Evaluating the usability and instructional design quality of Interactive Virtual Training for Teachers (IVT-T)

Elisa S. Shernoff, Katherine Von Schalscha, Joseph L. Gabbard, Alban Delmarre, Stacy L. Frazier, Cédric Buche & Christine Lisetti

Educational Technology Research and Development
A bi-monthly publication of
the Association for Educational
Communications & Technology

ISSN 1042-1629
Volume 68
Number 6

Education Tech Research Dev (2020)
68:3235-3262
DOI 10.1007/s11423-020-09819-9

Springer

Education Tech Research Dev (2020) 68:3235–3262
https://doi.org/10.1007/s11423-020-09819-9

DEVELOPMENT ARTICLE

Evaluating the usability and instructional design quality of Interactive Virtual Training for Teachers (IVT-T)

Elisa S. Shernoff[1] · Katherine Von Schalscha[1] · Joseph L. Gabbard[2] · Alban Delmarre[3] · Stacy L. Frazier[4] · Cédric Buche[5] · Christine Lisetti[3]

Published online: 6 September 2020

Abstract
Simulation offers unique affordances over traditional training (e.g., remote access, mastery learning experiences, immediate feedback) relevant to teacher training in behavior management. This study describes a user-based evaluation of Interactive Virtual Training for Teachers (IVT-T). The study involved observing representative users (seven advanced education majors) perform benchmark tasks with the system, complete rating scales, and participate in interviews to evaluate the usability and instructional design quality of IVT-T. Global usability ratings based on established usability rating scales suggested IVT-T was adequately usable while observations of user performance and semi-structured interviews revealed design shortcomings that impeded effective user performance and informed ways to improve the interface. Observations of user performance, for example, identified 36% of usability problems related to learning, 19% = screen design, 17% = terminology; 3% = system capabilities and 25% = other problems. Cross analysis of user semi-structured interviews pointed to the system's ability to convey believable, visually appealing, realistic characters and classrooms. More contextual cues, multiple challenging behaviors featured at the same time, and changes to the visual appearance of the classroom would enhance realism. Revisions made to enhance the usability and instructional design elements of IVT-T are discussed. In addition, implications for teacher educators and researchers involved in the development of instructional technologies are summarized along with the potential value of including simulation in teacher training for behavior management.

Keywords Professional development · Educational technology · Teacher training · Technology integration · Mixed-methods

Introduction

Interactive simulations have emerged to support learners across many disciplines (Aslan and Reigeluth 2016; Graafland et al. 2015; Thompson and McGill 2017). Pilots are trained to navigate challenging flying conditions with simulation (Fletcher 2009), new therapists

✉ Elisa S. Shernoff
elisa.shernoff@rutgers.edu

Extended author information available on the last page of the article

learn to conduct risk assessments for suicidal clients using simulation (Beutler and Harwood 2004; Horswill and Lisetti 2011), and preservice teachers use virtual role play to identify and respond to student bullying (Schussler et al. 2017). A primary goal of simulation training is to immerse users in realistic scenarios to improve their work performance (Bellotti et al. 2010; Regalla et al. 2016).

Interactive Virtual Training for Teachers (IVT-T; Shernoff et al. 2018), which is the focus of the current study, is a simulation training model being developed and refined with funding from the Institute of Education Sciences (Grant # R305A150166). IVT-T is designed for early career teachers working in high poverty schools to bolster their behavior management skills and reduce turnover through simulated practice responding to disruptive characters in a virtual classroom. Simulation training models provide support in an area in which teachers indicate is their greatest professional development need—dealing with disruptive behaviors (Evertson and Weinstein 2006; Owens et al. 2018). Simulation training models can also minimize a trial-and-error approach to behavior management with real students. Effective prevention and management of disruptive behaviors should also theoretically improve classroom climate and student–teacher relationships, thus reducing the number of students referred for more intensive behavior problems (Epstein et al. 2008; Henry et al. 2000).

Development of simulation training models requires targeted evaluation of usability in addition to graphics, content, and instructional design elements to maximize learning and transfer. Thus, three goals guided the current study: (1) evaluating IVT-T usability by observing representative users interact with the system; (2) assessing the authenticity and realism of the characters, classrooms, and storylines; and (3) exploring instructional design quality, including practice, reflection, and feedback. We wanted to identify and address problems with usability, fidelity, and instructional design early in the design life cycle, before extensive resources were allocated to creating animations and programming (Bowman et al. 2002; Gabbard and Swan 2008; Hartson and Pyla 2012).

Simulation training models for teachers

Although the use of technology to support teacher professional development has tremendous promise to close the research-to-practice gap, few training models that use simulation technology to teach behavior management have been rigorously evaluated and more studies are needed to determine which approach for building and using virtual training systems is the most efficient and effective (Dawson and Lignugaris-Kraft 2017). Among the simulation training models that exist, their requirements and affordances differ and distinctive development decisions were made regarding graphics, user control, and system resources required.

SimSchool, for example, is designed for preservice teachers and relies on virtual students with unique profiles (Badiee and Kaufman 2015; Zibit and Gibson 2005). A recent usability study revealed that 76% of users rated the SimSchool content and curriculum as "good" or "very good" while 77% of users rated generalizability as "good" or "very good" (Rayner and Fluck 2014). Qualitative feedback indicated moderate support for its educational utility with three-quarters of the sample indicating concerns regarding realism of the simulated conversations (Rayner and Fluck 2014) which was echoed by teachers in a follow up usability study (Badiee and Kaufman 2015). VirtualPREX is a 3D classroom simulator built in the Second Life virtual world (Gregory and James 2011). VirtualPREX relies on 3D graphics and pre-scripted scenarios in which teacher trainees engage in interactive role plays that simulate challenging student–teacher interactions. Early studies by the developers indicate trainees experienced difficulties interacting with the system, with some requiring external notes to help them control the virtual student. Studies with SimSchool and VirtualPREX underscore the need to assess and ensure strong usability when developing a new technology for teachers.

TeachLivE (Dieker et al. 2017) and Breaking Bad Behavior (3B; Lugrin et al. 2016) use immersive technology that combines tangible, real-world elements (e.g., desks, chairs, whiteboards) with classroom scenarios displayed on a projector screen (Dieker et al. 2017). Both systems include a collaborative virtual reality environment in which the teacher (trainee) and instructor (operator of the system) interact in a shared virtual environment. In TeachLivE, teacher trainees stand in front of a screen and with 3B, trainees use virtual reality headsets to interact with virtual students who are controlled by a human instructor. Early usability studies with urban teachers suggest TeachLivE was realistic and compelling (Dieker et al. 2007) and a more recent study with middle school teachers (Dieker et al. 2017) indicated participants in the TeachLivE condition used more high-quality instructional questions at posttest (mean = 24%) than comparison teachers (Mean = 14%; $p = .002$). Another study of TeachLivE indicated transfer from the virtual to the live classrooms, with statistically significant improvements in special educators' use of proactive behavioral expectations ($p < .01$) and approval ($p < .01$) and reductions in student noncompliance ($p = .04$; Pas et al. 2016). These findings indicate promising results related to the TeachLivE platform and the potential benefit of human control of the virtual students which allows the simulation to adapt to users' actions. However, TeachLive and 3B cannot be used autonomously by teachers, as both require humans to operate the system, which imposes constraints as to when and how frequently each system can be accessed.

Interactive Virtual Training for Teachers (IVT-T; Shernoff et al. 2018), is a simulation training model being developed and refined for early career teachers working in high poverty schools to bolster their behavior management skills through simulated practice (see 'Methods' for detailed description of the system). IVT-T is being built with Unity3D (www.unity3d.com), a videogame platform in which teachers can access the system with their existing computing systems (laptops or desktops) which should maximize connectivity and access. IVT-T is also unique from other simulation training models in that it combines simulation technology with content and curriculum via programmed storylines and does not require human control of the characters. High poverty schools are also a targeted focus of IVT-T given teachers working in these contexts face unique stressors (e.g., overcrowding, high stakes accountability policies, and limited resources) that directly and indirectly contribute to classroom behavior problems (Atkins et al. 2015; Ouellette et al. 2018; Shernoff et al. 2011, 2016). Not only are annual turnover rates for new teachers higher in these settings (Atteberry et al. 2017; Guarino et al. 2006; Ingersoll and Strong 2011) but research also identifies struggles with disruptive behavior as new teachers' greatest professional development need (Evertson and Weinstein 2006; Owens et al. 2018; Shernoff et al. 2011, 2016).

Models and methods for assessing usability

Usability is broadly conceptualized as the quality of a user's experience when interacting with a product or system (Bowman et al. 2002). Given learnability, system functionality, and ease of use predicts adoption of new technologies (Ludwick and Doucette 2009), out first goal was to assess IVT-T usability. There are several standard approaches to assessing usability based on the stage of interface development and how usability problems are identified (Hartson and Pyla 2012; Mayhew 1999). A heuristic evaluation is an informal usability inspection technique for which an expert analyzes and identifies problems based on established usability guidelines (Nielsen 1993). Heuristic evaluations are typically conducted with early prototypes and supplemented with user-based evaluations in which representative users perform tasks in a laboratory with a more developed prototype (Nielsen 1993).

The current user-based evaluation of IVT-T was part of a four-year grant. At the start of Year 2, a usability engineer conducted a heuristic evaluation of IVT-T to identify surface-level design problems that could be remedied early in the design life cycle. During the latter part of Year 2, a user-based evaluation of IVT-T was conducted which involved observing representative users performing benchmark tasks with the system to identify usability problems not identified by the prior heuristic evaluation (Gabbard et al. 1999).

The role of fidelity in simulation training models

Representation fidelity was important to assess given it is a distinguishing feature of virtual training models that supports learning transfer and predicts the speed at which users adopt new technologies (Alessi and Trollip 2001; Ludwick and Doucette 2009; Whyte et al. 2015). Therefore, our second goal was to evaluate the authenticity and realism of the characters, classrooms, and storylines. The creation of authentic and realistic characters and classrooms was deemed important in order to immerse teachers in an environment that closely resembled their classrooms and the students with whom they interact. Logical and realistic storylines and realistic behaviors of students was theorized to support transfer of learning from one setting (e.g., virtual classroom) to a new setting (e.g., live classroom) if teachers confront similar situations and can carry forward knowledge and skills that are applicable in both contexts (Annetta et al. 2014; Sitzmann 2011).

