The Effect of Virtual Humans Making Verbal Communication Mistakes on Learners' Perspectives of their Credibility, Reliability, and Trustworthiness

Jacob Stuart* University of Florida
Karen Aul† University of Florida
Michael D. Bumbach ‡ University of Florida
Anita Stephen § University of Florida
Alexandre Gomes de Siqueira ¶ University of Florida
Benjamin Lok ‖ University of Florida

Figure 1: Images depicting patients: wrong medication error on a patient with a non-visual head injury (left) and wrong site error on a patient with leg and abdomen injuries (right). Patient responses to the virtual nurse indicate the non-visual head injury while the visual injuries are seen in the patient on the right (See Section 3.3).

ABSTRACT

Simulating real-world experiences in a safe environment has made virtual human medical simulations a common use case for research and interpersonal communication training. Despite the benefits virtual human medical simulations provide, previous work suggests that users struggle to notice when virtual humans make potentially life-threatening verbal communication mistakes inside virtual human medical simulations. In this work, we performed a 2x2 mixed design user study that had learners (n = 80) attempt to identify verbal communication mistakes made by a virtual human acting as a nurse in a virtual desktop environment. A virtual desktop environment was used instead of a head-mounted virtual reality environment due to Covid-19 limitations. The virtual desktop environment experience allowed us to explore how frequently learners identify verbal communication mistakes in virtual human medical simulations and how perceptions of credibility, reliability, and trustworthiness in the virtual human affect learner error recognition rates. We found that learners struggle to identify infrequent virtual human verbal communication mistakes. Additionally, learners with lower initial trustworthiness ratings are more likely to overlook potentially life-threatening mistakes, and virtual human mistakes temporarily lower learner credibility, reliability, and trustworthiness ratings of virtual humans. From these findings, we provide insights on improving virtual human medical simulation design. Developers can use these insights to design virtual simulations for error identification training using virtual humans.

Index Terms: Human-centered computing—Empirical studies in HCI——

*e-mail: jacobstuart@ufl.edu
†e-mail: kaul@ufl.edu
‡e-mail: mbumbach@ufl.edu
§e-mail: astephen@ufl.edu
¶e-mail: agomesdesiqueira@ufl.edu
‖e-mail: lok@cise.ufl.edu

1 INTRODUCTION

The ability to simulate real-world experiences in a safe environment has made virtual human (VH) medical simulations a common use case for research and training, particularly in areas that focus on training medical communication skills [5, 6, 9, 13, 17]. VHs have been used to train learners to identify medical mistakes by having virtual healthcare providers commit intentional verbal communication mistakes [31]. However, previous work suggests that despite VH simulations' benefits for training communication skills, learners struggle to notice potentially life-threatening verbal communication mistakes that VHs commit inside these simulations [25, 31]. White et al. found that 87% of students failed to notice a spoken medication error committed by a VH, and Stuart et al. found that 98.55% of students failed to notice a wrong site error communicated by a VH [25, 31]. If translated to the real world, these students' inability to identify verbal communication mistakes could lead to patient injury or death.

Unfortunately, the mistakes made by the VH in Stuart et al.'s work were design mistakes rather than intentional design decisions. More importantly, both prior work by White et al. [31] and Stuart et al. [25] did not investigate how error recognition rates may be affected by VH factors. Stuart et al. suggest that high trust may have hindered learners' abilities to identify the accidental error in their work. However, Stuart did not use a pre-trust survey and was unable to confirm this [25]. Factors such as learner perceptions of a VH's credibility, reliability, and trustworthiness (CRT) are found to influence learner reliance on virtual agents [8, 20]. Learner over reliance on virtual agents can lead to failure to properly monitor automation [8, 20]. In this context, a failure to monitor would include a learner's failure to take proper notice of deteriorations, injuries, or other medical negligence.

Therefore, we performed a dedicated study where VHs intentionally made scripted verbal communication mistakes. In this study, the medical training simulation is a triage assessment, and the VH is performing the role of an emergency department nurse. This work focuses on two different medical mistakes: wrong-site surgery er-

978-1-6654-9617-9/22/$31.00 ©2022 IEEE
DOI 10.1109/VR51125.2022.00065

rors and wrong-medication errors (See section 3.6). In the medical field, wrong medication errors frequently occur with a rate of 5 per 100 medication administrations, while wrong-site surgery errors infrequently occur with a rate of 1 in 112,994 operations [11]. By exploring errors of different frequencies, we hope to better train learners to identify errors by improving VH medical simulation design. This research aimed to answer the following questions:

- RQ1: How frequently are VH mistakes identified by learners?
- RQ2: Do learner perceptions of the VH's credibility, reliability, and trustworthiness affect learners' abilities to identify life-threatening mistakes?
- RQ3: How do VH mistakes impact learner perceptions of the VH's credibility, reliability, and trustworthiness?

The work presented here is part of a larger research effort to improve error recognition within VH medical training simulations. This work contributes the following to the IEEEVR community by building on previous VH medical training simulation research such that:

- (1) It provides evidence that learners struggle to identify uncommon verbal communication mistakes committed by VHs that could have life-threatening consequences. It goes beyond prior work, providing evidence that error recognition can be improved by improving VH trustworthiness.
- (2) It provides evidence that VH mistakes, when recognized, negatively impact learner perceptions of the VH. However, the negative perceptions caused by life-threatening mistakes are temporary.

2 Related Works

Teaching healthcare via simulation uses a broad spectrum of educational approaches. The fidelity of these approaches range from traditional paper-based case study simulations to high fidelity mannequin simulators [25, 31]. Recently, virtual human medical simulations have become more common in medical education [5, 6, 9, 13, 17]. Simulation is often used to expose learners to situations where they can learn from failures. One area that researchers are beginning to explore using simulation is recognizing errors [31]. This section aims to situate our work among previous work done involving virtual humans and trust, virtual human errors, medical simulations that train error identification, and the impact of errors in the real world.

2.1 Virtual Humans and Trust

The relationship of trust and virtual agents has been studied through aspects of appearance (attire, anthropomorphism, fidelity), voice quality, performance levels, and feedback [12, 16, 27, 33]. However, since this work focuses on trust in relation to a virtual nurse's performance, this section focuses on performance and feedback.

