Game design elements in serious games: Analysis of a chatbot for medical education

- Kumulative Arbeit -

Inaugural-Dissertation

zur Erlangung der Doktorwürde

der

Philosophischen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität

zu Bonn

vorgelegt von

Alexandra Aster

aus

Neuss

Bonn, 2026

Gedruckt mit der Genehmigung der Philosophischen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Prof. Dr. Fani Lauermann
(Vorsitzende)

Prof. Dr. Tobias Raupach
(Betreuer und Gutachter)

Prof. Dr. Ulrich Ettinger
(Gutachter)

Prof. Dr. André Beauducel
(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 03. November 2025

Table of contents

Summary

Serious games are defined as games pursuing a primary objective beyond mere entertainment. Serious games are often used within educational contexts to foster intrinsic motivation in students, enhance their learning experience and thus their learning outcomes. Responsible for affecting users' intrinsic motivation are the inherent characteristics of games, known as game design elements. A frequently referenced motivational theory underlying serious games is the self-determination theory. It proposes that the three basic psychological needs for autonomy, competence, and social relatedness have to be addressed to enhance intrinsic motivation. Previous literature has already attempted to assign specific game design elements to the respective needs they address. However, the list of possible game design elements can be arbitrarily augmented with further elements. Serious games are already used in different educational settings, especially in the realm of medical education. They are known to be safe environments for training a wide range of skills and competencies allowing for a training without posing real-life risks to, e.g., patient safety. Training history taking is an integral part in the medical curriculum and serious games with their safe character lend themselves for training it in a risk-free environment. A chatbot for training history taking could serve as a training tool embedded in a serious game. Therefore, this thesis aimed to examine the theoretical foundation of using a chatbot for medical history taking to extend the list of theory-based game design elements with the element chatbot.

Following the proposed research framework, a first study developed a serious game with an inherent retrieval-based chatbot relying on free open-text entries and conducted a first usability and user experience evaluation. Building on this foundation, the second study compared the newly established chatbot with a previously existing keyword-based retrieval-based chatbot. For assessing autonomy, two measurement methods were fused. The objective autonomy in terms of students' exploratory behaviour was assessed during history taking in terms of the serious games' process data while the subjective autonomy was assessed via questionnaires after the session. Results indicated that the chatbot relying on free open-text entries allowed

for more exploratory behaviour while no significant differences were found for the subjective autonomy measures. The history taking data was analysed with regard to what students asked their virtual patients but not how their questions were phrased. A third study was conducted focusing on the effectiveness of material for self-directed learning on history taking in chatbots. Therefore, the two chatbots of the second study were used and it was assessed whether students profited from an interposed guideline between two sessions. Results indicated that students using the chatbot with free open-text entries achieved overall but not significantly higher scores for their history taking in both sessions. However, only students using the keyword-based retrieval-based chatbot improved significantly between the sessions. A fourth study was conducted, in which students took histories with a generative chatbot and their entries were analysed with focus on indicators of empathy. Additionally, students' perception on subjective autonomy was assessed. The results of the fourth study indicated that the generative chatbot enabled students to show written empathic reactions and reported high levels of autonomy afterwards. However, due to the small sample size these results have to be regarded as preliminary.

In conclusion, this dissertation provided initial findings on the theoretical foundation of chatbots as game design elements. Building on the definitions of game design elements, it can be assumed that a chatbot can be regarded as a game design element due to its interactive character. Chatbots have to be chosen based on the context they are used in, yet it is worth mentioning that the more degrees of freedom a chatbot has the more it is able to address students' subjective autonomy. It is further discussed whether other needs, such as the need for competence, may be addressed and ideas for consequent research are proposed.

Keywords: Serious game; Game design element; Chatbot; Medical education

Acknowledgements

I am sincerely grateful to Prof. Dr. Tobias Raupach for opening the door to this doctoral project and for his steady support, open discussions, and thoughtful input at every step of the journey. Additionally, I would like to thank Prof. Dr. Ulrich Ettinger for his initial encouragement and for making the beginning of this journey possible. Especially, I would like to thank Matthias Carl Laupichler, who has supported me from day one, not only in completing this doctoral thesis but also in finishing two half marathons. Thanks to all my current and former colleagues at the Institute of Medical Education of the University Hospital Bonn, as well as to all my co-authors. Without your engagement, this project and all its studies would not have come to life.

Special thanks to all my friends, friends-in-law, my family, and family-in-law. Thank you for always believing in me and never growing tired of asking how things were going. Moreover, my gratitude extends to all the companions along the way who have supported me not only with encouraging conversations but also with music.

Lastly, I have to express my deepest gratitude to my wife, Annika. Thank you for being my lighthouse in the dark and for guiding me safely through troubled waters. I truly would not have made it to this point without you. This is not only mine but also yours.

Glossary

AI	Artificial Intelligence
ChatGPT	Chatbot Generative Pre-trained Transformer
GATE framework	Game Design Elements, Theory, Effectiveness framework
GBL	Game-Based Learning
GDE	Game Design Element
LLM	Large Language Model
SDT	Self-Determination Theory
SG	Serious Game
SP	Simulated Patient
VP	Virtual Patient

List of Tables

List of Figures

List of references for Studies 1 to 4

Study 1

Aster, A., Hütt, C., Morton, C., Flitton, M., Laupichler, M.C., & Raupach, T. (2024). Development and evaluation of an emergency department serious game for undergraduate medical students. *BMC Medical Education*, *24*(1), Article 1061. https://doi.org/10.1186/s12909-024-06056-z

Study 2

Aster, A., Lotz, A., & Raupach, T. (2025). Theoretical background of the game design element "chatbot" in serious games for medical education. *Advances in Simulation, 10*(1), Article 10. https://doi.org/10.1186/s41077-025-00341-7

Study 3

Aster, A., Lotz, A., Laupichler, M.C., & Raupach, T. (2025). Impact of providing a customized guideline on virtual medical history taking in two serious games for medical education. *Medical Education Online, 30*(1), Article 2527175. https://doi.org/10.1080/10872981.2025.2527175

Study 4

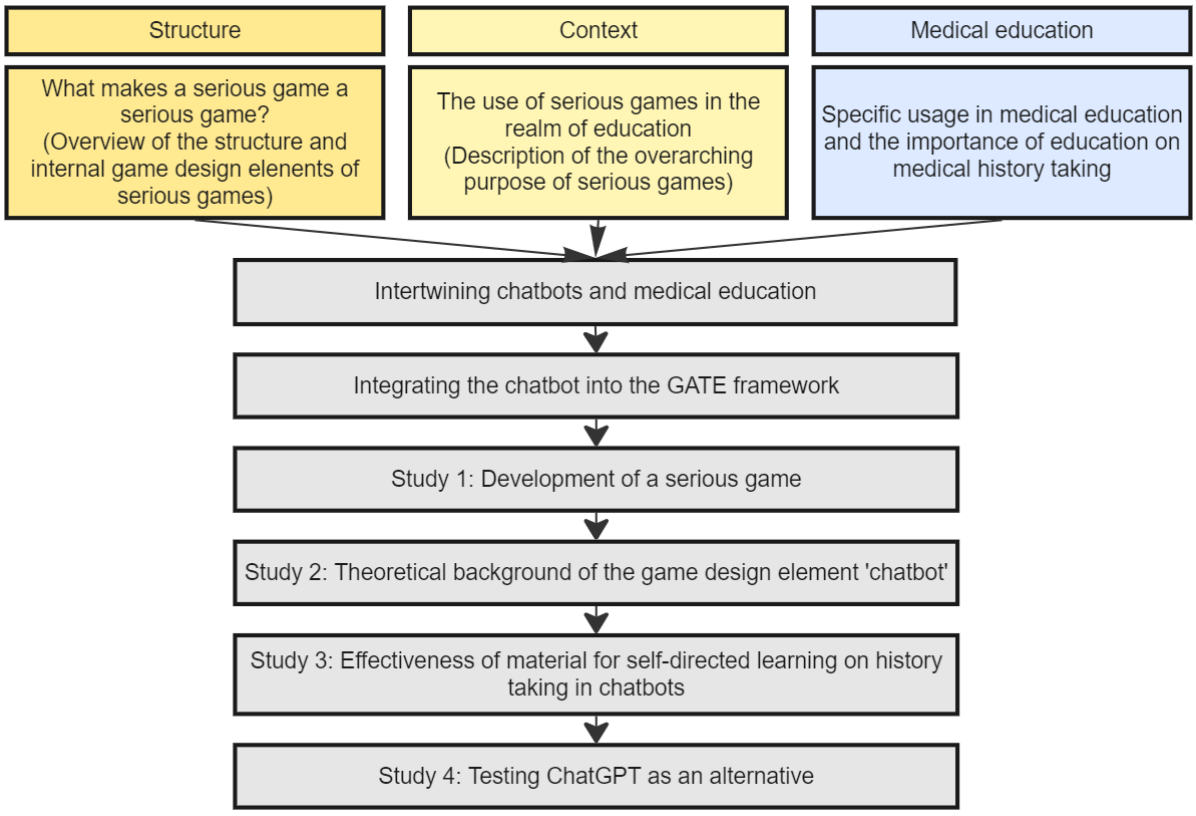
Aster, A., Ragaller, S.V., Raupach, T., & Marx, A. (2025). ChatGPT as a virtual patient: Written empathic expressions during medical history taking. *Medical Science Educator, 35*(3), 1513-1522. https://doi.org/10.1007/s40670-025-02342-7