Assessing instructional design to maximize learning

The third goal of the current study was to evaluate three instructional design elements embedded within the system. *Practice* provided opportunities for users to interact with disruptive characters and make decisions about how to respond to provocative and off task behavior. *Reflection* prompted users to consider their responses to disruptive behavior and reflect on ways to improve their approach in the future. *Feedback* allowed users to receive information on the effectiveness of their responses and suggestions for improving their choices in the future. A more detailed description of the IVT-T prototype and how trainees interacted with the storylines and how reflection and feedback were integrated into the prototype is described in the method.

These instructional design elements were informed by experiential learning theory (Kolb et al. 2000; Lindsey and Berger 2009) which emphasizes the critical role that extended practice plays when mastering complex skills. Practice forms the basis of reflection and problem-solving regarding how to improve future performance (Kolb et al. 2000; Lindsey and Berger 2009). Instructional design research further underscores that practice alone is unlikely to promote transfer without reflection and performance feedback (Richey et al. 2011; Tracey et al. 2014). Simulation training models provide a powerful context for transfer given a core design component includes user feedback while solving authentic work-based scenarios and problems (Dede 2009; Gresalfi and Barnes 2016). Practice, reflection, and feedback were also theorized to address limitations to existing teacher training in behavior management for which didactic strategies (e.g., presentation, demonstration, discussion) pervade despite research suggesting active learning (e.g., practice, reflection, and feedback) is critical to teacher skill acquisition and transfer (Desimone et al. 2002; Shernoff et al. 2015).

Research design

This study used qualitative methods (i.e., observations of user performance and semi-structured interviews) to examine user's in-depth experience with IVT-T. This included identifying usability problems and design flaws and eliciting constructive feedback to improve upon the characters, classrooms, and storylines in subsequent prototypes. Established quantitative usability measures supplemented qualitative methods to assess the global usability of the system. To summarize, this study sought to address the following research questions: (1) What types of usability problems emerged when users performed tasks within IVT-T? (2) What were user impressions of the authenticity and realism of the classrooms, characters, and storylines? and, (3) What were user impressions of instructional design elements of practice, reflection, and feedback?

Methods

Participants

This work was conducted with Institutional Review Board approval and in accordance with ethical guidelines for the protection of human subjects. Advanced education majors at a large university in a South Atlantic state were recruited to participate in the user-based evaluations. Virzi (1992) estimates sample sizes required to conduct formative evaluations, with five participants typically needed to identify 80% of the usability problems and nine users needed to identify 95% of usability problems in an interface. Therefore, our goal was to recruit between six and nine users who met the following inclusion criteria: (1) current education major, (2) interest in working in high poverty schools, and (3) completion of some field placement. Participants were recruited via campus list serves and further screened by project staff to ensure they met inclusion criteria. Eighteen potential users responded to original recruitment efforts, seventeen participants met inclusion criteria (one did not have any field experience), and the first seven who responded to follow up recruitment efforts were consented.

The demographic characteristics of participants are illustrated in Table 1. All participants were female, their mean age was 22.14 (SD = 3.13). Four participants were seeking Bachelor's Degrees and three were seeking Master's Degrees in Education. Two participants self-identified as Middle Eastern, the remainder self-identified as European American. With regards to behavior management training, one participant reported "none," five reported "some," and one reported "a great deal" of training in behavior management. Overall, the sample had limited experience playing videogames. Five participants indicated they were "non-gamers," spending less than one hour, on average, per week playing video games. One participant reported spending 5–7 h per week and one reported spending 1–3 h per week playing videogames.

Table 1 Descriptive data for each participant

	Participant	Description
1	Rachel	European American, Female, 29 years old, Seeking Master's Degree, No Prior Behavior Management Training, Gamer (5–7 h/Week Playing Video Games)
2	Janna	Middle Eastern, Female, 22 years old, Seeking Bachelor's Degree, Some Behavior Management Training, Non-Gamer (< 1 h/week Playing Video Games)
3	Dianne	European American, Female, 21 years old, Seeking Master's Degree, Some Behavior Management Training, Non-Gamer (< 1 h/week Playing Video Games)
4	Rania	Middle Eastern, Female, 22 years old, Seeking Master's Degree, A Great Deal of Behavior Management Training, Gamer (1–3 h/Week Playing Video Games)
5	Suzanna	European American, Female, 21 years old, Seeking Bachelor's Degree, Some Behavior Management Training, Non-Gamer (< 1 h/week Playing Video Games)
6	Stacey	European American, Female, 20 years old, Seeking Bachelor's Degree, Some Behavior Management Training, Non-Gamer (< 1 h/week Playing Video Games)
7	Betsy	European American, Female, 20 years old, Seeking Bachelor's Degree, Some Behavior Management Training, Non-Gamer (< 1 h/week Playing Video Games)

IVT-T design

IVT-T includes three components: (1) characters (first and sixth graders) who engage in off task and aggressive behaviors; (2) classrooms (one first grade, one sixth grade); and (3) storylines illustrating challenging interactions between teachers and students (Shernoff et al. 2018). The computer science team developed and refined thirty racially and ethnically diverse characters and two classrooms based on advisory board feedback during Year 1 of the funded grant (see Shernoff et al. 2018 for a description of the development and refinement of the graphics). Storylines were written during Years 1 and 2 by the first and fifth authors and guided by evidence-based behavioral strategies that inform the prevention and management of challenging behaviors (e.g., praise, ignore, redirect, proximity, instructions, punishment; Kazdin 2005; Simonsen et al. 2008). This approach to developing the IVT-T content was driven by the instructional potential of merging simulation technology with evidence-based behavior management strategies which has been limited to date. The user-based evaluation of IVT-T (focus of the current manuscript) took place at the end of Year 2, after the 3D graphics and prototypes of the storylines had been developed, and after the usability engineer had conducted a heuristic evaluation of IVT-T to identify early design problems.

Figure 1 illustrates an example storyline that was presented during the user-based evaluation of IVT-T. The blue hexagons represent teacher response options, the green rectangles depict how the character responds, and the salmon ovals indicate the end of the storyline. In this particular example, Jordan arrives late to class and is having difficulty getting started on the assigned Do Now. The user is provided with three options: (1) give Jordan a few minutes to settle into class. Smile and give him a thumbs up when you can catch his eye (i.e., wait time and nonverbal praise), (2) walk to Jordan's desk and say, "Copy down the answers for the Do Now" (i.e., proximity and instructions), and (3) say to Jordan, "Get

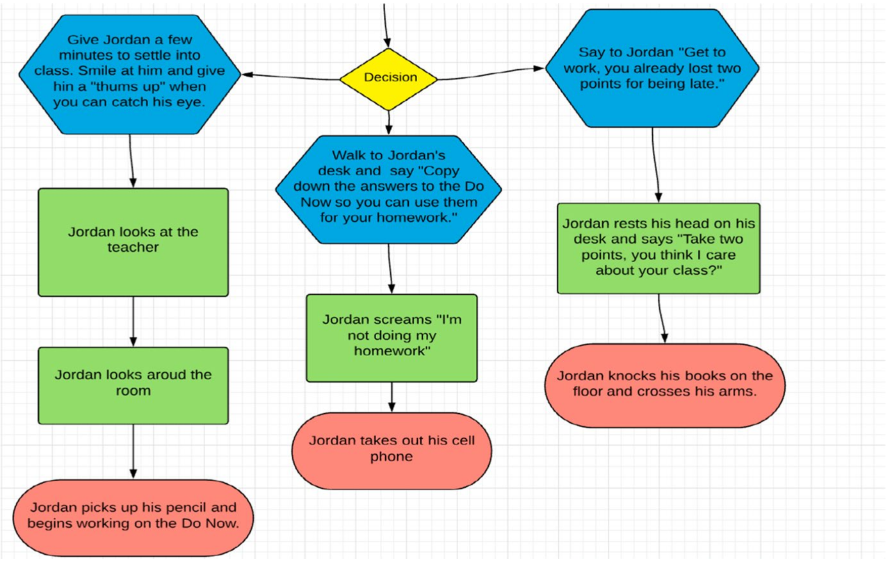

Give Jordan a few minutes to settle into class. Smile at him and give hin a "thums up" when you can catch his eye.
Decision
Say to Jordan "Get to work, you already lost two points for being late."
Walk to Jordan's desk and say "Copy down the answers to the Do Now so you can use them for your homework."
Jordan looks at the teacher
Jordan rests his head on his desk and says "Take two points, you think I care about your class?"
Jordan screams "I'm not doing my homework"
Jordan looks aroud the room
Jordan knocks his books on the floor and crosses his arms.
Jordan takes out his cell phone
Jordan picks up his pencil and begins working on the Do Now.

Fig. 1 Example Jordan storyline

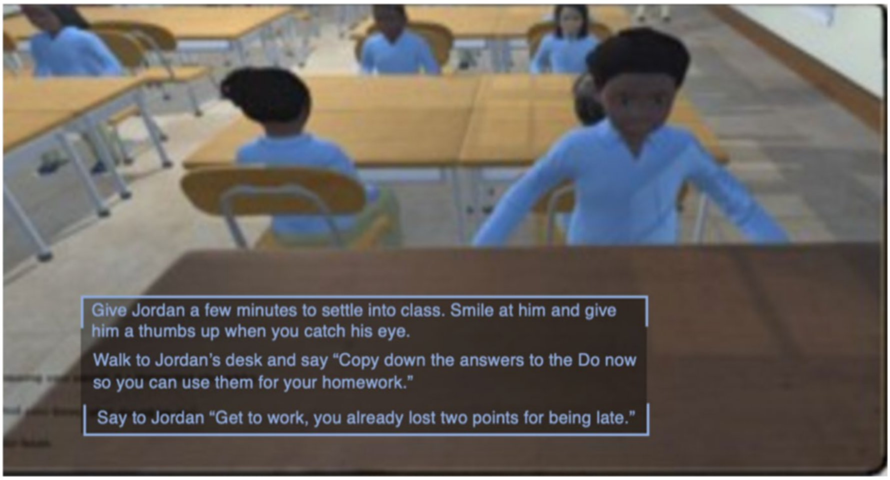
to work, you lost two points for being late (i.e., punishment). Storylines were intentionally written such that the character becomes more or less disruptive contingent on the teachers' response, with the goal of portraying the interactive (antecedent-behavior-consequence) cycle of disruptive behaviors in addition to the important role that the environment plays in maintaining those behaviors (Kazdin 2005; Shernoff and Kratochwill 2007).