Wang et al. found that the performance level of a robot agent is positively correlated with learner trust [29]. The learners interacted with either a high-ability robot, which always made the correct decision or a low-ability robot that occasionally made the incorrect decisions regarding a building's safety. The authors found that self-reported trust in the low-ability robot was highly correlated with understanding its decision and decision-making processes. The authors speculate that participants did not question the high-ability robot since it was always correct. Since the low-ability robot was not always correct, the explanations became more influential.

Jensen et al. found that learner trust in a virtual agent increased with the agent's accuracy and that trust varied based on the feedback given [12]. The authors found that the high reliability (90% accuracy) led to greater behavioral trust in the system than low reliability (60% accuracy). They also found that framing errors as the developer's mistake rather than the system's led to decreased learner trust in the system.

Hafizoglu et al. explored how the reputation of a VH's performance in a task will affect a learner's trust in the agent and reliance on the agent [7]. The researchers found that providing an endorsement of trustworthiness before interacting with the virtual agent increased learners' trust in the agent and their reliance on the agent to complete the given tasks.

Then in a follow-up study, rather than telling the participants what the agent's reputation was, the authors had participants interact with a trustworthy/untrustworthy agent and then measured how this past experience with the agent impacted learner trust, and reliance [8]. The agent's trustworthiness was determined by whether the agent would complete its fair share of the work in the given tasks (trustworthy) or not (untrustworthy). The authors found that a positive experience with a virtual agent would increase trust and reliance and that a negative experience decreases trust and reliance.

The work of Wang, Jensen, and Hafizoglu suggests that improving agent performance can increase trust and reliance on virtual agents. But, previous work did not investigate if these effects exist for VHs in medical simulations, if the effects persist, or how errors affect perceptions of reliability and credibility. Our work aims to confirm findings that learners struggle to identify life-threatening errors in medical simulations and explores the relationship between VH errors and learner CRT perceptions.

2.2 Virtual Human Errors

Much of the work regarding VH verbal communication errors focuses on VHs making conversational errors due to technical limitations of VHs such as rendering errors, inability to handle unexpected input, misrecognition, overanswering, or failures resulting from out of domain/expertise queries to the VH [3]. Investigating technical errors have allowed the community to improve areas such as dialogue systems and even allow us to measure virtual human experiences in different ways [4, 23].

However, VH technical limitation errors differ from the errors discussed in this work where VHs make what we will refer to as content mistakes. Content mistakes are virtual human errors that simulate errors made by real humans in realistic scenarios. These types of mistakes can be unintentional or intentional (typically part of educational learning objectives). While fewer works explore these content mistakes, two examples include the works of Stuart and White.

White et al.'s work is an example of an intentional content error [31]. White created a virtual human medical simulation where learners work with VHs to treat a patient. The simulation assessed learners' information transfer and communication skills. The simulation designers intentionally had a VH make a spoken drug dosage mistake that could be fatal if not caught during the simulation. It was the user's responsibility to identify this mistake. They found that approximately 87% of learners did not notice a fatal drug mistake committed by the VH.

Stuart et al. exemplifies works with unintentional content errors [25]. Stuart created a VH medical simulation intended to help learners practice communication skills. They later found that an unintentional verbal communication mistake was made by the VH where they stated the wrong leg to the learners acting as the nurse observers. Despite learners being able to visually see the patient's injury, 98.55% of learners did not identify the mistake.

Content errors are commonplace in more traditional healthcare education, such as medical case studies. However, content errors are only more recently being used in VH medical simulations. Therefore, this work builds on existing research that explores VH verbal communication errors in medical simulations and medical simulations error recognition rates.

2.3 Recognizing Mistakes in Medical Virtual Simulation

In looking at different medical simulation approaches, we see a range of recognition rates for content errors. On the low end of error recognition rates, Stuart et al.'s work found that less than 2% of learners identified the error, while on the high end of error recognition rates, Warholak et al. found that approximately 60% of learners identified an error [25, 30]. The fidelity of these approaches varied with Warholak using traditional paper-based case study simulations and Stuart using VH medical simulations. In this section, we will briefly describe how the error recognition rates of previous approaches used in healthcare education and how error recognition rates compare to those found in VH medical simulations with content errors.

Warholak et al. used a written questionnaire-style methodology to investigate healthcare learners' prescribing error recognition rates [30]. They found that approximately 60% of nursing learners noticed a drug dosage mistake made using this written format. Using a similar written methodology to Warholak, Whitehair et al. found that 44.2% of learners noticed a drug dosage mistake made using this written format and that 57% of learners identified a mistake where a drug a patient was allergic to was given to a patient [32].

VH medical simulation work reports lower learner error recognition rates than rates reported when using more traditional writing formats [25, 30–32]. When controlling for error type, the VH medical simulation error recognition rate (13%) is lower than the approximately 60% noted by Warholak and the 44.2% reported by Whitehair. All of these errors were related to the amount of a drug prescribed to a patient. Stuart et al. reported an even lower error recognition rate but the authors were unable to find a similar error type using traditional approaches for comparison.

In the VH medical simulation literature explored, both of the errors reported in the virtual simulations were verbal communication errors made by a VH, suggesting learners struggle to identify verbal communication mistakes more so than written content errors. Therefore, we performed a dedicated study where VHs intentionally made scripted verbal communication mistakes to help us understand how frequently learners identify verbal communication mistakes.

2.4 The Impact of Medical Errors in The Real World

In 2000 the Institute of Medicine stated that up to 96,000 people in the US die each year due to medical errors [15]. These numbers have increased since this report with some experts stating that up to 440,000 people die each year in the US due to medical errors [18]. Additionally, Medical errors are estimated to cost as much as $17.1 billion per year and Never Events make up approximately $3.7 billion of that cost [28]. Never Events are adverse events that should never happen because the events are clearly identifiable, measurable, serious, and preventable [1]. It is critical that we understand how to improve learner identification of these errors to reduce the unnecessary deaths, injuries, and costs that arise because of poor medical error recognition.

3 Study Design

Using a 2x2 mixed design user study, we explored how factors such as the real-world frequency of the mistakes made by the VH and learner perceptions of CRT affected learners' abilities to identify the mistakes made. The user study has learners observe a VH make either a mistake that frequently occurs in healthcare (wrong medication) or a mistake that infrequently occurs in healthcare (wrong site). Acting as a nurse observer, the learner will be asked to record any mistakes that occurred in the scenario after they have finished viewing the assessment. Injury type and order were manipulated between-subjects to control for ordering effects and potential discrepancies between injury types. The effect that VHs making mistakes has on learners' self-reported CRT perceptions was investigated within subjects.

