technical competences. We found a relation between self-evaluation of competences and the number of interactions with non-player characters. A higher frequency of interactions with non-player characters was related to a greater increase in the self-evaluation of technical competence measures. As non-player characters were designed to deliver verbal persuasion, these findings are in line with research of Wright et al. 2016 who found verbal persuasion to be an effective source of self-efficacy. Informed by design principles, our findings suggest that (a) tasks (i.e. game quests) that allow mastery experiences, and (b) feedback strategies that help players adjust self-evaluation of competences, are effective in promoting task-specific perceptions of technical competences. These design principles were derived from Bandura's social-cognitive theory of self-efficacy (1977, 1997), and from prior metacognitive research on the role of feedback for developing accurate self-evaluations (e.g., Callender et al. 2016; van Loon and Roebbers 2017).

The finding that scores on the more general measure of self-evaluations of technical abilities remained unchanged in the first study (short version of the game), but increased in the second study (full version of the game), is worthy of discussion in light of prior theoretical and empirical work investigating self-evaluation processes on distinctive levels. Bong and Skaalvik (2003) argued that more general self-concept of abilities is characterized by a relative temporal stability and develops over a longer period of time, while more task-specific self-evaluations such as perceived competence or self-efficacy are considered to change relatively quicker (see also Marsh et al. 2019; Muenks et al. 2018).

With regard to the debate about the generality versus specificity of metacognitive processes and how these processes are executed (e.g., Veenman et al. 1997), our findings indicate that novices and/or students who hold negatively-framed stereotypes need sufficient experiences

with tasks in a given domain, in order to base their self-evaluations on concrete task experiences, rather than their (inaccurate) competence beliefs. Future studies with serious games are encouraged to contribute further empirical data on the nature of the underlying metacognitive processes of self-evaluation of competences.

Limitations and challenges of assessing changes in self-evaluation of competence

The present research highlights the challenges of assessing changes in self-evaluation of competence. Firstly, we had to consider the trade-off between convenience (i.e., short scales) and sensitivity of the self-report instruments used to capture changes in self-evaluation of competences. We used a global standardized measure, because technical tasks in the field of renewable energies are inherently diverse. This global measure addressed the self-concept of one's abilities in technical domains with items similarly used in academic self-concept questionnaires (e.g., "In the domain of technology I am...", "not talented" (1) to "very talented" (6); e.g. Marsh 1990). There were no changes in students' self-evaluation of technical abilities when this global measure was used in the first study. Yet, the more specific measure of self-evaluation (i.e. perceived competence in technical tasks; see also Harter 1985) revealed changes in both studies. Players were more confident in their abilities to build and fix things, and in their understanding of technical processes after playing the game in the controlled setting. Evaluation of the serious game *Serena Supergreen* thus revealed challenges within the field of self-evaluation, including the need to assess whether measures are adequate and sensitive enough to render concrete changes in self-evaluation of technical competences.

Another methodological debate refers to how and when to measure self-efficacy, perceived competences in tasks, and self-concept. The timing, frequency, and manner in which questionnaires are presented is often debated when designing serious games. On the one hand, games provide the possibility to embed self-evaluation measures, and connect them to log-files and behavioral data from the game. This is especially promising for investigating the dynamic role of self-evaluation (Bernacki et al. 2015). On the other hand, it is crucial that the game experience, interest and enjoyment is not jeopardized by interruptions of the measurement process.

Limitations and challenges related to authentic settings

In both studies, participants played the game in an authentic classroom setting. Methodological limitations are associated with this setting, for instance, all students are in one room and interactions between them may not always be prevented. More control over the setting would be desirable in future studies, such as screens to limit interactions. As the game has the potential to be used in far more open and uncontrolled settings, such as at home for homework, the effects of the game may differ in other settings and contexts. A current challenge facing researchers is the trade-off between investigating the effects of the game in controlled versus uncontrolled settings.

A related methodological challenge refers to the complexity of the game as an intervention. The game gives autonomy for players to navigate the quests themselves, and combines a mixture of different technical tasks with specific feedback strategies and game features. The average completion time is approximately four hours. Although *Serena Supergreen* gives tutorial feedback, initiates internal feedback and provides verbal persuasion, it is difficult to

infer conclusions when testing the game as a complete intervention. That is, individual effects of the feedback strategies could not be partialled out. The log-files allowed us to gain insights into metacognitive processes during play. Still, these insights are limited to the first and the second level of the game and in case of the performance index to the first five quests as we could only collect a sufficient amount of log files for these quests. It remains a considerable challenge in applied research projects to systematically test every possible variation of what constitutes effective feedback strategies. Future studies should address the impact of particular feedback strategies on singular elements of the game.