IVT-T also included a three-phase training sequence. Phase 1 (Practice) was designed for users to interact with disruptive characters and make decisions about how to respond to provocative and off task behaviors. Phase 2 (Reflection) enabled users to review the responses they selected during the Practice Phase and describe via text entry how they could improve on their decisions in future training sessions. Phase 3 (Feedback) offered users numeric and descriptive information on the effectiveness of their responses and how they could improve their response selections in the future.

Procedures for conducting the user-based evaluations

After participants completed informed consent, the user-based evaluation began with a brief orientation to the system by the usability engineer followed by a concurrent think aloud protocol (CTA) to evaluate user experience with IVT-T and to identify instructional design and usability strengths and problems. After users completed the CTA protocol, they participated in a semi-structured interview and completed ratings scales.

A combination of Axure (www.axure.com) a rapid prototyping specification software tool, and PowerPoint was used to create the user-based evaluation prototype. Axure housed the practice and reflection interfaces while the feedback interface was presented in a series PowerPoint slides. The user-based evaluation prototype included a 2D version the storyline for Jordan, the sixth-grade character with aggressive and noncompliant behaviors in a



Give Jordan a few minutes to settle into class. Smile at him and give him a thumbs up when you catch his eye.

Walk to Jordan's desk and say "Copy down the answers to the Do now so you can use them for your homework."

Say to Jordan "Get to work, you already lost two points for being late."

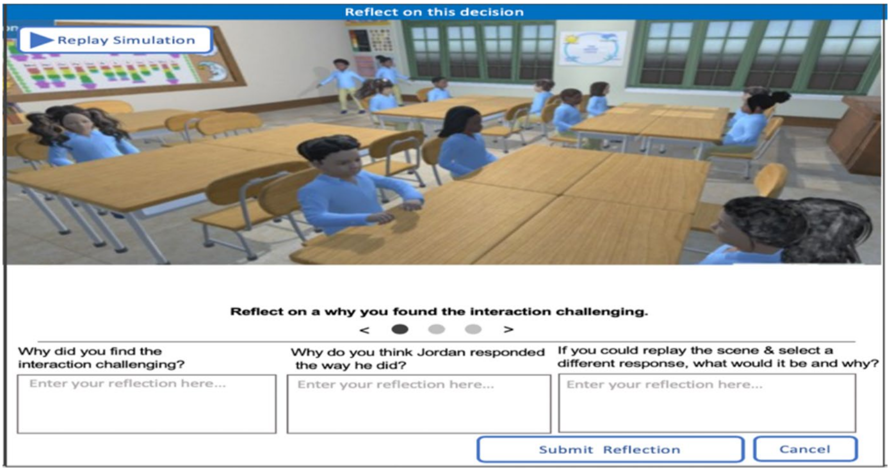
Fig. 2 Practice interface

classroom with 15 peers (see Fig. 1). Jordan arrives late to class and is having difficulty getting started on his work.

Figure 2 illustrates how users interacted with the system during the Practice Phase. This included user selecting from among three options for how to respond to Jordan's behavior.

Figure 3 illustrates how users interacted with the system during the Reflection Phase. This included prompting users to respond to the following three questions: (1) Why they found the interaction challenging, (2) Why they thought Jordan responded the way he did, and (3) If given the opportunity to select a different response, what would that response be and why.

Figure 4 illustrates how feedback was integrated into the system, including the quantitative feedback (e.g., effective = +1 and not effective = 0) and qualitative feedback along with the gold star indicating the total points awarded for that decision point.



Reflect on a why you found the interaction challenging.

Why did you find the interaction challenging?

Enter your reflection here...

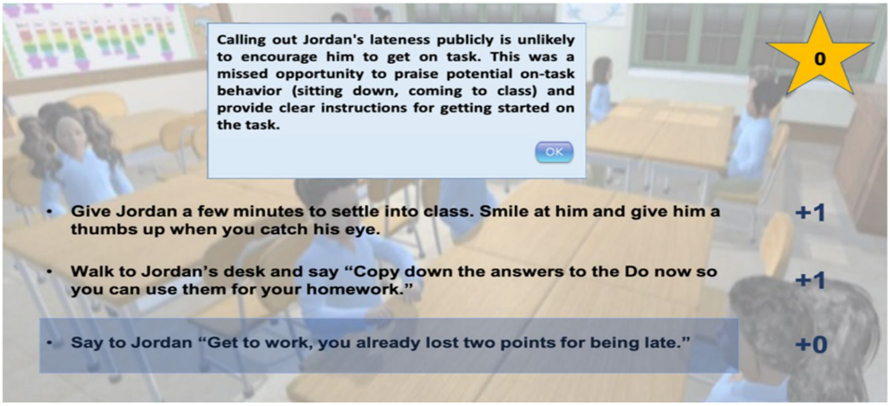
Why do you think Jordan responded the way he did?

Enter your reflection here...

If you could replay the scene & select a different response, what would it be and why?

Enter your reflection here...

Fig. 3 Reflection interface



Calling out Jordan's lateness publicly is unlikely to encourage him to get on task. This was a missed opportunity to praise potential on-task behavior (sitting down, coming to class) and provide clear instructions for getting started on the task.

- Give Jordan a few minutes to settle into class. Smile at him and give him a thumbs up when you catch his eye. +1
- Walk to Jordan's desk and say "Copy down the answers to the Do now so you can use them for your homework." +1
- Say to Jordan "Get to work, you already lost two points for being late." +0

Fig. 4 Feedback interface

Data collection

Concurrent think aloud protocol

The CTA prompted users to think aloud as they interacted with IVT-T (Cooke 2010; Jaspers 2009) while completing 40 pre-determined tasks (e.g., find items in the classroom, select and progress through a storyline, select decisions). The usability engineer provided the following instructions: "*I am giving you a list of tasks to complete. Please talk out loud while performing each task. If you stop talking for a while, I will remind you to keep talking*." The research team observed through a one-way mirror and took field notes summarizing usability issues while an audio and video track captured user interactions with the system. The average length of the CTA protocol was 89.28 min (SD = 15.04 min; Range = 65–115 min).

Semi-structured interviews

Users participated in semi-structured interviews which provided the development team with feedback on the authenticity and realism of the characters, classrooms, and storylines and key instructional design elements of practice, reflection, and feedback. An interview protocol was derived from the professional development and instructional design literatures and began with general questions, "*What was it like to use the IVT-T system?*" and "*What did you like the most/least about the system*?" These general questions were followed by open-ended probes for each specific area of inquiry (i.e., "*How realistic were the behaviors, actions, and dialogue of the main character*" "*How much did the vignettes [the interactions between the avatars and you the teacher] hold your attention"*). Mean interview length was 31 min (SD = 0.26, range 24–39 min).

System Usability Scale

The System Usability Scale (SUS; Bangor et al. 2008; Brooke 1996) is a standardized usability measure that includes 10 items related to ease of use and functional consistency (e.g., "*I found the system unnecessarily complex*" "*I would imagine that most people would learn to use this system very quickly*" and "*I felt very confident using the system*"). SUS items are rated on a 5-point scale (1 = *Strongly disagree* to 5 = *Strongly agree*) with negative valence items recoded. A factor analysis of the SUS indicated one significant factor (α = 0.90); hence, the measure is used as an index of overall usability (Bangor et al. 2008). Standardized coding procedures include multiplying the summed score by 2.5 to obtain an overall system usability value. Scores between 50 and 70 are considered marginal and scores above 71 are considered acceptable (Bangor et al. 2008; Brooke 1996; Finstad 2006).

Questionnaire for User Satisfaction

The Questionnaire for User Satisfaction (QUIS; Chin et al. 1988; Harper and Norman 1993) measures subjective satisfaction with specific aspects of human computer interfaces and guides the redesign of interactive systems in usability evaluations (Slaughter et al. 1995). QUIS items are rated on a 10-point scale ranging from 0 to 9 and anchored at both endpoints with adjectives (e.g., Difficult-Easy) positioned so that the scale moves from negative to positive. "Not Applicable" was also included for each item to tailor the evaluation to the particular interactive system being evaluated. The QUIS yields one Overall Reaction to the Software Scale which includes six items (i.e., Terrible-Wonderful, Difficult-Easy, Frustrating-Satisfying, Inadequate-Adequate, Dull-Stimulating, Rigid-Flexible). The QUIS also includes four additional subscales: (1) Screen Design and Layout Subscale (e.g., "*organization of information*" [0 = confusing …9 = very clear]); (2) Terminology and Systems Information Subscale (e.g., "*position of messages on screen"* [0 = inconsistent …9 = consistent]; (3) Learning Subscale (e.g., "*learning to operate the system"* [0 = difficult …9 = easy]); and (4) System Capabilities Subscale (e.g., "*system speed"* [0 = too slow …9 = fast enough]). Prior usability studies indicate average usability ratings range from 4.72 to 7.02 (Chin et al. 1988; Harper and Norman 1993; Slaughter et al. 1995). Prior research indicates the QUIS has an overall α = 0.94 (Chin et al. 1988; Harper and Norman 1993).

Data analysis

Descriptive analyses summarized overall usability ratings. To evaluate the types of usability problems that emerged when users performed tasks within IVT-T (Research Question #1), a collated list of usability problems identified during the CTA was summarized by the usability engineer through a review of merged field notes. The usability engineer then coded each usability issue identified during the CTA into one of the four QUIS subscales (i.e., Screen Design and Layout, Terminology and Systems Information, Learning, and System Capabilities) or an "Other" category and these informed subsequent design modifications.

To evaluate authenticity and realism of the characters, classrooms, and storylines (Research Question #2) and the instructional design elements of practice, reflection, and feedback (Research Question #3), a content analysis (Drisko and Maschi 2015) was conducted. The goal of these analyses included moving from a single case to a cross-case analysis to determine patterns of positive and negative impressions that could guide subsequent revisions. Interviews were coded and analyzed by one faculty member and a team of undergraduate and graduate students using Dedoose software (Lieber 2009).