1 Introduction

The thesis follows several parallel strands to adequately integrate the research into the proposed framework. It starts with setting the superordinate structure by defining the nature of serious games (SG), defining their general structure as well as their inner life reflected by game design elements (GDE). Secondly, insights in the contexts in which SGs can be used are provided with examples for the application in fields distant to medical education as well as more closely associated with medical education. Thirdly, the use of SGs in medical education and the importance of education on medical history taking are specified. Merging the strands results in the elaboration of chatbots in general and in medical education in particular. Before the four studies, which constitute the core of the thesis are discussed, the GDE chatbot will be categorised in the Game Design Elements, Theory, Effectiveness framework (GATE). Figure 1 illustrates the outline of this thesis.

Figure 1

Depiction of the outline of the thesis



2 What makes a serious game a serious game?

According to Botella et al. (2011) the term *serious game* was initially associated with educational games in a book by Abt (1970), who described SGs as games following an educational objective beyond being primarily entertaining. This aligns with the later proposed definition by Michael and Chen (2005), particularly in the aspect that a SG mainly pursues a predefined learning objective and therefore exceeds mere entertainment. Although not explicitly stated, it can be assumed that the learning environment in a SG simultaneously evocates enjoyment while reaching the learning objective. Abt (1970) referred to analogue SGs, while in the current use SGs are mostly referred to as digital games (Dörner et al., 2016).

To date, there is no uniformly stated definition of SGs yet researchers adhere to different definitions which resemble one another. The afore mentioned definition by Michael and Chen (2005) can be recognised as a common, if not the most commonly cited definition for SGs and will therefore be used as a fundament in this thesis. However, this definition is lacking crucial aspects such as an elaboration of the question whether commercial games that are mostly used for recreation should be used as SGs for educational purposes. Although entertainment games are also assigned to the category of SGs by some authors (Susi et al., 2007), using commercial games for educational purposes may be better included under the umbrella term *game-based learning* (GBL) as already assigned in earlier research (Qian & Clark, 2016). Therefore, a new working definition was developed, incorporating the criterion that SGs should be specifically designed for educational purposes. Conclusively, the working definition reads as follows: Serious games are non-commercial, specifically designed for educational purposes helping learners to reach predefined learning objectives while incidentally inducing enjoyment.

While SGs can be examined from technical to psycho-educational perspectives (Krath et al., 2021), this thesis adopts a psycho-educational approach.

2.1 Two concepts going hand in hand or being fundamentally different? – Distinguishing the concepts gamification and serious game

A lack of conceptual clarity in the literature is reflected in examples such as the interchangeable use of GBL and gamification (e.g., Hartt et al., 2020). Due to ambiguities as heterogeneously use of the term gamification, it is difficult to rely on a consistent body of literature (Seaborn & Fels, 2015). Thus, clear distinctions and consistent use of these terms throughout the literature are important. In the following, the concepts' definitions will be discussed. The chapter concludes with a synthesised conceptualisation of existing definitions, highlighting both the relationships and distinctions among the concepts.

Overarching concept: Game-based learning

The concept of GBL can be understood as an overarching concept, in the sense of an umbrella term subsuming the concept of learning with games. When synthesising the most frequently cited definitions for GBL (Emerson et al., 2020; Higham & Guzel, 2012; Krath et al., 2021; Plass et al., 2015; Qian & Clark, 2016; Wu et al., 2011), it becomes evident that regardless of their elaborateness and differences, all definitions have one thing in common: applying a game for learning purposes.

Krath et al. (2021) explicitly stated SGs as an operationalisation of GBL, while differentiating gamification as a distinct concept due to it solely using game elements and not full-fledged games. Hamari and Keronen (2017) called it "instrumental purpose" (p. 126) when games are used aside from leisure purposes and name SGs, simulations, as well as gamification as examples for using games for instrumental purposes. Early ideas of GBL primarily focused on the use of commercial games for educational purposes (Gee, 2003, 2005; Prensky, 2001). Gee (2005) provided explicit examples of how different learning principles apply in different commercial games and how they could be used for educational purposes accordingly. However, these remarks have to be viewed critically as they are not based on theoretical foundations but on deliberations and should be therefore rather seen as recommendations (Krath et al., 2021).

Specific application: Gamification

Gamification can best be defined as "the use of game design elements in non-game contexts" (Deterding et al., 2011, p. 9) or, more specifically, as "the intentional use of game elements for a gameful experience of non-game tasks and contexts" (Seaborn & Fels, 2015, p. 17). These two definitions provide insight into what can be understood as gamification, namely the addition of GDEs to non-gaming contexts. One example that accurately illustrates the differentiation between GBL and gamification can be found in Plass et al. (2015). The authors used the example of a mathematics homework. In the context of gamification, it would be equipped with additional GDEs that reward the learner for finishing the homework. To be used in the context of GBL, the homework has to be transformed in a more game-like activity by adding conflicts or rules that influence the execution of the homework.

Concrete Operationalisation: Serious games

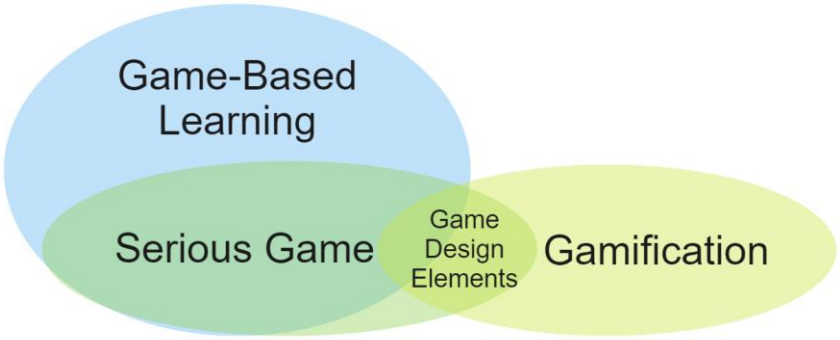
The concepts gamification and SG differ in the manner of learning environment. While gamification solely adds GDEs to non-gaming contexts (Deterding et al., 2011), SGs are fully-fledged games that address at least one learning objective (Michael & Chen, 2005). In the already mentioned definition by Michael and Chen (2005) SGs are defined as "games that do not have entertainment, enjoyment, or fun as their primary purpose" (p. 21).

Alongside the frequently cited definition by Michael and Chen (2005), there are several other definitions to be considered. According to Susi et al. (2007) all definitions for SGs share the common ground of focusing on educational aims instead of entertainment purposes and are more or less precise in stating how much entertainment can be simultaneously experienced during playing a SG. The authors further argued, that, in certain cases, entertainment games can function as SGs depending on the context of use (Susi et al., 2007). Similarly, a SG is not inherently a SG; its perception depends on the circumstances, and it may occasionally be perceived as an entertainment game (Susi et al., 2007). Following the definition by Michael and Chen (2005), a SG needs to have a learning objective and it can be assumed that

entertainment games are lacking this since their primary aim was to induce fun and enjoyment. Figure 2 illustrates the conceptualisation of the three constructs as derived from the literature.

Figure 2

Illustration of the conceptualisation of the three constructs GBL, SG, and gamification



Note. GDEs are borrowed elements from gamification that are used within serious games. Therefore, this can be understood as the overlapping point between both constructs although they have to be generally considered as distinct and demarcated.

In conclusion, as can be seen in Figure 2, the concepts SG and gamification go hand in hand by sharing the overlapping aspect of GDEs which SGs adopt from gamification. Nonetheless, they differ in their structure and application as further explained in Chapter 3.

2.1.1 Theoretical underpinnings of serious games

Now that SGs are defined, it is important to introduce the theoretical foundation of those games, since a thorough theoretical basis is connected with positive outcomes (Wu et al., 2012). Extensive work has been conducted to explore the theoretical underpinnings of SGs, GBL, and gamification (e.g., Krath et al., 2021; Qian & Clark, 2016; Sailer et al., 2017). Due to the ambiguity in defining and demarcating the previously mentioned constructs, it is similarly difficult to determine respective underlying theories for each construct. Despite those ambiguities, all three constructs go to some extent hand in hand. The following section provides insights into the current body of literature on underlying theories for SGs irrespective of their application area. In case theories apply specifically to SGs, they are highlighted.