Figure 2: The laceration and compound fracture injuries.

3.1 Modality

This work was initially intended to use virtual reality with 360 environments. This experience was designed using Unity and deployed using mobile VR. Unfortunately, due to the COVID-19 pandemic, it became unfeasible to have students use shared VR headsets. As a result, we converted the experience to an online desktop experience that students could complete from anywhere. This was done by pre-recording the Unity experiences that students would observe and showing the videos in multiple stages of a Qualtrics survey. The interaction is intended to help students improve their communication skills regarding observations the virtual nurse makes during the interaction. By moving the interaction from a VR experience to a desktop experience, we were still able to allow students to address the interactions learning objectives. Additionally, this allowed us to continue to answer questions regarding simulation design and provide a better understanding of how learners perceive VH mistakes. However, future work may investigate the benefits of using a more interactive VR experience to provide learners an opportunity to directly intervene in the triage assessment when an error is noticed or investigate if higher immersion levels increase error recognition rate.

3.2 Character Design

Three VHs were developed for this study. The authors created the VHs using Adobe Fuse, and animations applied to the VHs were created using Autodesk Motionbuilder. Two black male patients were developed to maintain race consistency during the interaction and a white female virtual nurse was developed to lead the triage assessment. A separate voice actor recorded each VH voice. These characters were evaluated by nursing educators and deemed suitable for educational use.

3.3 Injury Design

Two injury conditions were designed for this study: A non-visual head injury condition and a visual injury condition with a laceration to the abdomen and a compound leg fracture (see figure 2). Patient responses to the virtual nurse indicate the non-visual head injury. The visual injuries were created from real images of these types of injuries and applied to the VH body textures using Adobe Photoshop. The injuries applied to patients vary based on a participant's group (See Table 1). Three nursing collaborators assessed the VH injuries and deemed them to appear accurate.

3.4 Environment Design

The environment is a 360 image recorded using a Virb 360 camera. The 360 image displays an image of the clinical simulation room that the participants use during simulation training in their health assessment course. Initial development intended users to be able to move closer to the patient in a VR environment. Once the interaction became a desktop experience due to COVID-19, the user's point of view was placed at the foot of the bed and made static to focus on the triage assessment.

3.5 Triage Assessment Design

In this study, learners view a virtual nurse performing two triage assessments. We divided each triage assessment into two parts: An observational video of a virtual nurse assessing a patients vitals (vitals video) and an observational video of a virtual nurse performing a physical assessment of a patient (triage assessment video). This was done so that perceptions of the nurse could be measured after giving correct vitals information to the learner and after incorrect triage assessment information is given. The vitals videos include the virtual nurse stating the emergency medical services report, which provides a general overview of the type of injuries the patient has, and the virtual nurse reporting the current vital signs of the virtual patient. The triage assessment videos have the virtual nurse introduce themselves to the patient, ask a series of questions to determine the patient's state of mind (e.g., What brought you here today?, What is your name, etc.), ask about the patient's medical information (e.g., Do you have any allergies, Are you on any medications?, etc.), perform a brief physical examination of the patient, and then the virtual nurse tells the patient what course of action they plan to take to help them. For further detail on the processes followed by the virtual nurse, please see the supplemental video.

3.6 Medical Error Design

Two errors were created for this study. The first error is a wrong site error. This error has the virtual nurse state the wrong leg to the learner acting as the nurse observer. If the injury site is recorded incorrectly by the nurse observer, there can be dire consequences such as surgery being performed on the wrong limb. The second error is a wrong medication error. This error has the virtual nurse provide a medication (morphine) for which the patient previously stated they have an allergy. If a nurse observer does not identify a medication error, a patient can have an allergic reaction that can lead to longer recovery times, injury, or even death. Nursing collaborators chose the wrong site and wrong medication errors for this study as they are considered Never Events. Never Events are adverse events that should never happen because the events are clearly identifiable, measurable, serious, and preventable [1].

3.7 Accelerated Future Design

Accelerated futures were used to let participants know of the VH verbal communication mistake and its consequences in case the learner did not identify the mistake during the triage assessment. The accelerated future video included a segment showcasing the patient's condition two years into the future as a result of the error and a debrief from the instructor (See fig. 3). The accelerated future occurred for both the correct and incorrect assessments, but the correct assessment had a notice before the accelerated future that the following video would have happened if the VH had made a mistake. For the patient with the leg injury, the accelerated future depicts the patient, who is missing both legs rather than just one, discussing the anger and distress they are experiencing due to their medical care. This scenario is possible in real life if surgery is performed on a wrong limb due to a wrong site error. For the patient with the head injury, the accelerated future depicts the patient being angry and complaining of the harm caused to them due to an allergic reaction to the medication given to them. This scenario is a possibility in real life if a wrong medication is given to a patient and the patient suffers an allergic reaction.

4 Participants

This study included 80 nursing students enrolled in a health assessment course offered at a local university. The health assessment course had 124 students enrolled in total, with 12% male and 88% female. Students were recruited through an announcement posted on the nursing course's learning management system, and the students

Figure 3: Example images from the accelerated futures of the patients with the leg injury (top) and head injury (bottom).

were offered course extra credit to participate. Students could participate in the study by using the provided Qualtrics link to review and agree to the online informed consent form. Students who agreed to participate were enrolled electronically in the study and could begin the simulation.

5 Measures

This study had participants answer questions regarding the accuracy of the triage assessments and complete six CRT questionnaires. Other data were collected during this study but will not be referenced in this work to maintain simplicity.

The CRT questions were modeled after questions from Kalyanaraman et al. [14]. The CRT questions asked *"Please rate your level of agreement with the following adjectives describing the virtual nurse presented in the simulation, with (1) representing 'Strongly Disagree' and (9) representing 'Strongly Agree.'"* for the adjectives credible, reliable, and trustworthy. CRT was measured using six identical surveys throughout the simulation.