Qualitative analyses involved several steps: transcribing the interviews, excerpting interview transcripts, developing the codebook, independent coding, establishing inter-rater agreement, and conducting a cross-case analysis (Miles and Huberman 2019). First, verbal content from the semi-structured interviews was transcribed verbatim and checked against digital recordings to verify accuracy (see McLellan et al. 2003). Second, interviews were excerpted into complete, coherent thoughts or ideas (segments) in which meaning could be extracted in isolation (Saldaña 2015). Third, coders created a preliminary codebook which included a start list of codes developed by the coding team via consensus through one round of open coding of one transcript (Miles and Huberman 1994). This step included creating codes (i.e., graphics quality, IVT-T instructional components) and subcodes (indicating positive and negative comments connected to the broader code). Fourth, individual coders were trained to review transcripts independently and apply codes to excerpted text. As the research team independently coded each transcript in pairs, the codebook was refined (six revisions total). Coding disagreements were discussed as a group, codes were refined, and transcripts were independently recoded. The final codebook included operational definitions of the codes, exclusion and inclusion criteria, coding hierarchies, and coding instructions (Fonteyn et al. 2008). Results suggest that interrater reliability between independent coders was high (Kappa = 0.92; range 0.76–1.0). Fifth, after coding all data, the team conducted a cross-case analysis using guidelines outlined by Miles and Huberman (2019). This step included synthesizing coded data across users to identify patterns or clusters of positive and negative impressions across users. The goal of the cross-case analysis was to organize and compare patterns of positive and negative impressions across the domains of interest to inform subsequent revisions to the system.

Results

Descriptive analyses of the System Usability Scale (SUS; Bangor et al. 2008; Brooke 1996) are reported in Table 2. SUS sum scores for individual participants ranged from 75 to 100, suggesting acceptable overall usability (i.e., 50–70 is considered marginal, > 71 is considered acceptable (Bangor et al. 2008; Brooke 1996; Finstad 2006).

Further descriptive analyses of the QUIS (Chin et al. 1988; Harper and Norman 1993) that comprise the Overall Reaction to the Software Scale (i.e., Terrible-Wonderful, Difficult-Easy, Frustrating-Satisfying, Inadequate-Adequate, Dull-Stimulating, Rigid Flexible)

Table 2 System usability scale (SUS) sum scores illustrating global usability

Participant	SUS sum score
Rachel	83
Janna	100
Dianne	75
Rania	80
Suzanna	85
Stacey	90
Betsy	95
Mean of sum score	86.79 (8.75)

Items are rated on a 5-point scale (1 = *strongly disagree* to 5 = *strongly agree*); users rated frequency of use, ease of use, and extent to which a system is functionally consistent reflect overall positive ratings related to critical usability requirements. The Overall Reaction to the.

Software Scale includes six items rated on a 10-point scale (ranging from 0 to 9) with anchored endpoints (e.g., Terrible-Wonderful) moving from negative to positive. Overall, user ratings were positively skewed on this scale (QUIS ratings ranged from 4 to 9) with no users rating these six items lower than four. Generally, scores clustered in the 5–9 range suggesting average to high ratings for overall usability. Rigid-flexible was rated lower relative to the other items, including Suzanna, who rated Rigid-Flexible as low as a 4 on the 10-point scale. Janna and Betsy rated Rigid-Flexible a 5, Rachel and Stacey rated a 6, Diane rated this a 7 and Rania rated this an 8. Dull-Stimulating was rated a 7 by Rachel and Suzanna, an 8 by Janna and Dianne, and a 9 by Rania and Betsy. For the remaining four items on the Overall Reaction to the Software Scale (Terrible-Wonderful, Difficult-Easy, Frustrating-Satisfying, Inadequate-Adequate) user scores all ranged from 5 to 9. For the remaining four items on the Overall Reaction to the Software Scale (Terrible-Wonderful, Difficult-Easy, Frustrating-Satisfying, Inadequate-Adequate) user scores all ranged from 5 to 9.

IVT-T usability problems

Seventy-two usability problems extracted from merged field notes during the CTA were identified by users. Table 3 reports the percentage of problems identified using the four remaining QUIS subscales.

Merged field notes from the CTA indicated that the fewest percentage of usability problems (3% or 2/72) were related to System Capabilities (e.g., system loading too slowly, replay button not working), and the greatest percentage of usability problems (36% or 26/72) were related to Learning. For example, CTA data indicated some users had difficulty moving through the empty classroom using the keyboard, particularly when navigating through the optional *Tour the Classroom* feature. CTA data indicated user confusion when navigating through the reflection and the feedback interface. For example, users requested the option to "go back" and replay the previous segment before selecting a decision to reflect upon. CTA data indicated that 19% (14/72) of usability problems were related to Screen Design and Layout (e.g., location of user interface widget to select reflection question was unclear, font size was too small, decision points not rendered correctly). Seventeen percent (12/72) of usability problems were related to the Terminology and System Information provided by domain experts, including inconsistencies in terminology and lack of confirmation around submitting information. Given users were able to interact with the system without major difficulties, evaluation sessions allowed for in depth analysis and valuable feedback on graphics and content described next.

Authenticity and realism of the characters and classrooms

Authenticity and realism of the characters and classrooms was operationalized in the codebook as user comments regarding the visual appearance of the characters and classrooms and the quality of the graphics with subcodes indicating positive impressions, negative impressions, and recommended changes. Audio components connected to the characters and classroom were coded here as well. Table 4 illustrates the number and percent of segmented text falling into each subcode to illustrate the representativeness of those comments within the subcodes (Miles and Huberman 2019).

Table 3 Percent of usability problems observed during CTA categorized by QUIS subscales

QUIS subscale	Percent of usability problems (%)	Number of usability problems	Example observations
Learning subscale	36	26	Example: Difficulty learning the controls to move around the classroom Example: Forgetting what happened before, during, and after response options were selected made reflection difficult
Other design issues	25	18	Example: Inadequate background information provided for the character Example: Lack of information about teacher and student tone (e.g., encouraging, stern, sarcastic) made interpretation of responses challenging
Screen design and layout subscale	19	14	Example: Location of user interface widget to select reflection question unclear Example: Font size on the confirmation page too small
Terminology and system information subscale	17	12	Example: Decision point selected and not selected was unclear Example: The label "submit" for reflection was not the appropriate terminology and created confusion about who would receive user reflections
System capabilities subscale	3	2	Example: Classroom scene took too long to load Example: Proceed to replay button did not consistently work
Total usability problems identified	100	72	

Usability problems observed when the seven users performed tasks during the user-based evaluations were documented by the research team via field notes and merged by the usability engineer. The usability engineer coded each usability problem into one of the four QUIS subscales (i.e., Learning, Screen Design and Layout, Terminology and Systems Information, System Capabilities or Other Design Issues) to identify the percent and number of usability problems identified by QUIS subscales (72 total were identified)

Table 4 Percentage and number of segmented text coded for each subcode

Subcode	Percentage	Number of segments
Authenticity/realism—characters and classrooms		
Positive impression—characters and classrooms	54	30
Negative impression—characters and classrooms	23	13
Recommended changes—characters and classroom	23	13
Total (authenticity/realism—characters and classrooms)	100	56
Authenticity/realism—storylines		
Positive impression—storylines	57	21
Negative Impression—Storylines	16	6
Recommended changes—storylines	27	10
Total (authenticity/realism—storylines)	100	37
Instructional design quality		
Positive impression—instructional design	29	50
Negative impression—instructional design	5	11
Recommended changes—instructional design	66	113
Total (instructional design quality)	100	174

Overall, more segments were coded into positive impressions (30 or 56%) than negative impressions of the characters and classrooms (13 or 23%). The following sections present results from the cross-analysis of semi-structured interviews to reflect patterns converging across users with direct quotes illustrating feedback.

Positive impressions of characters and classrooms

Two patterns of positive impressions of the characters and classrooms emerged through the cross-analysis: (1) age/developmental level accurately depicted, and (2) character voice enhanced realism. The first pattern, **age/developmental level depicted accurately**, focused on comments made regarding the characters appearing visually consistent with their chronological age and the classrooms appearing realistic for the particular age group. Janna shared: "*You portrayed exactly what I thought kids would look like*" while Stacey indicated: "*They [the characters] looked real. I thought the classroom itself was represented well with the student population*." Comments also focused on the physical appearance of the classroom, including desk arrangement and bulletin boards featured, for example, Janna explained, *"[I] like the set-up with the desks was good. I paid important attention to that like [sic] what was on the walls …The sixth grade one seemed pretty fine.*" Betsy also noted, *"I thought that the classroom seemed very realistic with the decorations and the set up of the rooms, and everything like that…and when I was up close or I was working in the front of the classroom, it did look very realistic to me."* One user also noted that they could discriminate between the first and sixth grade classroom based on the visual appearance, for example, Dianne shared, "*When we talked about like the differences between the sixth-grade classroom and the first-grade classroom, I think you could definitely tell. The sixth-grade classroom had things that would be [in] a sixth-grade classroom*."