Qian and Clark (2016) analysed in their literature review the applied learning theories for the overarching construct GBL and found that 76% of the included studies reported theories for

the game design or the research design with *constructivism* being the most frequently mentioned. However, it is not clear whether the theories were all applied to the game design or solely for the research design. In their systematic review, Krath et al. (2021) extensively explored the theoretical basis of gamification, GBL and SGs. Although the authors considered these constructs together, they found intriguing results on the use of a plethora of theories. In sum, they found 118 theories and allocated them to the four areas of “Affect and Motivation”, “Behaviour”, “Learning”, and “Other” (p. 10). In the first category “Affect and Motivation” the *Self-Determination Theory* (SDT; Deci & Ryan, 1985; Deci & Ryan, 2000) was used the most frequently, directly followed by the *Flow Theory* (Csikszentmihalyi, 1975). Both theories were not only the most frequently referenced theories within this category but across all categories throughout all underlying studies. The second category “Learning” comprised theories that were all generally cited less often than the SDT or the flow theory. Some examples falling into this category are: *Cognitive Load Theory* (Sweller, 2010), *Constructivist Learning Theory* (Jonassen, 1999), or *Experiential Learning Theory* (Kolb, 1984). The third category “Behaviour” comprised theories that were even less cited than the theories mentioned in the previous categories. Examples falling into this category are the *Theory of Planned Behaviour* (Ajzen, 1991), or the *Reinforcement Theory* (Moore, 2011; Skinner, 1953). Explicitly naming and stating all of the named theories in the review and being more precise about the mentioned theories, would exceed the scope of this thesis. As the SDT was mentioned most frequently in the underlying studies and also aligns with the later stated research framework for this thesis, the following explanations will focus on this theory and explain it in more detail.

The self-determination theory initially established by Deci and Ryan (1985, 2000) is a psychological theory focusing on motivation in humans and originally arose from the research on why and how external rewards compromise intrinsic motivation (Deci & Ryan, 2012). It states that the fulfilment of the three basic psychological needs for autonomy, competence, and social relatedness enhance intrinsic motivation depending on (social) environmental factors (Deci & Ryan, 2012). The need for autonomy is addressed when one feels their own

behaviour can be executed in a self-determined and volitional manner (Niemiec & Ryan, 2009). The need for competence relates to the subjective feeling of successfully executing a (goal-directed) behaviour (Niemiec & Ryan, 2009). The last basic psychological need, the need for relatedness, refers to humans' inherent need to belong and to be connected to significant others (Ryan & Deci, 2000). To date, the overarching SDT entails six sub-theories, each focusing on different aspects of motivation (Ryan & Deci, 2017). The first sub-theory, namely *Cognitive Evaluation Theory*, focused only on autonomy and competence and defined external factors (e.g., rewards or feedback) that increase or decrease intrinsic motivation (Deci & Ryan, 2012). Furthermore, social environmental factors were characterised that support autonomy or have a controlling effect as well as their influences on intrinsic motivation (Deci & Ryan, 2012). Lastly, it states how social environmental factors moderate intrinsic motivation by the interaction with external factors (Deci & Ryan, 2012). It is reasonable to assume that the cognitive evaluation theory informed GDE research. Further complementary sub-theories are the following: *Causality Orientations Theory* (focusses on individual differences), *Organismic Integration Theory* (focusses on the internalisation of extrinsic motivation), *Basic Psychological Needs Theory* (focusses on the effects of needs satisfaction on psychological wellbeing), *Goal Content Theory* (focusses on pursuits and life goals), and *Relationships Motivation Theory* (focusses on the effect of close relationships) (Ryan & Deci, 2017). All theories combined result in the content of the overarching SDT.

Knowledge about social environmental factors that affect intrinsic motivation is especially relevant for educators or others aiming to motivate people as this knowledge enables them to adapt the contexts accordingly in order to promote an environment that fosters intrinsic motivation (Ryan & Deci, 2000). These changes cannot only be conducted through basic changes in the environment but also through adding GDEs as they can alter the earlier mentioned social environmental factors (Alexiou & Schippers, 2018). Ryan and Deci (2000) stated that "[…] people will be intrinsically motivated only for activities that hold intrinsic interest for them, activities that have the appeal of novelty, challenge, or aesthetic value." (p. 71). In

this sense, SGs including their GDEs, which originally stem from gamification, can be regarded as such activities yielding the stated values. Educators may enable the satisfaction of the learners' basic psychological needs by skilfully changing the environmental factors through embedding gamification elements in non-gaming contexts or by using SGs with deliberately chosen theory-based GDEs. Although there has already been research conducted combining gamification with the cognitive evaluation theory (Hsu, 2022), research concerning SGs however focused on the SDT in its entirety (e.g., Krath et al., 2021).

In particular, GDEs can be viewed as the carrier medium for motivation within SGs as they can address different motivational factors in the player (Plass et al., 2015), e.g., by addressing the basic psychological needs (Sailer et al., 2017).

2.2 Collecting points to climb the leaderboard – Embedding game design elements in serious games

Seaborn and Fels (2015) reviewed several definitions of games and concluded that games commonly share the aspects of "rules, structure, voluntariness, uncertain outcomes, conflict, representation, [and] resolution" (p. 16), although these aspects may not be equally pronounced and can appear in various combinations. According to Plass et al. (2015) games are mostly structured consisting of "a challenge, a response, and feedback" (p. 262) with feedback facilitating a circle by emerging new challenges or requiring altered responses. Seaborn and Fels (2015) build on the work by Deterding et al. (2011) for defining GDEs as "[…] the pieces that comprise the game – dynamics, mechanics, components […]." (p. 17). Therefore, following Deci and Ryan (2012), GDEs can be regarded as extrinsic features affecting intrinsic motivation. GDEs have already been used effectively in entertainment games to facilitate learning and have been proven to enhance players' motivation and involvement (Qian & Clark, 2016). Conclusively, GDEs are relevant and needed in SGs for several reasons. On the one hand, GDEs transport the learning contents and are therefore relevant for gaining knowledge and practicing skills (Arnab et al., 2014). On the other hand, they are the core for making the learning experience playful (Plass et al., 2015). Several researchers have designed

different frameworks attempting to ease the classification and application of GDEs in different contexts (Alexiou & Schippers, 2018; Arnab et al., 2014; Blohm & Leimeister, 2013; Plass et al., 2015).

Characterization of different GDEs

The GDEs used in educational games and in GBL in general, do not only originate from the field of entertainment games (Arnab et al., 2014), but also from behaviourism, cognitivism, and constructivism (Plass et al., 2015). GDEs can compile "patterns, objects, principles, models, and methods" (Seaborn & Fels, 2015, p. 17) that can be summarized in categories as mentioned before. However, neither defining frameworks for the compilation of GDEs nor the specification of GDEs is straightforward (Seaborn & Fels, 2015). Different endeavours to compiling lists of GDEs have been conducted, but due to the heterogeneity of GDEs, they cannot be regarded as exhaustive (Sailer et al., 2013). Nonetheless, there are GDEs throughout the literature that are used and referenced more frequently than others. In the following, an overview of the most commonly known GDEs along with their definitions is presented (see Table 1). Attempting to establish an exhaustive list of GDEs would extend the scope of this thesis. Due to the heterogeneous nature of GDEs and them stemming from the area of entertainment games, the list can be arbitrarily extended.

Table 1

List of GDEs along with their definition

GDE	Definition	Source
Points	"Numerical units indicating progress" (p. 20) that also serve as feedback	Seaborn and Fels (2015) Sailer et al. (2017)
Badges	"Visual icons signifying achievements" (p. 20)	Seaborn and Fels (2015)
Leaderboards	"Display of ranks for comparison" (p. 20) "Combination of assessment, conflict/change, and rules/goals" (p. 758)	Seaborn and Fels (2015) Landers (2015)
Progression	"Milestones indicating progress" (p. 20)	Seaborn and Fels (2015)

GDE	Definition	Source
Status	“Textual monikers indicating progress” (p. 20)	Seaborn and Fels (2015)
Levels	“Increasingly difficult environments” (p. 20)	Seaborn and Fels (2015)
Rewards	“Tangible, desirable items” (p. 20)	Seaborn and Fels (2015)
Roles	“Role-playing elements of character” (p. 20)	Seaborn and Fels (2015)
Performance graphs	Display of “the player’s own performance over time” (p. 373) without social comparison to other players	Sailer et al. (2017)
Meaningful stories / Storyline	Narrative of the game that is detached from player’s performance	Sailer et al. (2017)
Avatars	Self-generated or selectable virtual player representations	Sailer et al. (2017)
Teammates	Characters within the game, creating cooperative or competitive environments	Sailer et al. (2017)
Collaboration	Teamwork between fellow players	Aster, Laupichler, et al. (2024)
Competition	“Competing against each other individually or in groups” (p. 1831)	Aster, Laupichler, et al. (2024)
Feedback	Acknowledgement of players’ actions in verbal, numerical, or visual form	Aster, Laupichler, et al. (2024)
Time limit	“Time restriction on gameplay” (p. 1831)	Aster, Laupichler, et al. (2024)
Additional life	“Receiving additional life, e.g. as a reward” (p. 1831)	Aster, Laupichler, et al. (2024)
Hints / Tips	“Support during game play” (p. 1831)	Aster, Laupichler, et al. (2024)

Note. List of the most common or best known GDEs along with their definitions and the respective references of which the definitions were extracted.