Participant performance in spotting the error was measured via the ratings and explanations of accuracy assigned to the virtual assessments. The accuracy questions asked *"How accurate was the nursing's report for your patient? (1) representing 'Not Accurate At All' and (9) representing 'Completely Accurate/ No Errors'"*. If participants rated accuracy less than 9, they were asked to explain why they provided the rating they did in an open ended response.

6 Procedure

Overall, the study procedure took participants approximately 30 minutes to complete. To begin the study, learners were sent a link by their instructor that led them to an informed consent form. If the learners agreed to participate, the participants then completed a pre-test to assess their situation, background, assessment, recommendation (SBAR) and triage knowledge (data not used in this paper). After completing the pre-test, the participants were then pre-briefed on the scenario via a pre-recorded video. The pre-brief informed the participants on the accident that occurred, the patients they were about to see, the learning objectives of the scenario, and their role as a nurse observer during a triage assessment conducted by a VH nurse. Once the pre-brief was over, participants observed the first of two triage assessments performed by a virtual nurse. Each triage assessment consisted of four parts: An observational video of a virtual nurse assessing a patients vitals (vitals video), an observational video of a virtual nurse performing a physical assessment of a patient (triage assessment video), a student evaluation, and an observational video showing the patient's accelerated future.

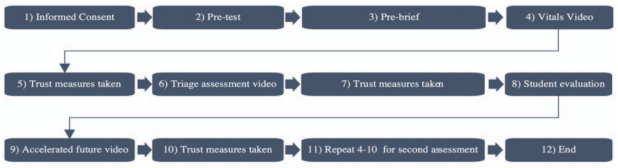

Figure 4: The study flow experienced by learners.

CRT ratings were recorded a total of six times. Measurements were taken after each vitals video, assessment video, and accelerated future video. The observational videos of the virtual nurse assessing a patients vitals, which were correct for each assessment, acted as controls for the CRT measurements. The observation of the physical assessment is where the error occurred in the first assessment. The student evaluation asked participants to rate the accuracy of the triage assessment. This process can be seen in figure 4.

7 Groups

Participants were randomly assigned to one of four groups to control for ordering effects and potential discrepancies between injury types. Patient order is described in Table 1.

Random assignment was completed using Qualtrics randomization features which resulted in uneven group sizes. However, it is unlikely that the uneven group sizes affected the results. This is because the pairings SE1/SE2 (N = 41) and ME1/ME2 (N = 39) are identical groups until triage assessment 2 (See Table 1). The error recognition rate and initial learner perceptions are unaffected by triage assessment 2. The CRT perceptions over time may be affected by the second triage assessment. However, the groups exhibited similar trends showing increases in CRT perceptions over time so the authors do not see the varying group sizes as an impactful factor (See Fig. 6 for visual and Section 9.3 for discussion).

8 Results

The main goal of this study is to improve VH medical simulation design by providing insight on why life-threatening mistakes made by VHs go unnoticed. This can be done by reducing monitoring failures. To understand why monitoring failures occur, we explored the impact of mistake frequency and learner CRT perceptions in the VH. Therefore, measures to analyze CRT in the VH and student error recognition rate were used. The VH CRT questions were designed off of questions from Kalyanaraman et al. [14]. To compare learner CRT ratings between those who discovered mistakes against those who did not unpaired t-tests were used. Additionally, to analyze the non-parametric Likert scale data, the Wilcoxon Signed-Rank non-parametric statistical hypothesis test was performed using the R statistical software (See Table 2) [21].

8.1 Error Recognition Rate

Error recognition was measured by having learners rate the VH's accuracy on a 1-9 scale. If choosing less than 9, learners were asked to explain why they believed the VH's assessment was inaccurate. The number of learners who reported errors, reported the correct error, and who did not report errors were tallied. 41 learners saw the wrong site error and 39 saw the wrong medication error. The error recognition rate for the wrong site error was 4.89% (2/41). The error recognition rate for the wrong medication error was 56.41% (22/39). For the wrong site error, 34 learners did not report any errors, 6 learners reported errors, and only 2 learners reported the wrong site error. For the wrong medication error, 9 learners did not report any errors, 30 learners reported errors, and 22 learners reported the wrong medication error.

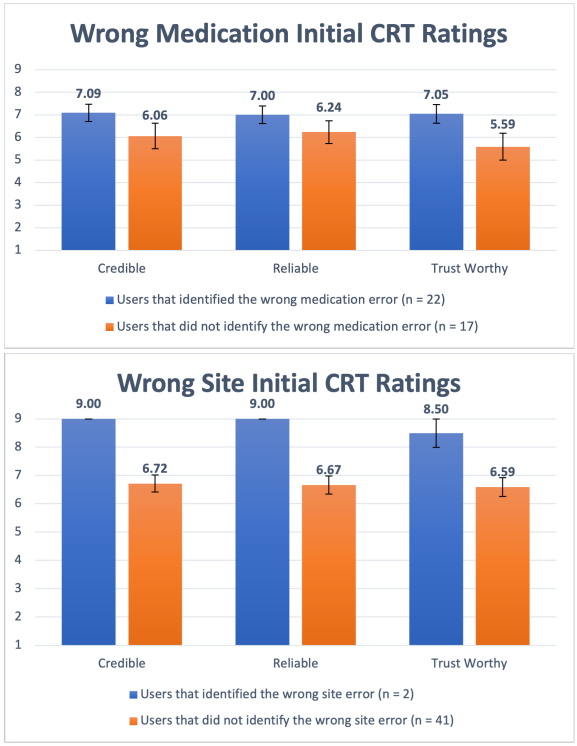

Figure 5: Graphs displaying the average initial CRT ratings of learners broken up by error type and if users identified the mistakes.

8.2 Initial Learner Perceptions

Learners were asked to rate whether they felt the VH was credible, reliable, and trustworthy using a 1-9 scale (1 being strongly disagree and 9 being strongly agree). This data was then separated by the error type the learners saw and whether they discovered the error (See Fig. 5). An unpaired T-test was used to compare learner perceptions of those who found the wrong medication error against those who did not. This was not done for the wrong site group as the number of learners who noticed the wrong site error was limited (n = 2).

The 2 learners who noticed the wrong site error reported (M = 9.0, SD = 0.0), (M = 9.0, SD = 0.0), and (M = 8.5, SD = 0.707) for credibility, reliability, and trustworthiness respectively. This is compared to the 41 learners who did not notice the medication error who reported (M = 6.72, SD = 1.90), (M = 6.67, SD = 2.03), and (M = 6.59, SD = 2.11) for credibility, reliability, and trustworthiness respectively.