The second pattern, **character voice enhanced realism**, focused on user comments regarding the value of hearing the character speak and how that made the simulation feel more realistic. These comments included what the characters said combined with their tone of voice. Janna stated: "*Yeah, he did absolutely [sound like a sixth grader*], Rania shared, "*You can tell in his voice that he was pretty sassy and that's good because they [sixth graders] are sassy*" while Betsy noted, "*Yes, that was actually one of my reactions, when Jordan had an attitude about something, and I was like 'oh my gosh, he has a real attitude' and I think that makes it more realistic…[hearing his voice] made me actually internalize 'well if a kid is going to say it in that tone, how would I actually react rather than just seeing the words up on the screen' and so, it made it more real."*

Negative impressions of characters and classrooms

Negative impressions of the characters and classrooms clustered into two patterns: (1) Location/setting difficult to place, and (2) More visual cues and detail needed. The first pattern, **location/setting difficult to place**, focused on confusion regarding the location and setting of the virtual classroom within a larger context and community and concerns regarding lack of realism of the classroom. Several users noticed that racial and ethnic diversity was represented in the characters. For example, Dianne shared, "*Yeah, it was a multiracial classroom*" while Rania indicated, "*I definitely noticed a lot of different skin colors."* However, the character uniforms and appearance of the classroom was visually inconsistent with some users' experiences of high poverty public schools. Some users indicated that the uniforms suggested a wealthy, private school. Rachel, for example, asked: *"Everyone was wearing what seemed to be similar clothing and I didn't know if that was like ease of constructing the system or if you wanted to model a private school?"* Betsy further shared: *"It almost looked private because they were all wearing the same color shirt. Yes, I did notice that. So, I would assume that it would be a private school just from the dress of the students."* When Rania was asked to identify the geographic location of the classroom, she responded "*Affluent…the uniforms, the cleanliness, all the organization*." When the interviewer asked what would be needed to make the classroom appear more urban, Stacey replied: *"… the classroom looks nice… in the urban schools I have seen, there are more blank walls and bare desks*." Rachel shared that the classroom was unrealistic looking because it appeared artificially clean and new: *"It seemed to be really clean, and nothing was broken, the windows were all solid panes of glass, they weren't repaired, the ceiling didn't have wet spots, the floor didn't have stains. And that's not necessarily an indication of high poverty, it's just an indication of the space has been used…all the posters look new and so, it just looked like a virtual context as opposed to a realistic context."*

The second pattern, **more visual cues and detail needed**, suggested that some users needed more meaningful visual detail or a closer personal view of the classroom or character to interpret the classroom environment and character responses. Several comments related to the need for more visual details. Betsy, for example, stated: "*facial expressions of characters would improve IVT*" and *"Include more visual detail."* Janna explained how a *zoom in* feature that provides a detailed view of an object or action would be helpful: "*I wish I could've like been able to have a closer view in the classroom. Like of what they were actually doing at their desk. There could be cool zoom in feature and you could see what the worksheet that they were doing. You know I said that I liked the hand on desk attentive position. I'm huge fan of that because kids like to bang in their desks but like you don't know if they're like banging on their desk causing chaos in the classroom or if their like hands were actually on their desk to do work."* In this particular case, the zoom feature was considered critical to helping this user interpret whether the students were on task and the type of work they were completing.

Authenticity and realism of the storylines

Authenticity and realism of the storylines was operationalized as the degree to which the storylines, behaviors and actions of the characters, and teacher response options were logical and realistic. Table 3 illustrates that more segments were coded into positive impressions (21 or 57%) than negative impressions of the storylines (6 or 16%).

Positive impressions of the storylines

The cross-analysis identified one consistent positive impression of the storylines: **realistic situations**. Specifically, feedback suggested that the scenarios and behaviors of Jordan and the non-disruptive characters in the background were generally realistic. For example, Janna explained: "*It was very realistic in my opinion. That's very pre-teen of him like to be you know like trying to be difficult almost like the things he was saying were like just being difficult, trying to distract the teacher and to get attention. At the same time all these things were very encompassing of a sixth-grade personality."* Rachel commented that the opening scene in which Jordan arrives late to class was true-to-life, "*It's like, well, I'm already late, what, how much worse can it be? … yeaaah it felt familiar. Both as a peer in the classroom having been, you know, in class with Jordan, and having dealt with students… so that's where the reactions themselves seemed really real*." Dianne shared, "*I guess. I liked it because it kind of gave you like a real-life situation that you probably would encounter as a teacher. But, it let you actually practice it…I feel like it was pretty realistic I guess because, it does seem like a lot of situations you encounter…I thought it kind of did a good job at making it kind of feel real."* Suzanna commented in response to the interviewer asking for feedback on the storylines, *"Yes, behavioral problems, definitely, and the slamming of the books."*

Negative impressions of the storylines

Negative impressions of the storylines clustered into two patterns: (1) only one disruptive student featured in the scenario, and (2) selecting one response option was difficult. First, having **only one disruptive student featured in the scenario** at a time was deemed unrealistic. Feedback suggested illustrating one teacher-student interaction does not accurately reflect the complexities of classrooms, how disruptive behavior spreads, and how other students in the classroom can become distracted by disruptive behavior. Rania suggested that the system should do a better job featuring how other students in the classroom respond to disruptions: *"As a teacher, I would not want to see myself, but almost be able to see the classroom easier rather than maybe just seeing Jordan's perspective… if you see another child's reaction to what is happening with me and Jordan."* Stacey shared: "*I think it would be important to include how the class would respond…*" and then went on to explain, *"Disruption should spread to other students…most importantly, it would be beneficial to have one scenario where one or two other children start to be disruptive because it is more difficult to handle two people who are really disruptive versus one child who is disruptive*." Suzanna also explained, "*how one student reacts is going to affect the entire classroom. I* *think the bigger issue is the class, overall, and how they are going to respond. That is much more difficult to respond to than a one-on-one interaction with the child."*

The second pattern, **selecting one response option was difficult**, suggested that some users found the forced response option format to be restrictive. Betsy indicated that the available response options did not represent how they would respond to disruptive behavior "*I am definitely more of the empathic sort of teacher. … some of the choices were more emotional. I would have approached him and told him what to do.*" Rania shared that she wanted the option to generate her response to the character rather than the system generating it on her behalf: *"Some teachers would say "I would do that, definitely' some teachers would say, 'Hmm, I would not do any of these things' and then write in what they would do."* Stacey noted that she wanted to combine across the available responses, "*The hardest part, for me, was wanting to combine two answer choices*." Other users indicated that the fixed response options forced them to avoid ineffective responses and select the obvious effective responses to perform well in the system. Suzanna, for example, explained, "*I wouldn't want to [select the bad option] …because I'd want to pass. I would probably be focused on success*." Other users highlighted that forced decision points left them curious as to how a character would respond to the ineffective responses. For example, Dianne noted, *"I was really curious- I did want know what would happen [if I selected the worst option]."* Betsy also noted her interest in knowing what would have happened if she picked the ineffective option, "*How would've the kid reacted. What would have gone down?"*.

Instructional design quality

Instructional design quality was operationalized as user impressions regarding practice, reflection, and feedback and beliefs regarding translation of learning from the virtual to the live classroom. Table 3 illustrates that most segments related to instructional design were coded into recommended changes (66%), followed by positive impressions (29%) and fewest coded into negative impressions (5%).

Positive impressions related to instructional design

Positive impressions related to instructional design clustered into two patterns: (1) the practice/playing phase was entertaining and engaging, and (2) feedback promoted user knowledge of why their decision was effective and how to improve their performance. The first pattern suggested that some users found **practice/playing phase entertaining and engaging** because their decisions influenced the progression of the storyline and allowed them to view character responses and how the scenario unfolded. Janna stated: "*I enjoyed making the decisions/choices"* Rania noted: "*The first [part], going into the classroom…cause it's fun"* and Suzanne shared: *"The most fun is like being able to go through it virtually with him and like choosing which option you would choose …and seeing what he was going to say…"* Suzanne further explained: *"I thought that it was most helpful going through and seeing the different options for like what I could possibly say … I thought it was helpful to see … what he was saying and having that opportunity to … click next to see if what he was saying was puzzling or … took me aback or like if I could possibly predict what was going to happen next."*

The second pattern, **feedback promoted user knowledge of why their decision was effective and how to improve their performance** focused on the unique contributions of qualitative and quantitative feedback, how IVT-T feedback contrasted with current feedback systems in place in schools, and how the combination of positive and constructive feedback supported learning by affirming good choices and noting opportunities for improvement. Examples of quotes reflecting the value of qualitative feedback included Rachel, who noted: "*Everyone ends up using points and rubrics and scoring things to show progression, but the qualitative information, like, you choose the right answer, and we're not just telling you it's the right answer, here's why. I think that's a really important component of what you're doing*…. Rachel went on to explain: "*Short, immediate feedback loop especially because sometimes it can be really hard to get, you know, someone to come in and observe your class! And if they do, they might be using a frickin' rubric with numbers…"* Suzanne explained, *"It [qualitative feedback] taught me a lot about the situation about what to do and what not to do but it just affirmed, which was nice. As a teacher you want to hear affirmation on stuff that you are doing well and stuff that you are not doing well. You want to know why or what could have been better."* Betsy explained how the combination of qualitative and quantitative feedback that included positive and constructive feedback supported her learning: *"I liked the words…but then for the numbers, I thought the numbers were encouraging having the points and you get the points when you give the right answers…and I referred to the words for that feedback.* Betsy also noted, "*Seeing the choices that I made and the feedback gave positive feedback as well as negative feedback, which I thought was very helpful….because I felt I did make the right choice in my situations so the positive feedback reinforced that choice and my confidence in my choice…but then the negative feedback was more 'this is how you can improve,' and so I actually learned from it…rather than just getting positive feedback for making the right choice, or only negative feedback for making the wrong choice. So, I really really liked the feedback section*." Janna similarly explained how positive feedback was affirming: "*The feedback is extremely important because that's like you know while I was going through the program like at the two different reflection points I was like I had the thoughts that they gave me in the feedback like I had the thoughts when I was choosing the options that they gave me in the feedback so I was on the right track. So, it was nice seeing that at the end like reaffirming. Like 'Oh you were you had the right frame of thought.'*".