2.2.1 Which one to choose? – The theoretical foundation of game design elements

It is crucial to avoid overloading the SG with GDEs. Therefore, the most relevant GDEs in accordance with the respective learning objective have to be chosen to support the required learning needs (Plass et al., 2015). As described in Chapter 2.1.1., the SDT is often used as the theoretical foundation of SGs. Given that GDEs are an integral part of games, it is likely that the same theoretical foundation likewise pertains to them. GDEs are the key factors in games, and GBL in general, which lead to engagement (Plass et al., 2015). Another addressed factor besides engagement is motivation and from studies of entertainment games it is already known that GDEs address motivation in players (Qian & Clark, 2016). Accordingly this has also been incorporated into the research of educational games and different authors have assigned specific GDEs to the needs of the SDT (Alexiou & Schippers, 2018; Aparicio et al., 2012; Sailer et al., 2017). Table 2 compiles different GDEs, their addressed needs, the respective references the information were extracted from along with the manner how the authors of the primary studies derived the relationship between the GDEs and their respective mentioned theories. In this sense, the term *empirical* refers to empirical derivations, while *theoretical* refers to a theoretical guided derivation, and *intuitive assumption* describes a derivation of needs based on the authors' deliberations. As with Table 1, Table 2 also lists some exemplary GDEs and does not aim to exhaustively cover the literature. The vast majority of studies reported that GDEs affect intrinsic motivation, yet there are also studies that suggest that extrinsic motivation is addressed by certain GDEs, such as points, leaderboards and levels (Mekler et al., 2017). This is in line with earlier research on motivation, as Deci and Ryan (2012) showed that some external factors are able to increase intrinsic motivation (such as feedback or choice) while others decrease it by being attainable (such as external rewards or competition).

Because GDEs play an important role in the achievability of learning objectives, it is essential to specify the concrete learning objectives the SG aims to achieve and to choose the appropriate GDEs accordingly and thoughtfully during the design phase (Plass et al., 2015). SGs directly target learning outcomes while also addressing motivation and engagement,

whereas gamification influences learning indirectly by altering motivation or engagement (Landers, 2015).

The research on GDEs mostly stems from the area of gamification (e.g., Seaborn & Fels, 2015) and is often not directly conducted in SGs. However, GDEs are frequently embedded in SGs without theoretical foundations or effectiveness testing. Since both constructs refer to the same GDEs, it can be helpful to consider research on gamification when researching SGs (Landers, 2015). In this sense it can be considered that the GDEs in SGs address the needs from the SDT to enhance students' motivation (i.e., change of behaviour or attitudes in the sense of the theory of gamified learning) which in turn affects the impact of the instructional content on the learning outcome, but whether this coherence is mediating or moderating depends on the respective context (Landers, 2015).

Table 2

List of GDEs and the addressed needs along with their respective reference

GDE	Addressed Need of the SDT	Derivation	Reference
Points	Competence	Intuitive assumption	Aparicio et al. (2012)
Levels	Competence	Intuitive assumption	Aparicio et al. (2012)
Performance graph	Competence	Empirical	Sailer et al. (2017)
Badges	Competence	Empirical	Sailer et al. (2017)
Leaderboards	Competence	Empirical	Sailer et al. (2017)
		Intuitive assumption	Aparicio et al. (2012)
Progressive information	Competence	Intuitive assumption	Aparicio et al. (2012)
Intuitive controls	Competence	Intuitive assumption	Aparicio et al. (2012)

GDE	Addressed Need of the SDT	Derivation	Reference
Challenge	Competence	Theoretical	Alexiou and Schippers (2018)
		Intuitive assumption	Aparicio et al. (2012)
Feedback	Competence	Theoretical	Alexiou and Schippers (2018)
Positive feedback	Competence	Intuitive assumption	Aparicio et al. (2012)
Avatars	Autonomy	Empirical	Sailer et al. (2017)
		Intuitive assumption	Aparicio et al. (2012)
Profiles	Autonomy	Intuitive assumption	Aparicio et al. (2012)
Macros	Autonomy	Intuitive assumption	Aparicio et al. (2012)
“Increased levels of perceived control” (p. 2561) (no GDE in the narrower sense)	Autonomy	Theoretical	Alexiou and Schippers (2018)
Control of privacy and / or notification	Autonomy	Intuitive assumption	Aparicio et al. (2012)
Alternative activities	Autonomy	Intuitive assumption	Aparicio et al. (2012)
“Opportunities for self-representation and expression” (p. 2561) (no GDE in the narrower sense)	Autonomy	Theoretical	Alexiou and Schippers (2018)

GDE	Addressed Need of the SDT	Derivation	Reference
Configurable interface	Autonomy	Intuitive assumption	Aparicio et al. (2012)
Meaningful stories / Storyline	Autonomy & social relatedness	Empirical	Sailer et al. (2017)
Teammates	Social relatedness	Empirical	Sailer et al. (2017)
Online communities	Social relatedness	Theoretical	Alexiou and Schippers (2018)
Groups	Social relatedness	Intuitive assumption	Aparicio et al. (2012)
Messages	Social relatedness	Intuitive assumption	Aparicio et al. (2012)
Blogs	Social relatedness	Intuitive assumption	Aparicio et al. (2012)
Connection to social networks	Social relatedness	Intuitive assumption	Aparicio et al. (2012)
Chat	Social relatedness	Intuitive assumption	Aparicio et al. (2012)

Note. Derivation refers to the manner how the referenced studies derived the related needs for the respective GDEs with empirical referring to empirical derivations, theoretical referring to theoretical guided derivation, and intuitive assumption to the derivation of needs based on the authors' deliberations.

3 The use of serious games in the realm of education

Now that the foundation for understanding the concept of SGs is laid, a detailed look can be taken at the usage of SGs in different educational settings.

3.1 General educational usage

Within their respective fields, SGs provide safe learning environments where failure and mistakes can happen without severe consequences (Plass et al., 2015; Susi et al., 2007). Since their invention, SGs have been frequently used learning environments which enable different groups to learn a plethora of differing contents within various teaching contexts. This

allows SGs to be used in countless settings. In the following, SGs in various contexts are described. Starting with SGs in fields distant and near to medical education before approaching SGs that are concretely used in medical education. All GDEs used in the respective SGs are underlined for improved readability and to support visual clarity.

3.1.1 Use of serious games in educational settings distant to medical education

Energy education is one exemplary field distant to medical education in which SGs are applied. The game “Energy Chicken” as described by Orland et al. (2014) was used to elicit changes in energy-consciousness of office workers in a “mid-size commercial office complex” (p. 44) in the U.S. aiming for more energy efficient behaviour. The real-life energy consumption of the players was mirrored in the health condition of chickens living on a customisable farm within the game. A reduction in real-life energy consumption lead to a good health of the chicken, where they grow and lay eggs, which in turn could be exchanged against accessories for the chickens or the farm. In this sense, the eggs can be understood as rewards while the accessories can be understood as awards or badges. In case the real-life energy consumption increases, the chickens get ill and their health declines. A graph shows an overview of the energy consumption throughout the playing time. Additionally, a “Mountain View” (p. 45) allows players to view other players’ farms and chickens from above, without providing an opportunity for interaction. The energy consumption was used as a measurement for change and playing the game led to a decreased energy use compared to baseline.