The 22 learners who noticed the medication error (M = 7.05, SD = 1.94) compared to the 17 learners who did not notice the medication error (M = 5.59, SD = 2.48) demonstrated significantly higher ratings of VH trustworthiness, $t(37) = 2.06$, $p = 0.0462$.

The same is not true for credibility or reliability ratings for the VH. The 22 learners who noticed the medication error (M = 7.09, SD = 1.82) did not demonstrate significantly higher ratings of VH credibility compared to the 17 who did not (M = 6.06, SD = 2.33), $t(37) = 1.55$, $p = 0.129$. Also, the 22 learners who noticed the medication error (M = 7.00, SD = 1.83) did not demonstrate significantly higher ratings of VH reliability compared to the 17 who did not (M = 6.24, SD = 2.08), $t(37) = 1.22$, $p = 0.229$ (See Fig. 5).

8.3 Learner CRT Perceptions Over Time

Each of the CRT measures were taken 6 times throughout the simulation (See Fig. 6). To show the changes in CRT perceptions over

Group	N	Triage Assessment 1	Triage Assessment 2
Wrong Site Error Group 1 (SE1)	16	Patient 1: Leg Injury (With Wrong Site Error)	Patient 2: Head Injury
Wrong Medication Error Group 1 (ME1)	17	Patient 2: Head Injury (With Wrong Medication Error)	Patient 1: Leg Injury
Wrong Site Error Group 2 (SE2)	25	Patient 1: Leg Injury (With Wrong Site Error)	Patient 2: Leg Injury
Wrong Medication Error Group 2 (ME2)	22	Patient 1: Head Injury (With Wrong Medication Error)	Patient 2: Head Injury

Table 1: Ordering injury type for each group

time, we use three comparisons.

First, CRT ratings recorded after the first triage assessment's vitals video (T1) are compared against those taken after the first accelerated future video (T3) where learners are alerted of the error to see how VH mistakes affect CRT ratings. CRT ratings after the first patient safety report (T3) were statistically significantly lower than the CRT ratings measured after the first patient's vitals were stated (T1) (See Table 2). This shows that once all learners have been alerted of the VH verbal communication mistake their CRT ratings dropped.

Second, CRT ratings taken after the first accelerated future video (T3) are compared against ratings recorded after the second triage assessment's vitals video (T4) to understand if student CRT ratings were affected by a new triage assessment beginning. CRT ratings after the second patient's vitals were stated (T4) were statistically significantly higher than the CRT scores measured after the first patient safety report (T3) (See Table 2). This increase suggests that reduced CRT perceptions will quickly return to original levels. Finally, CRT ratings recorded after the second triage assessment's vitals video (T4) are compared against those taken after the second accelerated future video (T6) to see how CRT perceptions recover after viewing a correct assessment. CRT ratings after the second patient safety report (T6) were statistically significantly higher than the CRT ratings measured after the second patient's vitals were stated (T4) (See Table 2). This indicates that viewing a correct scenario after being alerted of a VH verbal communication mistake further increases learner trust in the VH.

Additionally, the wrong site error groups saw increases in CRT ratings from the first vitals video (T1) and the ratings taken after viewing the first triage assessment (T2) while the wrong medication error groups saw decreases in CRT ratings. This difference is likely due to the differences in error recognition rate in the two injury types (see Section 8.1). The wrong site groups likely saw increases in their CRT ratings because many did not notice the wrong site error during the assessment. In contrast, over half of the wrong medication groups noticed the error, leading to a decrease in their CRT ratings directly after viewing the assessment. These diverging CRT ratings provide further evidence that VH mistakes lower CRT ratings (See CRT changes over time in Fig. 6).

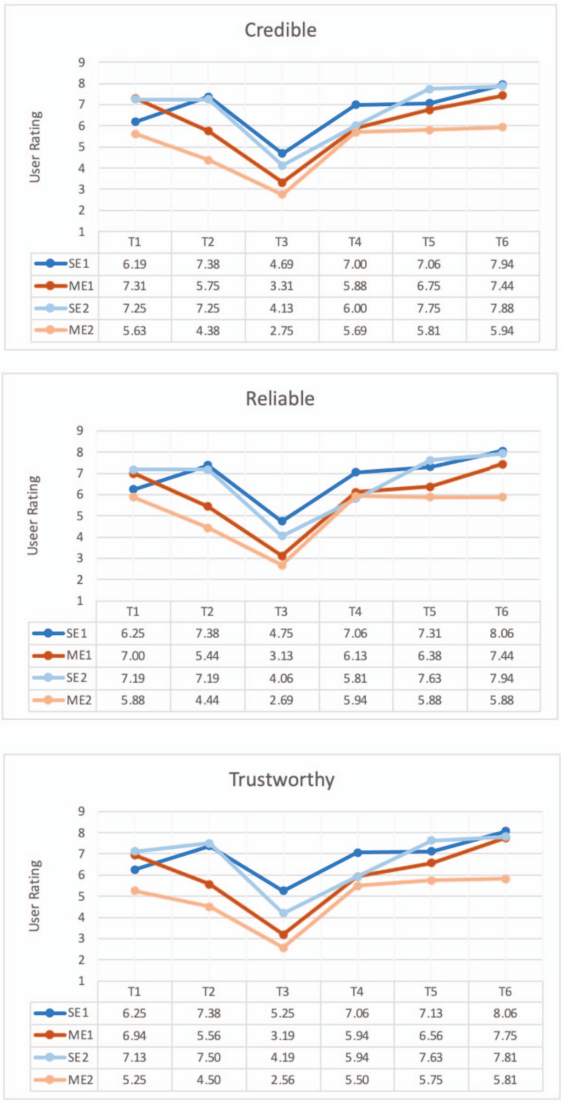

Figure 6: Figures showing the CRT ratings of learners over the course of the interaction. The graphs are broken up by groups and showcase the average measurements at each CRT measurement time point.

9 Discussion

This section discusses how frequently learners identify different types of VH mistakes (RQ1), how learner perceptions of a VH's CRT affect learners' abilities to identify VH mistakes (RQ2), how VH mistakes affect learner perceptions of the VH's CRT (RQ3), and draws an overall conclusion from our findings.