Negative impressions related to instructional design

Negative impressions clustered into two patterns: (1) reflection was confusing, and (2) quantitative feedback was unclear. Users made several positive comments related to the importance of reflection and its connection to learning. For example, Rachel shared: "*reflection provokes deeper thinking"* while Rania noted: "*as a teacher I would look back on that [reflection]…and if I'm having a huge issue with a student I could look back and say this is what I did with Jordan I could do that with another student*"). However, the cross-case analysis indicated that **reflection was confusing** with regards to what users were supposed to do and that the placement of text added to their confusion. Several participants noted confusion regarding the reflection interface, including Suzanna who stated, "*Yeah it seemed confusing [reflection]* and Rachel who further noted, *"Sometimes the screen seemed a little busy especially, 'here's all the things you're going to be asked to reflect on here for the instructions.*'" Users also found it difficult to have to recall the specifics of the decision point and how the character responded in order to complete the reflection. Stacey shared, "*Having decision point text reminds you what you picked but not how the scene played out*" and "*Difficult not to see/reflect upon alternate scenarios*" while Rachel noted, "*You only have your decision tree… because you don't necessarily have the visual of how the character responded…it's hard to respond [reflect], like, if you can't remember how they responded.*" Stacey went on to clarify, *"I would have altered which ones [reflection questions] I picked if I had a visual of how Jordan reacted."*

The second pattern, **quantitative feedback was unclear**, suggested that users found the quantitative feedback (e.g., effective = +1 and not effective = 0) confusing in terms of how they earned points, what the points meant, and for whom the points were intended. Janna noted: "*I mean there was a zero, one, and one next to the options so like were the two equally as constructive like the one and one…were they equally constructive… Or were they just better than the zero?"* Stacey shared*, "I felt indifferent about [the points]. I would like it if I could see why each one had the point they did so I can see why I got one point for this answer or why I got zero points for another answer."* Rachel did not agree with the point value assigned: *"[the zero is confusing] I think a negative score would be appropriate in a lot of places, especially with the sarcastic choices… because bad options are not necessarily neutral options.*" Finally, Rania indicated confusion regarding who earned the points as she thought the points were for character not the teacher: "*I thought that was for Jordan, because of the gold star. And I thought why is he getting points?"*

Discussion

Virtual learning environments represent a unique opportunity to support early career teachers to improve their behavior management skills. Leveraging instructional technologies such as IVT-T can also provide broad access to professional development for school districts interested in scaling up these supports (Hew and Brush 2007; Xie et al. 2017). The goal of this study was to assess usability early in the development lifecycle guided by principles of user-centered design to anticipate and address the needs of early career teachers (Antonenko et al. 2017; Gabbard et al. 1999; Hix and Hartson 1993). Findings suggest complimentary and distinct patterns in user experience.

Global numeric usability ratings based on the QUIS (Chin et al. 1988; Harper and Norman 1993) and SUS (Bangor et al. 2008) suggest IVT-T was adequately usable while the concurrent think aloud protocol and semi-structured interviews revealed design shortcomings that impeded effective user performance and informed ways to improve the interface. Similar findings have emerged in usability evaluations, including a recent formative evaluation of an augmented reality science game, with quantitative results indicating positive ratings for specific design features while interviews revealed critical usability issues regarding those same design features that would have gone unnoticed (Laine, Nygren et al. 2016). Comparable findings also emerged in usability evaluation of a multi-media system for distance education (Parlangeli et al. 1999). Earlier examination of the initial IVT-T prototype also indicated that numeric ratings of crude prototypes from advisory board members were generally high, while open-ended responses yielded constructive feedback to improve the technology (Shernoff et al. 2016, 2018).

Revisions made based on usability problems identified during the CTA

Observing representative users perform benchmark tasks with IVT-T indicated several avenues for improving the interface. The usability engineer's coding of usability issues identified during the CTA into QUIS subscales (i.e., Screen Design and Layout, Terminology and Systems Information, Learning, and System Capabilities) informed subsequent design modifications. Given users' location in the classroom is controlled by the software and usability problems related to navigating in the classroom only emerged during the optional stand-alone *Tour the Classroom* feature, the design team did not revise navigation features in subsequent prototypes. However, given participants had limited experience playing videogames and we expect future users will also need explicit instructions and didactic information regarding how to use the system, we added an online user guide, FAQ, and online support page for users to obtain technical support when they had questions or encountered problems learning how to use the system. Level 1 (untimed and unscored) also provided users with additional practice learning how to navigate the system without worrying about points earned.

Results also pointed to the need for more feedback to users as they performed tasks to optimize their learning (i.e., 17% of usability problems coded as Terminology and System Information). Therefore, we created cleaner segues between practice, reflection, and feedback using visual fade in. Design changes also included visually highlighting users selected choices and providing explicit feedback to users when they practiced with the characters. We also embedded more user choice related to completing tasks and interacting with the system (e.g., enabling keystrokes including numbers and arrows for choosing a decision point, allowing users to select ENTER to finalize their choices, allowing users to complete all reflections using the keyboard, using arrows to move up and down the list of options for responding to the character).

Very few problems (3%) were related to System Capabilities, and specifically that the system was loading too slowly. The IVT-Prototype was created in Axure to render a 2D format, however, the long-term goal is to use the unity game engine to enable state-of-the-art, web browser-independent, real-time rendering of 3D scenes and animated avatars with integrated audio. Given the expected variability in graphics cards, operating systems, and internet speeds available to teachers, we added visual cues throughout the system to make users aware that the system may take time to load. For instance, we replaced the gray screen with, "*Please wait while classroom is loading"* and plan to conduct additional user studies to assess the robustness of the system when accessed with a variety of graphics cards, operating systems, and internet speeds.

Revisions made to enhance realism of the characters, classrooms, and storylines

Results from the cross-analysis suggested that the characters were visually consistent with their chronological age and the classrooms appeared realistic for that age/developmental level. In addition, some feedback suggested that hearing the characters speak made the simulation feel more realistic. This feedback was encouraging given prior research indicates that the visual elements of simulations and serious games, and particularly sophistication and detail of graphics, predicts immersion (McLaughlin et al. 2010). Given the task of developing realistic 3D avatars is labor intensive, artistically complex, and repetitive, the development of initial avatars deemed as realistic is a contribution on its own that could be of interest to the virtual intelligent agent community. Results from this study also informed the software team in terms of replicating their approach to future graphics work.

We conceptualized a visually compelling interface as necessary but insufficient if usability problems prevent users from learning how to interact effectively and comfortably with the system or acquiring knowledge and skills related to behavior management. Several user concerns emerging from the interviews are worth noting. Users reported that the classroom did not resemble their notion of high poverty classrooms—instead they were perceived as well-resourced and artificially clean and polished. Character uniforms exacerbated this issue as some users associated the uniforms with private, affluent schools, making geographic locale difficult to discern. Although there is geographic variability regarding uniforms in public schools in the US, more than 50% of urban, high poverty school districts have a uniform policy (Brunsma 2005) and thus the team concurred that uniforms should remain. However, we added *Learn About Your School* which described the geographical and community context and students served. Given the labor-intensive nature of graphics, the team concurred that changes to the classroom to graphically portray the realities of under-resourced schools would be made with time and resources permitting.

Users also highlighted the importance of providing more contextual information to enhance their understanding of the system and students populated within it. Some users noted that the storylines only depicted one disruptive student at a time which did not match their experience of managing multiple students with challenging behaviors and watching disruptions spread. Indeed, studies document the negative impact of having aggressive peers in a classroom in terms of modeling negative behavior and promoting aggressive peer norms (Thomas et al. 2006). Therefore, storylines were adapted to more fully illustrate the spread of disruptive behaviors to other students when they were not addressed quickly. Auditory distractions (e.g., intercom interruptions, traffic outside) were added to reflect the frequent interruptions and high stress that characterizes live instruction. Contextual cues (e.g., enhanced visual detail regarding student work and character and teacher tone) were added to help visually communicate learning, behavior, and emotions.

Despite users finding the pre-scripted response options restrictive, the research team agreed that the planned prototype would include fixed response options given subsequent programming depended upon this. Given recent advancements in embodied conversational avatars for learning, future work may include moving beyond pre-scripted entities to fully autonomous agents who engage in conversations with the teacher and express their intentions.

Relatedly, users were motivated to perform well and to earn points and thus avoided obvious ineffective response options, which limited their exposure to storylines in which the characters escalated. These challenging interactions between teachers and students were conceptualized as important to teacher learning and thus we added *Meet your Colleagues* to the interface, which allowed users to watch (in third person) virtual colleagues struggle with behavior management. This addition allowed users to view a greater range of storylines, including those with negative endings, without impacting their score or ability to level up. Level 1 (untimed and unscored) also provided opportunities for users to select ineffective choices that would not count against their score.

Revisions made to enhance instructional design of IVT-T

Users identified instructional design problems most notably related to reflection and feedback. Although several users appreciated the value of reflection, others were confused by what they were supposed to do, found the visual design confusing, and indicated that the requirement of having to recall the specifics of the storyline in order to complete the reflection was difficult. Although users agreed that reflection needed revising, they did not agree on optimal redesign features (e.g., ideal number of reflection questions, type of reflection questions). User feedback still provided ample opportunities to redesign the reflection training phase. This included placing all reflection questions on one screen rather than using a widget to switch between questions. We also provided more choice around which decision points to reflect on and made user choices available during reflection accompanied by a visual representation of how the student responded.

Confusion regarding how points were earned was addressed in several ways. First, we revised the user dashboard to provide more detailed feedback about earning points and leveling up. Second, we planned for an in-person orientation during subsequent funding years to invite questions and clarify important pedagogical elements early. Third, we created a detailed IVT-T User Guide which provided visual and textual information specific to navigating practice, reflection, and feedback, in addition to criteria for leveling up, scoring rules, and scoring feedback. Finally, the feedback image was changed from a star to a trophy which users noted might enhance clarity regarding who was earning the points. These findings, taken together, point to the value and efficiency of investing time early in obtaining user feedback and making refinements before the labor-intensive process of programming commences.

Limitations

Several important limitations are worth noting. Our focus on usability and instructional design and our goal of synthesizing common patterns across users diluted idiographic needs and individual user preferences. We focused on summarizing common issues and ignored idiographic concerns that could impact user experience with the system. In addition, we relied on a relatively small and homogenous sample of educators with limited knowledge of and experience with virtual interfaces. Although sampling procedures were based on established guidelines by Virzi (1992) regarding optimal number of participants to identify usability problems, it is not clear the extent to which similar usability problems and patterns may have emerged with a larger sample of teacher educators, perhaps with more experience with videogames and different standards related to usability. Future studies with a larger, more diverse sample would improve the generalizability of findings from the current work and ensure more teacher voices are represented.

Implications

As the field of educational technology embraces more advanced technology, as an example but not limited to simulation training in behavior management, the relative effectiveness of these training programs will be increasingly dependent upon how well educators can use these systems. Therefore, system usability should be given high priority in the development and evaluation process, with the goal of implementing efficient and effective ways to facilitate usability engineering in the field of educational technology. Maximizing IVT-T usability can facilitate teacher learning by reducing usability distractions that prevent teachers from meeting the IVT-T instructional objectives and can also reduce the amount of time and effort that school districts must allocate to test use and training (Hartson and Pyla 2012; Varier et al. 2017; Verdú et al. 2017).