Whittaker et al. (2021) examined the effects of GDEs embedded in a SG for enhancing sustainability knowledge. Similar to the SG developed by Orland et al. (2014), this SG focused on promoting specific energy-saving behaviours in households in general. Whittaker et al. (2021) set up the two categories “reward-based game mechanics” and “meaningful game mechanics”. Points, badges, and trophies are considered “reward-based game mechanics” while the educational messages are considered as “meaningful game mechanics”. Throughout their session, players gained points by switching off lights, achieved badges by accomplishing given achievements, received trophies when having received a certain amount of badges, and

were provided with educational messages. The results showed that badges and trophies had a significant and direct effect on sustainability knowledge while points and educational messages had not. The authors concluded that the GDEs in the reward-based category can be further subdivided into higher and lower ranking GDEs with the significant GDEs being higher ranking GDEs and the not significant GDE being a lower ranking GDE. The authors assumed that since the higher ranking GDEs were awarded for more meaningful behaviour, they are more influential than lower ranking GDEs that are awarded for less meaningful behaviour. Furthermore, the authors concluded that the GDEs' influence stemmed from operant conditioning where learning is caused through rewarding or punishing for specific behaviour.

These exemplary SGs from educational settings distant to medical education demonstrate that SGs are equipped with GDEs regardless of their application area. It became apparent, that the "reward-based" GDEs (Whittaker et al., 2021) appeared across both SGs. The presented studies differed in their elaborateness. While Orland et al. (2014) focused on analysing the changes in behaviour after playing the game, Whittaker et al. (2021) explicitly analysed the influence of specific GDEs.

3.1.2 Use of serious games in educational settings close to medical education

Serious games do not only play a crucial role in educational settings distant to medical education but also in educational settings close to medical education, namely healthcare professions education (e.g., pharmacy, and nursing, among others). In a previously published systematic review I have identified studies focusing on SGs in healthcare professions education (Aster, Laupichler, et al., 2024) and will refer to some of them in the following along with other studies.

In the field of nursing, a recent systematic review and meta-analysis revealed that more SGs are concentrating on the training of practical skills than imparting theoretical knowledge (Lee et al., 2024). Even if most variables of interest referred only to level 1 or level 2 of Kirkpatrick's evaluation model (Kirkpatrick & Kirkpatrick, 2016) they showed positive results (Lee et al.,

2024). The distribution that most variables of interest were found on level 1 or 2 and few to none were found on level 3 or 4 also aligns with other reviews conducted in the realm of healthcare professions education (Aster, Laupichler, et al., 2024). An example of a SG for nursing students arose during the COVID-19 pandemic and teaches different infection prevention behaviours (Calik et al., 2022). Although Calik et al. (2022) mentioned to have based their game development on the learning mechanics – game mechanics (LM-GM) framework (Arnab et al., 2014) neither the specific structure of the game nor the GDEs were described. When analysing the included pictures, it became obvious that players chose an avatar and had to answer questions about diverse hygiene topics with varying graphics as answer options (Calik et al., 2022). Although the increase in knowledge was not significant after playing the SG compared to before, students reported feeling more confident in terms of patient safety afterwards (Calik et al., 2022).

Besides nursing, other disciplines have also recognised SGs as beneficial for teaching. Kayyali et al. (2021) designed a SG for nursing and pharmacy students to teach an improved application of the “British National Formulary” (p. 998). In contrast to the other previously described SGs, Kayyali et al. (2021) explicitly named the theory of player types by Bartle (Bartle, 1996) as the underlying theory for the selection of GDEs. Accordingly, achievements in form of medals, titles or ranks as well as scores on a public leaderboard were embedded to address killers and achievers (Kayyali et al., 2021). Kayyali et al. (2021) stated that the leaderboard addresses socializers; however, a function showing which other players are online was also embedded to address socializers. The last player type, explorer, was addressed by an interesting and relatable narrative within the SG. The GDEs were especially embedded and chosen to address the different player types. The SG entailed the opportunity to obtain clues, provided immediate feedback, clearly set the inherent goals and rules, and had a minimalistic aesthetic. The authors conducted only analysis on level 1 of Kirkpatrick’s evaluation model with usability testing and evaluation of students’ perceptions towards the game.

Two educational fields that require the training of practical and clinical reasoning skills similarly to medical education are physiotherapy training and dental education. Savazzi et al. (2018) developed a SG for training the clinical reasoning skills of qualified physiotherapists and physio therapy students for dual task rehabilitation. The game was designed as a virtual reality game in 3D where the player acted as a physiotherapist using the first-person perspective. Players were provided with a score and a conclusive feedback, but no other GDEs were presented. An evaluation of the game showed that players positively rated the user experience, and playing the game led to achieving the learning objectives as well as a decreased negative affect. Another SG concentrating on the training of clinical reasoning skills in dental students was developed by Wu et al. (2021). Throughout this game, players had to answer questions about several different dental educational topics to which the game provided the players with immediate feedback. Moreover, players were rewarded for choosing the right instruments. The referring study dealt with validating the game and assessing its usability.

In addition to the academic educational settings, patient education is also an area where SGs are frequently used. Thomas et al. (2023) reported about a SG for women with advanced breast or gynaecologic cancers to enhance their self-advocacy. The game's narrative contained different stories about avatar-women who were diagnosed with cancer with the players' decisions influencing the avatar. The game's aim was to help the avatar-women achieve self-advocacy and an improved quality of life by making several decisions throughout the game. Players received explicit and implicit feedback for their decisions. The study assessed the game's feasibility and acceptability among the players and computed a preliminary efficacy of using the game compared to a care-as-usual group. All results were positive in terms of feasibility and acceptability and a preliminary significant improved self-advocacy was measured between baseline and the 6 months follow-up. From a psychological point of view it can be argued that the game operates as observational learning in the sense that the avatars' actions or possible actions can be transferred and applied to the patient's own life (Klauß et al., 2024).

Despite these all being exemplary SGs, it becomes evident that the majority of games followed a basic design including a few, not theory-based, GDEs, while a minority of games reported a sophisticated and elaborated approach to design and GDEs. According to Whittaker et al. (2021), the mentioned GDEs within these SGs can be assigned to the category “reward-based game elements”.

3.2 Specific usage in medical education

The specific usage of SGs in medical education is an extensively and thoroughly researched field. Reviews often researched SGs in the broader realm of healthcare professions education. Yet, this chapter focuses on the specific usage of SGs in medical education and therefore, only reviews and studies that investigate undergraduate or postgraduate medical education are considered. In 2010, a best evidence medical education (BEME) guide was published which assessed the effectiveness of educational games in medical education with exclusion of all other healthcare professions and their respective education (Akl et al., 2010). Although the studies mostly indicated a positive impact of educational games, the systematic review led the authors to the conclusion that recommending the general use of educational games was not appropriate at this point due to the unsatisfactory methodological rigor of the five included studies (Akl et al., 2010). A later review concentrated solely on analogous SGs for medical education and stated that since studies mostly evaluated the outcomes on level 1 and level 2 of Kirkpatrick’s evaluation model (Kirkpatrick & Kirkpatrick, 2016), it cannot be assumed that those games can substitute traditional teaching scenarios but can supplement them (Edwards et al., 2025). Edwards et al. (2025) concluded that the SGs were used in a variety of medical specialties. Different contents are taught within SGs in medical education and they lend themselves for training hard skills as well as soft skills. Hard skills can generally be defined as quantifiable and therefore measurable abilities covering functional, task-specific aspects which can be more easily compared, while soft skills represent social, interprofessional, or interpersonal skills that are more abstract (Continisio et al., 2021; Klein et al., 2024). In the medical field, hard skills can comprise skills such as medical knowledge, patient care, or

practice-based learning, whereas soft skills can encompass skills such as interpersonal and communication skills, or professionalism (Klein et al., 2024). Owing to their game-like nature, SGs provide a safe environment for training various skills. In the following, some examples of SGs for the acquisition or training of knowledge and different skills are presented.

A SG for training laparoscopic surgery was found to have a higher impact on surgeons' skills than on medical students' skills but was perceived to be useful and realistic while no explicit GDEs or respective theories were mentioned (Kowalewski et al., 2017). Another SG in the field of surgical training focused on the situational awareness in terms of problem recognition and problem-solving competence for equipment-related issues during a minimally invasive laparoscopic surgical procedure (Graafland et al., 2017). Within the SG, players had to recognize and solve problems during a mini game that was not directly related to a surgical procedure (Graafland et al., 2017). However, a laparoscopic tower was embedded in the game where players had to recognize and solve real-life problems of a surgeon enhancing the closeness to reality (Graafland et al., 2017). The SG does not illustrate a specific skill training for the operating room but provides training of more abstract higher order skills like the ability to solve problems (Graafland et al., 2017). Although the authors did not explicitly state an underlying theory for the embedded GDEs, they stated the embedding of points and feedback and in a depiction of the SG it was also apparent that a time display as well as a deterioration bar were embedded, these were however not described in detail (Graafland et al., 2017). The evaluation of the SG showed that playing it led to an increased problem-solving ability regarding equipment-related issues in first- or second-year residents (Graafland et al., 2017).