Question	$T_1 vs. T_3$	$T_3 vs. T4$	$T_4 vs. T_6$
Reliable	n = 61, $\mu_1 = 6.74, \mu_2 = 3.9, M_1 = 7, M_2 = 3, Z = 6.08 **$	n = 54, $\mu_1 = 3.9$, $\mu_2 = 6.45, M_1 = 3, M_2 = 7, Z = -5.76 **$	n = 57, $\mu_1 = 6.45$, $\mu_2 = 7.34, M_1 = 7, M_2 = 9, Z = -3.47 **$
Credible	n = 65, $\mu_1 = 6.7$, $\mu_2 = 3.96, M_1 = 7, M_2 = 3, Z = 5.58 **$	n = 55, $\mu_1 = 3.96$, $\mu_2 = 6.41, M_1 = 3, M_2 = 7, Z = -5.71 **$	n = 58, $\mu_1 = 6.41$, $\mu_2 = 7.3, M_1 = 7, M_2 = 9, Z = -3.62 **$
Trustworthy	n = 64, $\mu_1 = 6.55$, $\mu_2 = 4.03, M_1 = 7, M_2 = 3, Z = 5.53 **$	n = 52, $\mu_1 = 4.03$, $\mu_2 = 6.4, M_1 = 3, M_2 = 7, Z = -5.73 **$	n = 57, $\mu_1 = 6.4$, $\mu_2 = 7.36, M_1 = 7, M_2 = 9, Z = -3.68 **$
$p < .05$ $p < .01$			

Table 2: n is the number of non-zero differences used in the Wilcoxon signed-rank test, M1 and M2 are the first and second groups medians, μ_1 and μ_2 are the means of the groups, and Z is the calculated Z-Score.

9.1 Error recognition rates were higher for frequent medical mistakes than infrequent medical mistakes (RQ1)

The error recognition rate results indicate that learners are more likely to identify frequent verbal communication mistakes made by VHs than infrequently occurring mistakes. The error recognition rate for the infrequent mistake was 4.76%, and the rate for the frequent mistake was 56.41%. Previous literature shows widespread uses of simulation types and error recognition rates when training learners to identify errors [30–32]. Our results for frequent verbal communication mistakes fall within the range of error recognition rates of written approaches [30,32]. Therefore, this work contributes to the growing literature that uses VHs to train error identification in virtual simulations by showing that VH medical simulations that use verbal communication can provide results similar to more traditional written approaches for identifying frequent medical errors.

However, learners identified infrequent verbal communication mistakes less often than frequent verbal communication mistakes. Unlike the frequent mistakes in this study, the training related to infrequent mistakes had yet to be covered in the learner's curriculum. Despite prior medical knowledge not being needed to spot the verbal communication mistakes that occurred in this study, the lack of training placed in the curriculum for the infrequent mistake may have made learners less aware of the infrequent verbal communication mistake, leading to a lower error recognition rate. This suggests that VH medical simulation training error recognition should be performed after education related to the medical errors involved. However, if virtual simulation designers choose to deploy VH medical simulation training for error recognition before education on the corresponding errors, previous work suggests that implementing a lecture focused on the medical error afterwards may improve learning [24,26].

9.2 Learners with Higher Perceptions of a VH's trustworthiness had Higher Error Recognition Rates (RQ2)

Initial CRT ratings show that learners who noticed the mistakes reported significantly higher initial ratings of trustworthiness in the VH than learners who did not notice the mistakes (see Fig. 5). While not significant, credibility and reliability ratings also exhibited similar trends to the trustworthiness ratings. This suggests that learner perspectives may influence their abilities to identify communication mistakes. This scenario suggests that trustworthiness may be a predicting factor but this may differ for other scenarios. Further, it is also unclear whether these perceptions continue to influence user's in multi-stage interactions as we will discuss in the next section.

9.3 VH Verbal Communication Mistakes Temporarily Lowered CRT Ratings (RQ3)

Evidence suggests that VH verbal mistakes reduce CRT ratings. Learners demonstrated significant decreases in CRT ratings once alerted of the VH error. CRT rating averages decreased from 6.74, 6.7, and 6.55 to 3.9, 3.96, and 4.03 ($p < .01$ for each CRT Value) after being alerted of the VH mistake. Additionally, groups with more learners that identified the VH verbal mistake during the VH's assessment reported lower CRT ratings before being alerted of the mistake than those that did not identify the mistake during the assessment. However, average CRT ratings increased from 3.9, 3.96, and 4.03 at the end of the first assessment to 6.45, 6.41, and 6.4 ($p < .01$) at the beginning of the second assessment suggesting these CRT rating reductions do not persist over time (See Table 2).

These CRT findings are important because research has shown that healthcare professionals that make mistakes such as this are prone to do so again [19]. This would suggest that learner CRT ratings should not increase at the beginning of the second assessment. Reasons for increased CRT ratings at the start of the second assessment may include learners treating the assessments as discrete scenarios, failure to establish an easy-to-follow storyline for learners, or the act of watching a VH state vitals correctly is enough to increase learner trust.

Learners' increased CRT ratings after the VH states details like vitals indicates that a reset effect may exist regarding learner feelings of CRT. This reset effect may cause participants to treat the next patient similar to a new level in a video game, disregarding much of what they saw from the VH in the previous assessment. Another reason for the increase in CRT ratings at the start of the second assessment may be that the virtual simulation fails to establish a fiction contract with the learners. A fiction contract is an implicit or explicit agreement regarding how learners are expected to interact with the simulated situation [22]. Despite using pre-briefing to describe the scenario and the use of VHs to provide visual detail, it is possible that the simulation failed to establish a fiction contract with the learners that would have them treat both assessments as a continuous scenario rather than two discrete scenarios.

Additionally, it is also possible that simply seeing the VH perform any action correctly after seeing them make a mistake is enough to increase CRT levels similar to where they began. Further research is needed to determine the reasoning for the immediate increase in CRT ratings to help make multi-stage interactions more cohesive.

9.4 Insights For Virtual Human Training

Overall, our results suggest that 1) the relationship between VHs and trust is more complicated than previous work suggests, and 2) developers should be aware of VH training's ability to lead to negative transfer of training [10].