Findings have important implications for researchers and teacher educators involved in designing and evaluating instructional technologies, particularly in the early phases of development and refinement. Findings from the current study indicate the value of combining a ground up, user-driven, qualitative evaluation of user experience with IVT-T (i.e., observations of user performance and semi-structured interviews) with a top-down, quantitative approach that relied on established usability heuristics and guidelines (i.e., established usability measures to assess global usability of IVT-T). From a pragmatic perspective, qualitative feedback was instrumental to the redesign of IVT-T and enhanced our understanding of how and why users liked the content and function of different design components related to practice, reflection, and feedback. To this end, findings illustrate how standardized qualitative research methods can provide development teams with deeper insight into why users like or dislike certain elements, and specific design changes to enhance usability. The combined expertise of end users and usability experts is also critical as it allows teachers to be involved in planning and adopting educational technologies that meet specific usability principles and design guidelines.

Positive user experience and satisfactory usability play a critical role in the acceptance, satisfaction, and use of educational technology training systems such as IVT-T (Harrati et al. 2016). The systematic evaluation of how preservice teachers interacted with this system, including early indicators that the system is appealing, even among a small sample of non-gamers, suggests the potential value of including simulation in teacher training for behavior management. Findings further highlight the important role that technological advances can play in teacher education, without losing sight of users and the human issues involved in these models. Moving forward, it will also be important for the development team to continue conducing usability studies as a means of further optimizing the system and ensuring that such technologies are sustainable within the constraints and resources available in school districts.

Conclusions

Despite the limited sample size, our first usability evaluation of IVT-T provided us with promising results regarding important usability requirements (i.e., ease of use, functional consistency, system capabilities) based on numeric ratings provided by users. Data from the concurrent think aloud protocol and semi-structured interviews revealed IVT-T design shortcomings that impeded effective user performance and informed ways to improve the interface. Findings from the current study highlighted the value of combining qualitative evaluations of user experience with quantitative approaches which has implications for researchers and designers of virtual training for teachers. Future research in educational technology development would benefit from continued reliance on mixed methods and leveraging the strengths of quantitative and qualitative research designs when evaluating virtual platforms.

Findings from the current study also point to the system's ability to convey believable, visually appealing, realistic classroom scenarios including disruptive behaviors conveyed by virtual characters. More contextual cues, multiple challenging behaviors featured at the same time, and changes to the visual appearance of the classroom would enhance realism. Although innovations in the area of VR (e.g., facial animations synchronized with text-to-speech, AI avatars who are autonomous and embodied) point to possibilities for enhanced technological innovations related to IVT-T, those advancements would require further evaluation of their incremental benefits to teacher learning and transfer and ensuring usability and instructional design features remain strong.

Data from the concurrent think aloud protocol and semi-structured interviews also pointed to critical instructional design improvements related to the reflection and feedback interface necessary to successfully integrate this virtual training into teacher education models. Future work will continue to integrate usability evaluations with assessment of crucial instructional design components of IVT-T to further optimize the system and increase the likelihood it will benefit educators in need of support in behavior management.

Funding The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A150166 to the Rutgers, the State University of New Jersey. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Compliance with ethical standards

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

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Alessi, S. M., & Trollip, S. R. (2001). *Multimedia for learning: Methods and development*. Boston, MA: Allyn & Bacon.

Annetta, L., Lamb, R., Minogue, J., Folta, E., Holmes, S., Vallett, D., et al. (2014). Safe science classrooms: Teacher training through serious educational games. *Information Sciences, 264*, 61–74.

Antonenko, P. D., Dawson, K., & Sahay, S. (2017). A framework for aligning needs, abilities and affordances to inform design and practice of educational technologies. *British Journal of Educational Technology, 48*(4), 916–927.

Aslan, S., & Reigeluth, C. M. (2016). Investigating "the coolest school in America:" How technology is used in a learner-centered school. *Educational Technology Research and Development, 64*(6), 1107–1133.

Atkins, M. S., Shernoff, E. S., Frazier, S. L., Schoenwald, S. K., Cappella, E., Marinez-Lora, A., et al. (2015). Re-designing community mental health services for urban children: Supporting schooling to promote mental health. *Journal of Consulting and Clinical Psychology, 83*(5), 839–852.

Atteberry, A., Loeb, S., & Wyckoff, J. (2017). Teacher churning: Reassignment rates and implications for student achievement. *Educational Evaluation and Policy Analysis, 39*(1), 3–30.

Badiee, F., & Kaufman, D. (2015). Design evaluation of a simulation for teacher education. *Sage Open, 5*(2), 1–10. https://doi.org/10.1177/2158244015592454.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction, 24*(6), 574–594.

Bellotti, F., Berta, R., & De Gloria, A. (2010). Designing effective serious games: Opportunities and challenges for research. *International Journal of Emerging Technologies in Learning, 5*(SI3), 22–35.

Beutler, L. E., & Harwood, T. M. (2004). Virtual reality in psychotherapy training. *Journal of Clinical Psychology, 60*(3), 317–330.

Bowman, D. A., Gabbard, J. L., & Hix, D. (2002). A survey of usability evaluation in virtual environments: classification and comparison of methods. *Presence: Teleoperators & Virtual Environments, 11*(4), 404–424.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry, 189*(194), 4–7.

Brunsma, D. L. (2005). *Uniforms in public schools: A decade of research and debate*. Lanham, MD: Rowman & Littlefield Publishing Group.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 213–218). ACM.

Cooke, L. (2010). Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication, 53*(3), 202–215.

Dawson, M. R., & Lignugaris-Kraft, B. (2017). Meaningful practice: Generalizing foundation teaching skills from TLE TeachLivE™ to the classroom. *Teacher Education and Special Education, 40*(1), 26–50.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*(5910), 66–69. https://doi.org/10.1126/science.1167311.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*(2), 81–112.

Dieker, L., Hynes, M., Stapleton, C., & Hughes, C. (2007). Virtual classrooms: Star simulator. *New Learning Technology SALT, 4*, 1–22.

Dieker, L. A., Hughes, C. E., Hynes, M. C., & Straub, C. (2017). Using simulated virtual environments to improve teacher performance. *Journal of the National Association for Professional Development Schools, 10*(3), 62–81.

Drisko, J. W., & Maschi, T. (2015). *Content analysis: Pocket guides to social work research methods*. Oxford, NY: Oxford University Press.

Epstein, J. L., Sanders, M. G., Sheldon, S. B., Simon, B. S., Salinas, K. C., Jansorn, N. R., … & Williams, K. J. (2008). *School, Family, and Community Partnerships: Youth Handbook for Action* (3rd ed.). Thousand Oaks, CA: Corwin Press.

Evertson, C. M., & Weinstein, C. S. (Eds.). (2006). *Handbook of classroom management: Research, practice, and contemporary issues*. Mahwah, NJ: Lawrence Erlbaum.

Finstad, K. (2006). The system usability scale and non-native English speakers. *Journal of Usability Studies, 1*(4), 185–188.

Fletcher, J. D. (2009). The value of expertise and expert performance: A review of evidence from the military. In K. A. Ericsson (Ed.), *Development of Professional Expertise* (pp. 449–469). NY: Cambridge University Press.

Fonteyn, M. E., Vettese, M., Lancaster, D. R., & Bauer-Wu, S. (2008). Developing a codebook to guide content analysis of expressive writing transcripts. *Applied Nursing Research, 21*(3), 165–168.

Gabbard, J. L., Hix, D., & Swan, J. E. (1999). User-centered design and evaluation of virtual environments. *IEEE Computer Graphics and Applications, 19*(6), 51–59.

Gabbard, J. L., & Swan, J. E., II. (2008). Usability engineering for augmented reality: Employing user-based studies to inform design. *IEEE Transactions on Visualization and Computer Graphics, 14*(3), 513–525.

Graafland, M., Schraagen, J. M. C., Boermeester, M. A., Bemelman, W. A., & Schijven, M. P. (2015). Training situational awareness to reduce surgical errors in the operating room. *British Journal of Surgery, 102*(1), 16–23.

Gregory, S., & James, R. (2011). VirtualPREX: Open and Distance Learning for pre-service teachers. In *Expanding Horizons-New Approaches to Open and Distance Learning. Presented at the 24th ICDE World Conference on Open & Distance Learning, Bali*.

Gresalfi, M. S., & Barnes, J. (2016). Designing feedback in an immersive videogame: supporting student mathematical engagement. *Educational Technology Research and Development, 64*(1), 65–86.

Guarino, C. M., Santibañez, L., & Daley, G. A. (2006). Teacher recruitment and retention: A review of the recent empirical literature. *Review of Educational Research, 76*, 173–208.

Harper, B. D. & Norman, K. L. (1993). Improving User Satisfaction: The Questionnaire for User Interaction Satisfaction Version 5.5. In: *Proceedings of the 1st Annual Mid-Atlantic Human Factors Conference* (pp. 224–228), Virginia Beach, VA.

Harrati, N., Bouchrika, I., Tari, A., & Ladjailia, A. (2016). Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis. *Computers in Human Behavior, 61*, 463–471.

Hartson, R., & Pyla, P. S. (2012). *The UX book: Process and guidelines for ensuring a quality user experience*. Waltham, MA: Elsevier.

Henry, D., Guerra, N., Huesmann, R., Tolan, P., VanAcker, R., & Eron, L. (2000). Normative influences on aggression in urban elementary school classrooms. *American Journal of Community Psychology, 28*(1), 59–81. https://doi.org/10.1023/A:1005142429725.

Hew, K. F., & Brush, T. (2007). Integrating technology into K-12 teaching and learning: Current knowledge gaps and recommendations for future research. *Educational Technology Research and Development, 55*(3), 223–252.

Hix, D., & Hartson, R. (1993). *Developing user interfaces: Ensuring usability through product and process*. New York, NY: Wiley & Sons.

Horswill, I., & Lisetti, C. L. (2011). On the simulation of human frailty. In *BICA* (pp. 146–150).

Ingersoll, R. M., & Strong, M. (2011). The impact of induction and mentoring programs for beginning teachers: A critical review of the research. *Review of Educational Research, 81*, 201–233.

Jaspers, M. W. (2009). A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. *International Journal of Medical Informatics, 78*(5), 340–353.