In the field of emergency skill training a SG simulating an emergency department was developed for residents to train their emergency care skills by means of VPs (Dankbaar, Roozeboom, et al., 2017). The game focused on training emergency care skills by providing six patient cases where the player had the opportunity to conduct actions like in a real-life emergency department (Dankbaar, Roozeboom, et al., 2017). The authors did not mention any theoretical foundation for their game design but embedded several GDEs, such as direct

feedback throughout the game play, an introductory tutorial, a timer covering 15 minutes, narrative feedback and an individual score at the end of the game as well as a leaderboard depicting all players scores simulating competition (Dankbaar, Roozeboom, et al., 2017). Two groups of residents were compared for an effectiveness testing of which one group received only the course manual while the other in addition to the manual also played the game (Dankbaar, Roozeboom, et al., 2017). The effectiveness testing showed that the combined group had improved emergency care skills and motivation to play the game, whereas the motivation did not differ between the groups prior to the intervention (Dankbaar, Roozeboom, et al., 2017). After two weeks of face-to-face training the skill level did not differ between the groups anymore (Dankbaar, Roozeboom, et al., 2017).

Another SG in the area of emergency medicine is a strategy card-board game that was designed for learning how to manage multiple patients in an emergency department environment (Tsoy et al., 2019). The authors clarified the design and development process of their SG without explicitly mentioning any theoretical background for the selection of the five embedded GDEs, namely characters, goals, mechanics/resources, feedback, and challenge (Tsoy et al., 2019). Players selected characters to form a complete emergency department team working an entire shift together, eliciting collaboration between the players (Tsoy et al., 2019). Moreover, the players received points which could be swapped for extra resources or to win the game (Tsoy et al., 2019). The game was lost when too many violations of patient-safety occurred and was won when players achieved a predefined amount of points by the end of the shift (Tsoy et al., 2019). The game was not only played and evaluated by postgraduates but also by nurses (Tsoy et al., 2019). A limitation is that the evaluation did not contain an effectiveness test but only ratings of different perceptions of the game play, yet overall it was rated positively (Tsoy et al., 2019). Only a third of the players conducted actions which they would also conduct in a real-life emergency department what leads the authors to the conclusion that the game might not elicit a fully immersive game play (Tsoy et al., 2019).

A SG developed by Dankbaar, Richters, et al. (2017) trained patient safety knowledge in fourth-year medical students. For an effectiveness test, three groups were compared regarding altered patient safety knowledge and motivation. One group played the SG, while the second used an e-module and the third was a historical control group (Dankbaar, Richters, et al., 2017). The authors did not refer to a theoretical framework, neither for the development of the SG nor for the selection of specific GDEs. Dankbaar, Richters, et al. (2017) mentioned to have embedded points which in turn lead to progress in the game, and collaboration with other stakeholders within the game which along with communication lead to improved teamwork. Effectiveness testing revealed that the knowledge about patient safety improved equally in the game group and in the e-module group compared to the historical control group, although the game group reported higher levels of motivation (Dankbaar, Richters, et al., 2017). This SG also aimed at training students' resilience and found that students reported stress was lower in the game group compared to students in the e-module group (Dankbaar, Richters, et al., 2017). Albeit this being a significant effect and since scores were generally relative low, the authors ascribed this to students having possibly less stress in this study phase (Dankbaar, Richters, et al., 2017).

Another important skill that encompasses hard as well as soft skills is history taking. A study using a SG for training history taking in cardiological patients was developed based on the cognitive load theory and assessed changes in the first three levels of Kirkpatrick's model affective, cognitive, and behavioural attitudes (Alyami et al., 2019). The history taking was conducted via selection of the correct answer to the main characters complaints out of a multiple choice menu (Alyami et al., 2019). Alyami et al. (2019) reported having embedded the GDEs rules, points, score, rank, feedback, narrative, and three difficulty levels. They compared two groups with one playing the game and one receiving a pdf-document covering the same contents as the SG. Both groups significantly improved over time in knowledge and self-efficacy yet no group differences were found on the different levels apart from the game group reporting significantly higher satisfaction (Alyami et al., 2019).

These findings reflect the results of a systematic review of SGs in medical and healthcare professions education showing that the most frequently used GDEs were points, storyline, and feedback (Aster, Laupichler, et al., 2024). Additionally, the abovementioned SGs frequently relied on social aspects such as collaboration or challenge, highlighting the importance of interprofessional teamwork in medical education.

Given that simulations train practical skills in realistic, high-fidelity scenarios (Molloy et al., 2021), SGs represent more abstract settings by conveying learning objectives in a more game-like manner. The presented SGs for medical education demonstrated that using abstract settings and game play is also feasible for medical education. Thus, these SGs train skills or impart knowledge on a more abstract level. Serious games enabling the training of medical skills in a safe environment are especially important for training critical medical encounters. Nonetheless, SGs can also be used in non-life-threatening situations to train patient interaction in an error-tolerant environment. One such area in which key competencies, such as communication skills, can be practiced through SGs is history taking.

4 Importance of training history taking in medical education

History taking can be understood as a process of gathering information from the patient (e.g., related to the symptoms, the individual and their psychosocial status) relevant for arriving at a diagnosis and for initiating a suitable medical treatment (Keifenheim et al., 2015). History taking comprises two overarching aspects: while one aspect focuses on using the necessary communication techniques (e.g., listening, or verbal and non-verbal communication skills), the other emphasises the importance of a clear structure for gathering all information necessary for drawing a comprehensive picture allowing diagnostic reasoning (Nardone et al., 1980). Following the three-function model, history taking aims at collecting information from the patient about themselves, conveying empathic responses to the patients' emotional state, and educating the patient about their disease and influencing subsequent behaviours (Bird & Cohen-Cole, 1990). According to Hampton et al. (1975) and Peterson et al. (1992) about one-third of correct diagnoses can already be made after history taking and further physical

examinations or laboratory diagnostics did not add value in terms of changing diagnoses. However, the physical examination helped physicians become more confident with the diagnoses in some cases (Hampton et al., 1975).

Due to these findings medical education should put great emphasis on teaching and training history taking in medical students (Hampton et al., 1975). While communication skills can be learned, short-term interventions are not sufficient and the acquired skills can easily be forgotten (Aspegren, 1999). More recent research has highlighted the need for curricular adjustments, particularly regarding the extension of learning interventions into longitudinal formats (Bachmann et al., 2017). It was shown that medical students at graduation indeed showed an increased ability in asking questions regarding differential diagnosis and in time management but, still showed shortcomings in systematically capturing patients' histories and capturing symptoms comprehensively, as well as in communication techniques, empathy, and investigating the patients' perspectives (Bachmann et al., 2017). It has been shown that specialized trainings lead to an improved history taking (Rutter & Maguire, 1976), increased communication proficiencies, and self-confidence (Bachmann et al., 2013). There is a plethora of heterogeneous teaching interventions for medical history taking that focus on different aspects of history taking (Keifenheim et al., 2015). In this regard, interventions concentrate not only on teaching one skill like interview structure or covering interpersonal or communication skills, but address more than one skill (Keifenheim et al., 2015). Regarding the investigated style of interventions, they ranged from interview simulations in small-group workshops via role-play, to simulated patients or virtual patients and real patient interviews (Keifenheim et al., 2015). Although a vast number of training methods exists, a frequently used safe training method is the use of simulated patients (SP) (Kaplonyi et al., 2017). This is unsurprising, as communication training is expected to be practical and experiential, enabling students to actively practise verbal and non-verbal communication skills rather than passively participating in bedside teaching or lectures (Sezer et al., 2023). Simulated patients are not only a frequently used but also an effective method in terms of eliciting changes in knowledge and behaviour

regarding communication at the same time they are valued positively by the learners (Kaplonyi et al., 2017). Yet, integrating SP programs into medical curricular is fairly resource intensive. Not only, suitable persons have to be trained adequately to become a SP and present the respective disease in a standardised manner, but also, stakeholders are required for managing the SPs (Cleland et al., 2009). To circumvent the often limited amount of SPs, another safe and standardisable training method arose with the invention of virtual patients (VPs). These allow even more sophisticated scenarios and can be trained regardless of restraints in time and teaching space (Stevens et al., 2006). Researchers have already demonstrated that VPs can be effectively used for training communication skills in medical students (Sezer et al., 2023). Furthermore, students attending a learning course for history taking and clinical reasoning with an embedded VP showed significant improvements in their competencies afterwards (Isaza-Restrepo et al., 2018). Virtual patients have also been used for training empathic communication skills. Compared to the training with SPs, third-year medical students showed higher levels of empathy during the interactions with a VP with which they communicated via a free text entry in a chatbot-like interface (Kleinsmith et al., 2015). Kleinsmith et al. (2015) supposed this to be due to the VP interaction eliciting less pressure and stress on the students since they could take more time for answering compared to the interaction with a SP. When training with a VP, non-verbal communication skills can barely be practiced; however, VP-based training can occur prior to training with SPs to learn factually "what to say" before concentrating on "how to say" it empathically with SPs (Kleinsmith et al., 2015). A qualitative study with third-year medical students also mentioned that they felt cognitively narrowed on how to take a medical history when interacting with a SP hindering them to empathically react to patients' emotions (Brodahl et al., 2022). Some interviewed students attributed their difficulty in showing empathy to the cognitive demands of history taking, particularly the need to recall the procedure and all relevant information (Brodahl et al., 2022). This finding supports the approach of providing prior training in the technical aspects of history taking before focusing on the training of empathic communication.