Previous research in virtual agents and trust indicates that higher user trust will lead to higher user reliance on a virtual agent [7,8,12]. However, we found that learners who reported higher initial ratings of trustworthiness were more likely to identify VH errors. This indicates that these high initial trust users were relying less on the information provided by the VH and more on their critical thinking skills. This finding suggests that further research is needed to understand the relationship between a VH and trust in a healthcare simulation. This need is further bolstered by the unclear reasoning for trust increasing at the start of the second triage assessment.

As for transfer of training, positive transfer of training occurs when a user successfully applies knowledge or skills to a new domain, and negative transfer of training occurs when a learner reacts correctly to a situation based on their previous training, but

incorrectly for how a real world scenario would require them to react [10]. When topics have been covered in a learner's curriculum, VH training can lead to positive transfer of training by offering a safe environment for learners to apply their knowledge and critical thinking skills. However, if applying VH training to topics that have not yet been covered, VH training may lead to negative transfer of training. VH training simulations have become common in medical education [5, 6, 9, 13, 17]. Despite simulations being commonplace in healthcare education, there are few examples of healthcare simulations that have VHs intentionally make mistakes. The correctness of VHs in previous simulations may lead learners to react inappropriately in simulations designed to have VHs make mistakes. This opens the opportunity for learners to form dangerous habits such as deferring to information provided to them by the VH rather than applying their own critical thinking skills. To avoid negative transfer of training, we suggest 1) that educators place VH training simulations after the skills and knowledge for a module have been taught, and 2) that developers of simulations highlight to users that intentional errors may exist in a VH simulation. By highlighting that errors may exist, developers can prevent users from thinking that errors that occur in a simulation are by developer accident rather than by design.

10 Limitations

The limitations include—(1) There is the potential that more participants rated the accuracy as a 9 (completely correct) so that they did not have to provide a response as to why they rated the assessment accuracy any lower. This seems unlikely as learners were aware they could stop the study at any time and still receive the extra credit and there was no other evidence indicating this occured. (2) The ordering used to analyze the impact of mistakes on CRT prevents a comparison to groups who did not experience mistakes at all. Though, based on previous work and the wrong site groups' upward trend when not noticing the error, we believe that learner CRT levels would have only increased for a group that did not experience mistakes [12, 29]. (3) While the low error recognition rate for the infrequent mistake suggests that using simulation before lessons may lead to wasted opportunities for learners to practice properly applying their skills in more realistic settings it is possible that setting up a lesson using simulation may improve future learning in lectures but this was not explored in this work. (4) This work was initially intended to use virtual reality with 360 environments. This experience was designed using Unity and deployed using mobile VR. Unfortunately, due to the COVID-19 pandemic, it became unfeasible to have students use shared VR headsets. As a result, we converted the experience to an online desktop experience that students could complete from anywhere. A head-mounted virtual reality experience may have led to higher levels of immersion which in turn could affect user error recognition rate or CRT perceptions of the virtual human. The effect of immersion levels on error recognition rate and CRT perceptions represent opportunities for future research. (5) It is possible that users struggled to differentiate credibility, reliability, and trustworthiness when rating the virtual humans leading the results to show similar trends. However, we recommend that future works continue to separate these metrics as it may help determine user concerns in particular healthcare scenarios better. For example, a patient may find the information given to them credible but not find the healthcare provider reliable enough to perform procedures; without separating the metrics, it would be more challenging to understand this in future works.

11 Future Work

This work has important implications for those in the IEEEVR community working in training simulations. Correct placement of simulation in curriculum, whether simulation is used before or after a lesson taught by an instructor, is not mentioned in simulation guidelines [2, 22]. Therefore, more research is needed to understand where simulations like the one discussed in this paper should fit into curriculum and what role they should play in optimizing learning opportunities. Additionally, this work suggests that further research is needed to determine the reasoning for the immediate increase in CRT ratings to help make multi-stage interactions more cohesive. Finally, this work purposefully has learners act as nurse observers. However, future work may benefit from a more scaffolded approach with more experienced learners that has them take a more interactive approach with the virtual human.

12 Conclusion

Low error recognition rates present serious concerns to the simulation community. If translated to the real world, the low error recognition rates found in this study could lead to severe injuries or death. We explored how frequently learners identify verbal communication mistakes in VH medical simulations and how factors such as perceptions of credibility, reliability, and trustworthiness in the VH affect error recognition rates. Our results provide evidence that learners are less likely to identify infrequent VH verbal communication mistakes. Learners with lower initial trustworthiness ratings are more likely to overlook potentially life-threatening mistakes. When these mistakes are recognized, we can observe a temporary decrease in learner CRT ratings for the VH. Additionally, we discuss how the placement/role of simulation may affect error recognition rates and highlight options for future exploration regarding our findings on VH credibility, reliability, and trustworthiness.

Acknowledgments

This work was funded by the National Science Foundation award numbers 1800961 and 1800947.

References

[1] Never Events, 9 2019.

[2] M. Alexander, C. F. Durham, J. I. Hooper, P. R. Jeffries, N. Goldman, S. ardong-Edgren, K. S. Kesten, N. Spector, E. Tagliareni, B. Radtke, and C. Tillman. NCSBN Simulation Guidelines for Prelicensure Nursing Programs. *Journal of Nursing Regulation*, 6(3):39–42, 2015. doi: 10.1016/S2155-8256(15)30783-3

[3] T. Bickmore, H. Trinh, R. Asadi, and S. Olafsson. Safety First: Conversational Agents for Health Care. 2018. doi: 10.1007/978-3-319-95579-7

[4] C. Bousquet-Vernhettes and N. Vigouroux. Recognition error handling by the speech understanding system to improve spoken dialogue systems. *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, 2003.

[5] J. C. Cendan and B. Lok. The use of virtual patients in medical school curricula. *American Journal of Physiology - Advances in Physiology Education*, 2012. doi: 10.1152/advan.00054.2011

[6] J. H. Chuah, B. Lok, and E. Black. Applying mixed reality to simulate vulnerable populations for practicing clinical communication skills. *IEEE Transactions on Visualization and Computer Graphics*, 2013. doi: 10.1109/TVCG.2013.25

[7] F. M. Hafizoglu and S. Sen. Reputation Based Trust In Human-Agent Teamwork Without Explicit Coordination. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. ACM, New York, NY, USA, 12 2018. doi: 10.1145/3284432.3284454

[8] F. M. Hafizoğlu and S. Sen. Understanding the Influences of Past Experience on Trust in Human-agent Teamwork. *ACM Transactions on Internet Technology*, 19(4), 11 2019. doi: 10.1145/3324300

[9] S. Halan, I. Sia, A. Miles, M. Crary, and B. Lok. Engineering social agent creation into an opportunity for interviewing and interpersonal skills training. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 2018.