Kazdin, A. E. (2005). *Parent management training*. New York: Oxford University Press.

Kolb, D. A., Boyatzis, R. E., & Mainemelis, C. (2000). Experiential learning theory: Previous research and new directions. In R. J. Sternberg & L. E. Zhang (Eds.), *Perspectives on cognitive, learning, and thinking styles* (pp. 227–247). Mahwah, NJ: Lawrence Erlbaum.

Laine, T. H., Nygren, E., Dirin, A., & Suk, H. J. (2016). Science spots AR: A platform for science learning games with augmented reality. *Educational Technology Research and Development, 64*(3), 507–531.

Lieber, E. (2009). Mixing qualitative and quantitative methods: Insights into design and analysis issues. *Journal of Ethnographic & Qualitative Research, 3*(4), 218–227.

Lindsey, L., & Berger, N. (2009). Experiential approach to instruction. In C. M. Reigeluth & A. A. Carr-Chellman (Eds.), *Instructional-design theories and models* (Vol. 3, pp. 117–142). New York, NY: Routledge.

Ludwick, D. A., & Doucette, J. (2009). Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries. *International Journal of Medical Informatics, 78*(1), 22–31.

Lugrin, J. L., Latoschik, M. E., Habel, M., Roth, D., Seufert, C., & Grafe, S. (2016). Breaking bad behaviors: A new tool for learning classroom management using virtual reality. *Frontiers in ICT, 3*, 1–26.

Mayhew, D. J. (1999). *The usability engineering lifecycle: A practitioner's handbook for user interface design*. San Francisco, CA: Morgan Kaufman.

McLaughlin, T., Smith, D., & Brown, I. A. (2010, June). A framework for evidence based visual style development for serious games. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games* (pp. 132–138).

McLellan, E., MacQueen, K. M., & Neidig, J. L. (2003). Beyond the qualitative interview: Data preparation and transcription. *Field Methods, 15*(1), 63–84.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Sage Publications, Inc.

Miles, M. B., & Huberman, A. M. (2019). *Qualitative data analysis: A methods sourcebook* (4th ed.). Thousand Oaks, CA: Sage.

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Academic Press.

Ouellette, R. R., Frazier, S. L., Shernoff, E. S., Cappella, E., Mehta, T. G., Maríñez-Lorad, A., et al. (2018). Teacher job stress and satisfaction in urban schools: Disentangling individual, classroom, and organizational level influences. *Behavior Therapy, 49*, 494–508.

Owens, J. S., Allan, D. M., Hustus, C., & Erchul, W. P. (2018). Examining correlates of teacher receptivity to social influence strategies within a school consultation relationship. *Psychology in the Schools, 55*, 1041–1055. https://doi.org/10.1002/pits.22163.

Parlangeli, O., Marchigiani, E., & Bagnara, S. (1999). Multimedia systems in distance education: Effects of usability on learning. *Interacting with Computers, 12*(1), 37–49.

Pas, E. T., Johnson, S. R., Larson, K. E., Brandenburg, L., Church, R., & Bradshaw, C. P. (2016). Reducing behavior problems among students with autism spectrum disorder: Coaching teachers in a mixed-reality setting. *Journal of Autism and Developmental Disorders, 46*(12), 3640–3652.

Rayner, C., & Fluck, A. (2014). Pre-service teachers' perceptions of SimSchool as preparation for inclusive education: A pilot study. *Asia-Pacific Journal of Teacher Education, 42*(3), 212–227.

Regalla, M., Hutchinson, C., Nutta, J., & Ashtari, N. (2016). Examining the impact of a simulation classroom experience on teacher candidates' sense of efficacy in communicating with English learners. *Journal of Technology and Teacher Education, 24*(3), 337–367.

Richey, R., Klein, J., & Tracey, M. W. (2011). *The instructional design knowledge base: Theory, research and practice*. New York, NY: Routledge.

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Washington, DC: Sage.

Schussler, D., Frank, J., Lee, T. K., & Mahfouz, J. (2017). Using virtual role-play to enhance teacher candidates' skills in responding to bullying. *Journal of Technology and Teacher Education, 25*(1), 91–120.

Shernoff, E. S., & Kratochwill, T. R. (2007). Transporting an evidence-based classroom management program for preschoolers with disruptive behavior problems to a school: An analysis of implementation, outcomes, and contextual variables. *School Psychology Quarterly, 22*(3), 449–472. https://doi.org/10.1037/1045-3830.22.3.449.

Shernoff, E. S., Mehta, T., Atkins, M. S., Torf, R., & Spencer, J. (2011). A qualitative study of the sources and impact of stress among urban teachers. *School Mental Health, 3*, 59–69. https://doi.org/10.1007/s12310-011-9051-z.

Shernoff, E. S., Lakind, D., Frazier, S. L., & Jakobsons, L. (2015). Coaching early career teachers in urban elementary schools: A mixed-method study. *School Mental Health, 7*, 6–20. https://doi.org/10.1007/s12310-014-9136-6.

Shernoff, E. S., Frazier, S. L., Marinez-Lora, A., Lakind, D., Atkins, M. S., Jakobsons, L., et al. (2016). Expanding the role of school psychologists to support early career teachers: A mixed-method study. *School Psychology Review, 45*, 226–249.

Shernoff, E. S., Frazier, S. L., Lisetti, C., Buche, C., Lunn, S., Brown, C., et al. (2018). Early career teacher professional development: Bridging simulation technology with evidence-based behavior management. *Journal of Technology and Teacher Education, 26*(2), 299–326.

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education & Treatment of Children, 31*(3), 351–380. https://doi.org/10.1353/etc.0.0007.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64*(2), 489–528.

Slaughter, L., Norman, K.L., Shneiderman, B. (1995, March) Assessing users' subjective satisfaction with the Information System for Youth Services (ISYS). In *Proceedings of the Third Annual Mid-Atlantic Human Factors Conference*. Blacksburg, VA: Virginia Tech.

Thomas, D. E., Bierman, K. L., & Conduct Problems Prevention Research Group. (2006). The impact of classroom aggression on the development of aggressive behavior problems in children. *Development and Psychopathology, 18*(2), 471–487.

Thompson, N., & McGill, T. J. (2017). Genetics with Jean: the design, development and evaluation of an affective tutoring system. *Educational Technology Research and Development, 65*(2), 279–299.

Tracey, M. W., Hutchinson, A., & Quinn, G. (2014). Instructional designers as reflective practitioners: Developing professional identity through reflection. *Educational Technology Research & Development, 62*(3), 315–334. https://doi.org/10.1007/s11423-014-9334-9.

Varier, D., Dumke, E., Abrams, L., Conklin, S., Barnes, J., & Hoover, N. (2017). Potential of one-to-one technologies in the classroom: Teachers and students weigh in. *Educational Technology Research and Development, 65*(4), 967–992. https://doi.org/10.1007/s11423-017-9509-2.

Verdú, E., Regueras, L. M., Gal, E., Castro, J. P., Verdú, M. J., & Kohen-Vacs, D. (2017). Integration of an intelligent tutoring system in a course of computer network design. *Educational Technology Research and Development, 65*, 653–677. https://doi.org/10.1007/s11423-016-9503-0.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors, 34*(4), 457–468.

Whyte, E. M., Smyth, J. M., & Scherf, K. S. (2015). Designing serious game interventions for individuals with autism. *Journal of Autism and Developmental Disorders, 45*(12), 3820–3831.

Xie, K., Kim, M. K., Cheng, S. L., & Luthy, N. C. (2017). Teacher professional development through digital content evaluation. *Educational Technology Research and Development, 65*(4), 1067–1111.

Zibit, M., & Gibson, D. (2005). simSchool: The game of teaching. *Innovate: Journal of Online Education, 4*, 1–7.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Elisa S. Shernoff, Ph.D. is Principal Investigator of Grant R305A150166 and an Associate Professor at Rutgers University. Her work focuses on leveraging technology to support teachers to prevent and manage challenging classroom behaviors. The overall goal of her work includes expanding mental health practice in schools to include supporting teacher effectiveness as a mechanism for promoting positive academic and behavioral outcomes for children living in poverty.

Katherine Von Schalscha is a Ph.D. candidate at the University of California-Santa Barbara and worked in the Rutgers IVT-T laboratory during publication of this manuscript.

Joseph L. Gabbard, Ph.D. is an Associate Professor of Human Factors at Virginia Tech and Director of the COGnitive Engineering for Novel Technologies (COGENT) Lab. His research explores methods of design and evaluation for augmented reality and virtual reality user interfaces and experiences.

Alban Delamarre is a Ph.D. candidate at Florida International University. His research focuses on how immersive virtual reality technology and emotional virtual agents can impact user experience and learning.

Stacy L. Frazier, Ph.D. is Professor at Florida International University. Her research examines tiered and technologyfacilitated models of workforce support for youth care systems, toward the goal of reducing health and education disparities.

Cédric Buche, Ph.D. is professor in computer science at the National Engineering School of Brest (ENIB, France). His work includes artificial intelligence, virtual reality, machine learning, robotics and video games.

Christine Lisetti, Ph.D. is an Associate Professor in the School of Computing and Information Sciences at Florida International University. Her research aims at creating diverse virtual social agents (VISAGEs) that can interact naturally with humans with verbal and non-verbal communication, in a variety of contexts involving socio-emotional content (e.g. virtual health counseling and coaches, virtual reality systems and educational games for training on domain-specific social skills).

Affiliations

Elisa S. Shernoff[1] · Katherine Von Schalscha[1] · Joseph L. Gabbard[2] · Alban Delmarre[3] · Stacy L. Frazier[4] · Cédric Buche[5] · Christine Lisetti[3]

[1] School Psychology Program, Rutgers University, 152 Frelinghuysen Road, Piscataway, NJ 08854, USA

[2] Grado Department of Industrial & Systems Engineering, Virginia Tech, 1145 Perry Street, Blacksburg, VA 24061, USA

[3] School of Computing and Information Sciences, Florida International University, 11200 S.W. 8th Street ECS 361, Miami, FL 33199, USA

[4] Department of Psychology, Florida International University, 11200 S.W. 8th Street, Miami, FL 33199, USA

[5] Centre Européen de Réalité Virtuelle, 25, rue Claude Chappe, 29280 Plouzane, France