Implications for further research and instructional practice

Serena Supergreen represents a design-based research approach in an applied field. Based on our studies, we conclude that serious games are valuable tools in addressing issues related to the promotion of self-evaluation of competences. However, we were not able to disentangle the specific effects of the various feedback strategies in the game. Several issues warrant further investigation. Firstly, what is the added value of combining the four self-efficacy sources within a serious game? Secondly, which of the four self-efficacy sources contribute most to the increase in positive self-evaluations? Existing research assumes that Bandura's sources of self-efficacy have different effects. Wright et al. (2016) report promising results for verbal persuasion but no specific effect for vicarious experiences. Thirdly, what kind of interaction with the game is necessary to positively affect self-evaluation (i.e., reduce negative and/or positive biases in self-evaluation), and how can these interactions be measured? Lastly, the present study only assessed short-term effects of the serious game. Whether these effects are retained and are observable over a longer period of time requires longitudinal study designs.

In sum, serious games provide rich environments in which learners can calibrate and increase their self-evaluation of competences through interactions with non-player characters and experiences of mastery after completing quests. Future research should investigate the underlying mechanisms, influencing factors, and effects of strengthened and more accurate self-evaluation of competences.

Acknowledgements We would like to express our sincere gratitude to Lisa Zhang for her native speaker advice and the help provided with the final editing of the manuscript. Furthermore, we are very grateful for the thoughtful comments of three anonymous reviewers.

Funding Information This work has been supported by two funds through the German Federal Ministry of Education and Research (BMBF; 01PD14005; 01PD17005).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24*, 1–3.

- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191–215.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs: Prentice Hall.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Bernacki, M. L., Nokes-Malach, T. J., & Alevan, V. (2015). Examining self-efficacy during learning: variability and relations to behavior, performance, and learning. *Metacognition and Learning*, *10*(1), 99–117.
- Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, *15*(1), 1–40.
- Bong, M., Cho, C., Ahn, H. S., & Kim, H. J. (2012). Comparison of self-beliefs for predicting student motivation and achievement. *The Journal of Educational Research*, *105*(5), 336–352.
- Bouffard, T., & Narciss, S. (2011). Benefits and Risks of Positive Biases in Self-evaluation of Academic Competence: Introduction. *International Journal of Educational Research*, *50*(4), 205–208. <https://doi.org/10.1016/j.ijer.2011.08.001>.
- Butler, R. (2011). Are positive illusions about academic competence always adaptive, under all circumstances: New results and future directions. *International Journal of Educational Research*, *50*(4), 251–256.
- Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, *11*(2), 215–235.
- DeSmet, A., Shegog, R., Van Ryckeghem, D., Crombez, G., & De Bourdeaudhuij, I. (2015). A systematic review and meta-analysis of interventions for sexual health promotion involving serious digital games. *Games for Health Journal*, *4*(2), 78–90.
- Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228–232.
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58.
- Eccles, J. S. (1989). Bringing Young Women to Math and Science. In M. Crawford & M. Gentry (Eds.), *Gender and Thought* (pp. 36–58). New York: Springer.
- Eccles, J. S. (1994). Understanding Women's Educational and Occupational Choices. Applying the Eccles et al. Model of Achievement-Related Choices. *Psychology of Women Quarterly*, *18*(4), 585–609. <https://doi.org/10.1111/j.1471-6402.1994.tb01049.x>.
- Eccles, J. S., & Wigfield, A. (2002). Motivational Beliefs, Values and Goals. *Annual Review Psychology*, *53*, 109–132.
- Eckert, C., Schilling, D., & Stiensmeier-Pelster, J. (2006). Einfluss des Fähigkeitsselfbsteckonzepts auf die Intelligenz- und Konzentrationsleistung [Influence of the self-concept of abilities on intelligence and concentration]. *Zeitschrift für Pädagogische Psychologie*, *20*(1/2), 41–48.
- Eklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, *46*(1), 6–25.
- Feather, N. T. (1969). Attribution of responsibility and valence of success and failure in relation to initial confidence and task performance. *Journal of Personality and Social Psychology*, *13*, 129–144.
- Federal Institute for Vocational Education and Training (Germany) (2017). Report on Vocational Education and Training 2017. Accessible through www.bmbf.de/upload_filestore/pub/Berufsbildungsbericht_2017_eng.pdf
- Ferla, J., Valcke, M., & Schuyten, G. (2010). Judgments of self-perceived academic competence and their differential impact on students' achievement motivation, learning approach, and academic performance. *European Journal of Psychology of Education*, *25*(4), 519–536.
- Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, *33*, 441–467.
- Gresham, F. M., Lane, K. L., MacMillan, D. L., Bocian, K. M., & Ward, S. L. (2000). Positive and negative illusory biases: Comparisons across social and academic self-concept domains. *Journal of School Psychology*, *38*, 151–175.
- Hacker, D. J. (2017). The role of metacognition in learning via serious games. In R. Zheng & M. K. Gardner (Eds.), *Handbook of Research on Serious Games for Educational Applications* (pp. 19–40). Hershey: IGI Global.
- Hadwin, A. F., & Webster, E. A. (2013). Calibration in goal setting: Examining the nature of judgments of confidence. *Learning and Instruction*, *24*, 37–47.
- Harter, S. (1985). Competence as a dimension of self-evaluation: Toward a comprehensive model of self-worth. In R. Leary (Ed.), *The development of the self* (pp. 55–122). New York: Academic Press.
- Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology*, *49*(5), 505–528.