[10] P. A. Hancock, D. A. Vincenzi, J. A. Wise, and M. Mouloua, eds. *Human Factors in Simulation and Training*. CRC Press, 12 2008. doi: 10.1201/9781420072846

[11] R. Hughes and E. Ortiz. Medication Errors: Why they happen, and how they can be prevented. *The American Journal of Nursing*, 105(3):14–24, 2005.

[12] T. Jensen, Y. Albayram, M. M. H. Khan, M. A. A. Fahim, R. Buck, and E. Coman. The Apple Does Fall Far from the Tree. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. ACM, New York, NY, USA, 6 2019. doi: 10.1145/3322276.3322349

[13] K. Johnsen, R. Dickerson, A. Raij, B. Lok, J. Jackson, M. Shin, J. Hernandez, A. Stevens, and D. S. Lind. Experiences in using immersive virtual characters to educate medical communication skills. In *Proceedings - IEEE Virtual Reality*, 2005. doi: 10.1109/vr.2005.33

[14] S. Kalyanaraman and J. D. Ivory. Enhanced Information Scent, Selective Discounting, or Consummate Breakdown: The Psychological Effects of Web-Based Search Results. *Media Psychology*, 12(3), 8 2009. doi: 10.1080/15213260903052232

[15] L. Kohn, J. Corrigan, and M. Donaldson. *To Err Is Human*. National Academies Press, Washington, D.C., 3 2000. doi: 10.17226/9728

[16] P. Kulms and S. Kopp. More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human–agent cooperation. In *ACM International Conference Proceeding Series*, 2019. doi: 10.1145/3340764.3340793

[17] B. Lok. Training with virtual operating room teammates to influence team behaviors. In *Proceedings - 2016 International Conference on Collaboration Technologies and Systems, CTS 2016*, 2016. doi: 10.1109/CTS.2016.115

[18] M. A. Makary and M. Daniel. Medical error—the third leading cause of death in the US. *BMJ*, 5 2016. doi: 10.1136/bmj.i2139

[19] W. Mehtsun, A. Ibrahim, M. Diener-West, P. Pronovost, and M. Makary. Surgical never events in the United States. *Surgery*, pp. 465–472, 2013. doi: 10.1016/j.surg.2012.10.005

[20] R. Parasuraman and V. Riley. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2), 2 1997. doi: 10.1518/001872097778543886

[21] R Core Team. R: A Language and Environment for Statistical Computing. 2017.

[22] B. Sittner, M. Aebersold, J. Paige, L. Graham, A. P. Schram, S. Decker, and L. Lioce. INACSL Standards of Best Practice for Simulation: Past, Present, and Future. *Nursing Education Perspective*, 36(5):294–298, 2015. doi: 10.5480/15-1670

[23] R. Skarbez, A. Kotranza, F. P. Brooks, B. Lok, and M. C. Whitton. An initial exploration of conversational errors as a novel method for evaluating virtual human experiences. In *2011 IEEE Virtual Reality Conference*. IEEE, 3 2011. doi: 10.1109/VR.2011.5759489

[24] J. Stefaniak and C. Turkelson. Does the Sequence of Instruction Matter During Simulation? *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, 9(1):15–20, 2014. doi: 10.1097/SIH.0b013e3182a8336f

[25] J. Stuart, K. Aul, M. D. Bumbach, A. Stephen, and B. Lok. Building a Handoff Communication Virtual Experience for Nursing Students Using Virtual Humans. *CIN: Computers, Informatics, Nursing*, 5 2021. doi: 10.1097/CIN.0000000000000760

[26] S. Thampi, C. C. M. Lee, R. V. Agrawal, B. Ashokka, L. K. Ti, S. Paranjothy, and G. G. Ponnamperuma. Ideal Sequence of Didactic Lectures and Simulation in Teaching Transesophageal Echocardiography Among Anesthesiologists. *Journal of Cardiothoracic and Vascular Anesthesia*, 34(5):1244–1249, 2020. doi: 10.1053/j.jvca.2019.12.011

[27] I. Torre, J. Goslin, L. White, and D. Zanatto. Trust in artificial voices. In *Proceedings of the Technology, Mind, and Society*. ACM, New York, NY, USA, 4 2018. doi: 10.1145/3183654.3183691

[28] J. Van Den Bos, K. Rustagi, T. Gray, M. Halford, E. Ziemkiewicz, and J. Shreve. The $17.1 Billion Problem: The Annual Cost Of Measurable Medical Errors. *Health Affairs*, 30(4), 4 2011. doi: 10.1377/hlthaff.2011.0084

[29] N. Wang, D. V. Pynadath, and S. G. Hill. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 3 2016. doi: 10.1109/HRI.2016.7451741

[30] T. L. Warholak, C. Queiruga, R. Roush, and H. Phan. Medication Error Identification Rates by Pharmacy, Medical, and Nursing Students. *American Journal of Pharmaceutical Education*, 75(2), 3 2011. doi: 10.5688/ajpe75224

[31] C. White, J. Chuah, A. Robb, B. Lok, S. Lampotang, D. Lizdas, J. Martindale, G. Pi, and A. Wendling. Using a Critical Incident Scenario With Virtual Humans to Assess Educational Needs of Nurses in a Postanesthesia Care Unit. *Journal of Continuing Education in the Health Professions*, 35(3), 2015. doi: 10.1002/chp.21302

[32] L. Whitehair, S. Provost, and J. Hurley. Identification of prescribing errors by pre-registration student nurses: a cross-sectional observational study utilising a prescription medication quiz. *Nursing Eduction Today*, 2013. doi: 10.1016/j.nedt.2012.12.010

[33] M. X. Zhou, G. Mark, J. Li, and H. Yang. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems*, 2019. doi: 10.1145/3232077