Communication trainings are recommended to be longitudinal to provide sufficient settings for practicing the respective skills (Bachmann et al., 2017) but since SPs as training method require several resources (Cleland et al., 2009), it is inevitable that alternative approaches to SPs are needed. Pursuing the idea of VPs further, it is possible that chatbots can be equipped adequately to sufficiently represent VPs for training history taking. Before intertwining the aspects of training medical history taking and how this can be conducted with the help of chatbots, the next chapter firstly concerns different types of chatbots to provide an understanding of which chatbots can be used for training history taking.

5 May I ask you a question? – Different types of chatbots

As stated in chapter 2.2, the list of GDEs can be arbitrarily expanded. In this sense it is conceivable that a chatbot embedded in a SG can also serve as a GDE. Although Aparicio et al. (2012) asserted that a chat may foster the need for relatedness, it is essential to explicitly define the scope and the functionality of a chatbot before the addressed need can be examined.

According to the Cambridge Dictionary (2025), a chatbot can be defined as “a computer program designed to have a conversation with a human being, usually over the internet”. This aligns with other definitions from researchers, like the one from Wollny et al. (2021) who defined chatbots as “[…] digital systems that can be interacted with entirely through natural language via text or voice interfaces” (p. 2). Adamopoulou and Moussiades (2020) presented a general architecture for a chatbot’s function that fundamentally contains three building blocks. The user submits a request, which the chatbot processes in various ways to prepare a response that is finally outputted. Adamopoulou and Moussiades (2020) proposed that chatbots can be classified according to various aspects, of which the four most important aspects for the present thesis are mentioned in the following: The knowledge domain classifies how the chatbot retrieves answers from its trained data, based on a closed domain (i.e., knowledge about a predefined topic) or an open domain (i.e., not being limited to one predefined topic). For the service provided it is mainly important to differentiate between

interpersonal chatbots providing services, like answering frequently asked questions, and intrapersonal chatbots that influence the users' personal radius and might be related to other chat apps. The goals of a chatbot can be subdivided into informing the user, providing a conversation, or performing required tasks. The input processing and response generation can be built upon rule-based, retrieval-based, or generative models (Adamopoulou & Moussiades, 2020). The differentiation between rule-based and generative chatbots is also common for research on chatbots in education (Huang et al., 2025). Moreover, chatbots can be categorized based on their application area into "teaching-oriented" and "service-oriented" chatbots (Quiroga Pérez et al., 2020). Based on the findings of their systematic review, Wollny et al. (2021) developed a concept map for chatbots in education comprising the categories of the evaluation and educational effects of chatbots as well as the categories "applications" and "designs". In terms of input methods, chatbots can be subdivided into button-based interfaces, interfaces based on keyword recognition, context, or voice-enabled inputs (Smutny & Schreiberova, 2020). In the interface based on keywords, the user enters words or a phrase to which the chatbot proposes an adequate predefined response, while the context-based input types work on artificial intelligence (Smutny & Schreiberova, 2020). Accordingly, a plethora of different chatbots can be created by mixing different properties. Depending on the context in which a chatbot should be used as a GDE, other aspects are of importance.

To date, several chatbots are already used in different settings. In the educational context, chatbots are frequently used for learning a language or programming (Wollny et al., 2021), for computer science, or with a general objective (Kuhail et al., 2022). Hwang and Chang (2021) found in their review that chatbots were mostly used in communicative educational settings like language learning instead of arts or design courses and concluded that hands-on educational settings might not profit from using chatbots for training specific skills. In educational settings, chatbots were mostly implemented as web-based applications, acted as a teaching agent (i.e., acting as a tutor or teacher to the students), relied on a chatbot-driven flow-based interaction (i.e., the chatbot initiating the interaction), followed the principle of

personalised learning, and were tested effectively in experiments (Kuhail et al., 2022). While in the chatbot-driven interaction style the conversation is led by the system, the user-driven interaction places the user in control (Kuhail et al., 2022). Regarding the interface, chatbots mostly relied on text-based inputs while only some established voice-based or speech-to-text inputs (Huang et al., 2025). For educational purposes, chatbots are mostly used for skill improvement and follow a learning role implying that chatbots impart knowledge or proficiencies (Wollny et al., 2021). Another review found that most chatbots followed the pedagogical aim to provide social interaction support, although other chatbots also facilitated information and knowledge acquisition (Huang et al., 2025). Chatbots are also used aside from education in different areas of which medicine is one. Since the release of ChatGPT by OpenAI (2022) in November 2022, the popularity of Large Language Models (LLMs) significantly increased among the general public. Early findings suggest that ChatGPT provided more empathic and higher-quality responses to patients' questions than physicians originally being queried and answered in a social media forum (Ayers et al., 2023). Moreover, ChatGPT could potentially be used for supporting physicians with clinical documentation (Nayak et al., 2023). Another important aspect in medicine is the difficult access to mental health treatment services, which a personalised chatbot for self-referral eased (Habicht et al., 2024). However, Chatbots play not only a role in referral but also in general mental health. A scoping review found that chatbots for mental health were mostly used for therapy, training, and screening, were built on a rule-based basis and that conversations were mostly chatbot-driven (Abd-Alrazaq et al., 2019). That chatbots relying on generative artificial intelligence (AI) become only more and more common after the launch of ChatGPT becomes evident since a review on chatbots in the time between 1996 and 2023 still found a vast majority of rule-based chatbots (Huang et al., 2025). When searching for research on chatbots published from 2023 on, one mostly finds research on AI-based chatbots (e.g., Labadze et al., 2023; Stöhr et al., 2024; Wu & Yu, 2023).

Authors have claimed that chatbots should not be built only technology-driven but also theory-driven with a sound pedagogical basis (Wollny et al., 2021). As one pedagogical approach,

research has been conducted on how the use of chatbots affect students' motivation (Huang et al., 2025). Huang et al. (2025) analysed in their scoping review 43 studies that examined the impact of chatbots in terms of motivation and found a generally positive impact although some included studies showed mixed results. The authors categorised the underlying theories, which served either the development of the chatbot or the explanation of its motivational impact, into four frameworks: motivational, learning-related, communication-related, and user behaviour-related (Huang et al., 2025). Among the motivational theories is also the previously mentioned SDT which contains the key elements need for autonomy, need for competence, and need for social relatedness (Huang et al., 2025).

The overlap of chatbots being used for medical aspects as well as for educational purposes lends itself to use them as VPs within medical education. This is conceivable as chatbots are already integrated into communication trainings in medical education (e.g., Liaw et al., 2023).

6 Intertwining the chatbot and its function to train history taking – Aim of the thesis

Chatbots are already used as VPs standalone for medical students (e.g., Holderried, Stegemann-Philipps, Herrmann-Werner, et al., 2024; Holderried, Stegemann-Philipps, Herschbach, et al., 2024; Lippitsch et al., 2024) or for students from other healthcare disciplines (e.g., Benfatah et al., 2024; Rädel-Ablass et al., 2025) with most studies using ChatGPT. However, chatbots have not yet been widely implemented in SGs for medical students, although isolated examples exist (Ziebarth et al., 2014).

As described in Chapter 5, chatbots are often used to impart knowledge or information. It is assumable that chatbots for conducting medical histories can also be regarded as knowledge brokers. However, only obtaining information is not sufficient when taking a medical history. Especially when learning how to take medical histories, the two proficiencies of *how* to ask and *what* to ask (Kleinsmith et al., 2015) have to be obtained. Thus, it can be reasonably assumed that a chatbot for training history taking in medical education follows a twofold purpose: It has

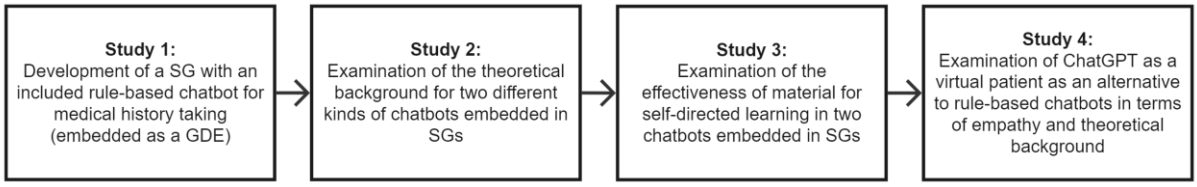
to provide all necessary information while simultaneously facilitating the training of history-taking skills. The question remains which kind of chatbot best elicits the combination of these two skills and which theoretical foundation suits best. It is reasonable to assume that a chatbot offering the opportunity of a conversation conducted with one's own words might foster students' autonomy and therefore enhance their motivation and consecutively learning performance (Huang et al., 2025). This might also apply for chatbots in which medical histories can be conducted. Therefore, this thesis project aimed to assess whether students' autonomy was addressed by different kinds of chatbots for history taking.