- Jansen, M., Scherer, R., & Schroeders, U. (2015). Students' self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology*, *41*, 13–24.
- Kanji, S., & Hupka-Brunner, S. (2015). Young women's strong preference for children and subsequent occupational gender segregation: What is the link? *Equality, Diversity and Inclusion: An International Journal*, *34*(2), 124–140. <https://doi.org/10.1108/EDI-05-2014-0041>.
- Kisielski, K., & Narciss, S. (2018). Development and Validation of a domain-specific Expectancy-Value Form of Learning Motivation. Paper presented at International Conference on Motivation (ICM), Aarhus, Denmark.
- Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*(2), 161–173.
- Klimmt, C. (2004). Computer- und Videospiele [Computer and Video Games]. In R. Mangold, P. Vorderer, & G. Bente (Eds.), *Lehrbuch der Medienpsychologie* (pp. 695–716). Göttingen: Hogrefe.
- Leduc, C., & Bouffard, T. (2017). The impact of biased self-evaluations of school and social competence on academic and social functioning. *Learning and Individual Differences*, *55*, 193–201.
- Lent, R. W., & Brown, S. D. (1996). Social Cognitive Approach to Career Development: An Overview. *The Career Development Quarterly*, *44*(4), 310–321. <https://doi.org/10.1002/j.2161-0045.1996.tb00448.x>.
- Lim, S., & Reeves, B. (2010). Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies*, *68*(1-2), 57–68.
- Liu, R. (2018). Gender-Math Stereotype, Biased Self-Assessment, and Aspiration in STEM Careers: The Gender Gap among Early Adolescents in China. *Comparative Education Review*, *62*(4), 522–541.
- Marr, A. C. (2010). *Serious Games für die Informations- und Wissensvermittlung - Bibliotheken auf neuen Wegen* (Vol. 28) [Serious Games for Information and Knowledge Transfer - Libraries on New Paths (Vol. 28)]. Wiesbaden: BIT Verlag.
- Marsh, H. W. (1990). The structure of academic self-concept: The Marsh/Shavelson model. *Journal of Educational Psychology*, *82*, 623–636.
- Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology*, *111*, 331–353.
- Meluso, A., Zheng, M., Spires, H. A., & Lester, J. (2012). Enhancing 5th graders' science content knowledge and self-efficacy through game-based learning. *Computers & Education*, *59*(2), 497–504.
- Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicut, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers and Education*, *70*, 29–40.
- Muenks, K., Wigfield, A., & Eccles, J. S. (2018). I can do this! The development and calibration of children's expectations for success and competence beliefs. *Developmental Review*, *48*, 24–39.
- Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, *38*(1), 30–38.
- Narciss, S. (2004). The impact of informative tutoring feedback and self-efficacy on motivation and achievement in concept learning. *Experimental Psychology*, *51*, 214–228.
- Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology* (3rd ed., pp. 125–144). Mahwah: Lawrence Erlbaum Associates.
- Narciss, S. (2013). Designing and Evaluating Tutoring Feedback Strategies for Digital Learning Environments on the Basis of the Interactive Tutoring Feedback Model. *Digital Education Review*, *23*, 7–26.
- Narciss, S. (2017). Conditions and effects of feedback viewed through the lens of the Interactive Tutoring Feedback Model. In D. Carless, S. M. Bridges, C. K. Y. Chan, & R. Glofcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 173–189). Singapore: Springer.
- Narciss, S., Koemdle, H., & Dresel, M. (2011). Self-evaluation accuracy and satisfaction with performance: Are there affective costs or benefits of positive self-evaluation bias? *International Journal of Educational Research*, *50*(4), 230–240.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125–173.
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543–578.
- Pajares, F. (2001). Toward a positive psychology of academic motivation. *Journal of Educational Research*, *95*, 27–35.
- Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition and Learning*, *4*(1), 3–31.
- Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, *50*(4), 258–283.
- Proske, A., Kördle, H., & Narciss, S. (2012). Interactive learning tasks. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (Vol I(9)), pp. 1606–1610. Heidelberg: Springer.

- Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction, 22*(4), 262–270.
- Rheinberg, F., & Engeser, S. (2018). Intrinsische Motivation und Flow-Erleben [Intrinsic motivation and flow]. In J. Heckhausen & H. Heckhausen (Eds.), *Motivation und Handeln* (pp. 423–450). Berlin, Heidelberg: Springer.
- Sagebiel, F. (2005). Gendered Organisational Cultures in Engineering. Theoretical Reflection on Women. Results and Future Research Perspectives. In A. Thaler & C. Wächter (Eds.), *Creating Cultures of Success for Women Engineers. Conference proceedings* (pp. 143–154). IFF/IFZ: Graz.
- Schöne, C., Dickhäuser, O., Spinath, B., & Stiensmeier-Pelster, J. (2012). *Skalen zur Erfassung des schulischen Selbstkonzepts [Scales for the assessment of the academic self-concept]* (2nd ed.). Göttingen: Hogrefe.
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45.
- Sikora, J., & Pokropek, A. (2012). Gender segregation of adolescent science career plans in 50 countries. *Science Education, 96*(2), 234–264.
- Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science, 29*(4), 581–593.
- Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193–210.
- Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin, 116*, 21–27.
- Taylor, S. E., Kemeny, M. E., Reed, G. M., Bower, J. E., & Gruenewald, T. L. (2000). Psychological resources, Positive illusions, and health. *American Psychologist, 55*, 99–109.
- Tellhed, U., & Adolfsso, C. (2018). Competence and confusion: How stereotype threat can make you a bad judge of your competence. *European Journal of Social Psychology, 48*(2), 189–197.
- Tellhed, U., Bäckström, M., & Björklund, F. (2017). Will I fit in and do well? The importance of social belongingness and self-efficacy for explaining gender differences in interest in STEM and HEED majors. *Sex Roles, 77*(1-2), 86–96.
- Thiede, K. W., Anderson, M. C. M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73.
- Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research, 78*(4), 751–796.
- van Loon, M. H., & Roebers, C. M. (2017). Effects of Feedback on Self-Evaluations and Self-Regulation in Elementary School. *Applied Cognitive Psychology, 31*(5), 508–519.
- Veenman, M. V., Elshout, J. J., & Meijer, J. (1997). The generality vs domain-specificity of metacognitive skills in novice learning across domains. *Learning and Instruction, 7*(2), 187–209.
- Wang, M. T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review, 33*(4), 304–340.
- Wang, M. T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review, 29*(1), 119–140.
- Watt, H. M., Shapka, J. D., Morris, Z. A., Durik, A. M., Keating, D. P., & Eccles, J. S. (2012). Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: A comparison of samples from Australia, Canada, and the United States. *Developmental Psychology, 48*(6), 1594–1611.
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. *Metacognition in Educational Theory and Practice, 93*, 27–30.
- Woodcock, A., & Bairaktarova, D. (2015). Gender-biased self-evaluations of first-year engineering students. *Journal of Women and Minorities in Science and Engineering, 21*(3), 255–269.
- Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249–265.
- Wright, B. J., O'Halloran, P. D., & Stukas, A. A. (2016). Enhancing self-efficacy and performance: an experimental comparison of psychological techniques. *Research Quarterly for Exercise and Sport, 87*(1), 36–46.
- Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., Simeoni, Z., Tran, M., & Yukhymenko, M. (2012). Our princess is in another castle a review of trends in serious gaming for education. *Review of Educational Research, 82*(1), 61–89.

Affiliations

Felix Kapp¹ · Pia Spangenberg² · Linda Kruse³ · Susanne Narciss⁴

Pia Spangenberg
pia.spangenberg@tu-berlin.de

Linda Kruse
linda@thegoodevil.com

Susanne Narciss
susanne.narciss@tu-dresden.de

¹ Department of Psychology and Ergonomics, Technische Universität Berlin, Marchstr. 23, D-10587 Berlin, Germany

² Institute for Vocational Education and Work Studies, Technische Universität Berlin, Berlin, Germany

³ the Good Evil GmbH, Game Studio, Köln, Germany

⁴ Psychology of Learning & Instruction, Technische Universität Dresden, Dresden, Germany