Based on a previous literature review, I have developed the GATE framework which can be used as an orientation to researchers when searching for theory-based and effectively tested GDEs for the embedment in SGs (Aster, Laupichler, et al., 2024). The GATE framework is explained in more detail in Chapter 6.1. A sound theoretical basis of GDEs allows for a convincing evaluation of them (Aster, Laupichler, et al., 2024). This is equally valid for chatbots as their educational groundwork should be aligned with the conclusive evaluation instead of only evaluating technical aspects (Wollny et al., 2021). Therefore, this thesis is informed by the GATE framework and primarily aims to assess a chatbots theoretical underpinning while also evaluating its effectiveness initially by examining students' history taking as process data throughout the chatbot use. This thesis research body is divided into four segments, these can be seen in Figure 3. It started with the development of a SG with an embedded rule-based open-entry chatbot. This resulted in Study 1: Development and evaluation of an emergency department serious game for undergraduate medical students (Aster, Hütt, et al., 2024). As a second segment of this thesis a study comparing the rule-based open-entry chatbot to an already previously examined keyword-based chatbot was conducted. This led to Study 2 titled "Theoretical background of the game design element "chatbot" in serious games for medical education" (Aster, Lotz, & Raupach, 2025) which assessed students' subjective and objective autonomy. The third segment examined the effectiveness of a guideline covering the topics of history taking as a material for self-directed learning in the two chatbots previously examined

in Study 2. This led to Study 3 titled “Impact of providing a customized guideline on virtual medical history taking in two serious games for medical education” (Aster, Lotz, Laupichler, et al., 2025). In the final study titled “ChatGPT as a Virtual Patient: Written Empathic Expressions During Medical History Taking” (Aster, Ragaller, et al., 2025), a generative AI chatbot (i.e., ChatGPT) was set up as a VP and students’ written empathy and their autonomy were assessed.

Figure 3

Sequence and content of the studies conducted for this thesis

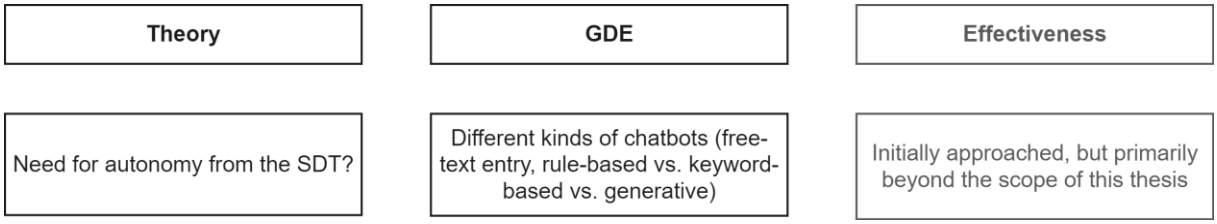


6.1 GATE framework – A framework for systematically choosing GDEs based on their theoretical foundations and effectiveness for the use in SGs

As a result of a systematic literature review regarding GDEs in SGs for medical and healthcare professions education, the GATE framework was proposed (Aster, Laupichler, et al., 2024). The framework contains GDEs used in SGs for medical and healthcare professions education along with their respective theory. In line with the previously described theories for GDEs and SGs, the GDEs within the GATE framework also referred to the needs postulated by the SDT. It was found that the need for competence was addressed by embedding educational material within the SG (in order to elicit a sense of connectedness to real-life) as well as by providing hints and tips (Aster, Laupichler, et al., 2024). Furthermore, the need for relatedness was addressed by collaboration between the players and the need for autonomy by including fewer rules within the SG (Aster, Laupichler, et al., 2024). Although theories were mentioned for the GDEs, no GDE was tested with regard to its effectiveness. This framework was established in the context of medical and healthcare professions education, it could likewise function as a blueprint for other, possibly broader, contexts. See Figure 4 for a depiction of the GATE framework adjusted for the GDE “chatbot”.

Figure 4

GATE framework adjusted for the present thesis focussing on the GDE "chatbot"



Note. Adapted from Aster, Laupichler, et al. (2024). The grayed out effectiveness testing extends the scope of this thesis but was included for the sake of completeness.

6.2 Integrating the game design element "chatbot" into the GATE framework

For this thesis, a chatbot is categorised as a GDE for training medical history taking. Accordingly, a chatbot is characterised as follows (references refer to the technological aspects and not to the connection with history taking):

- The conversation should be user-driven (Kuhail et al., 2022)
- Not a chatbot that provides support (e.g., chatbots for FAQs, Adamopoulou & Moussiades, 2020), but builds an independent learning environment in terms of being teaching-oriented (Quiroga Pérez et al., 2020) and takes on a learning role for skill improvement (Wollny et al., 2021)
- Can rely on both, closed or open knowledge domain (Adamopoulou & Moussiades, 2020), but may benefit from being open and not only answering to the predefined context but may have an extended and therefore possibly more realistic, patient-like knowledge
- Can rely on retrieval or generative output generation and can be best accessed via keyword-based, voice-enabled input (Adamopoulou & Moussiades, 2020; Smutny & Schreiberova, 2020; Wollny et al., 2021) or via free entries possibly enhancing users autonomy (Huang et al., 2025)
- To train communication in the sense of history taking two things should be trained simultaneously: the what (i.e., gathering all necessary information), and how (i.e., being empathic, listening actively, etc.) of taking a medical history (Kleinsmith et al., 2015).

In this sense, it provides the user with information while simultaneously providing a conversation (Adamopoulou & Moussiades, 2020)

- Should be distinguished from chatbots allowing communication between users (i.e., in the sense of GDEs in Table 2 this could be understood as ‘chat’)

7 Study 1 – Development of a serious game for medical education

Aster, A., Hütt, C., Morton, C., Flitton, M., Laupichler, M.C., & Raupach, T. (2024). Development and evaluation of an emergency department serious game for undergraduate medical students. *BMC Medical Education*, *24*(1), Article 1061. https://doi.org/10.1186/s12909-024-06056-z

7.1 Summary of Study 1

A SG was designed for medical students to train their clinical reasoning skills in a safe learning environment without compromising patients’ safety. Clinical reasoning is an essential competency to develop and consists of arriving at a diagnosis basically through generating and refining hypotheses and finally choosing and initiating appropriate treatments (Kassirer, 2010). Since these skills can be best learned with real cases (Kassirer, 2010), Serious games lend themselves to be used as a training environment as they hold the potential to provide a possibly unlimited number of real cases to students independent of time and place while being a safe environment allowing for errors. Serious games are already effectively used in the education of clinical reasoning for healthcare professions (Koelewijn et al., 2024; Middeke et al., 2018). In the present study, a SG representing a virtual emergency department, was developed according to the framework for SG development in medical education by Olszewski and Wolbrink (2017). All three steps of the framework (i.e., Preparation & Design, Development, and Formative Evaluation, Olszewski & Wolbrink, 2017) were conducted. The SG was designed and developed by a multidisciplinary team covering diverse professions with the aim of teaching undergraduate medical students clinical reasoning competencies but also

how to triage and treat patients under time pressure. The data basis of the SG relied on epidemiological data. Virtual patients were created by epidemiological probabilities and were equipped with AI-generated faces depicting the VP and their health condition. The main GDE that was used as operationalisation for the present thesis, was the chatbot. The chatbot provided the students with the opportunity to ask self-formulated questions regarding all relevant topics to be covered during history taking. Regarding history taking, clinical reasoning competencies should be trained by asking self-formulated questions. Compared with the chatbot characteristics defined in Chapter 6.2, the chatbot in the developed SG consisted of the following characteristics: user-driven (students have to enter the first question and the chatbot only answers to posed questions without asking further questions, what can be considered as illustrating real-life history taking), closed knowledge domain (only answers predefined questions), retrieval-based (i.e., rule-based) and is accessed via free entries, takes on a pedagogical learning role and provides users with information during conversation. Given the chatbot's characteristics, it can be assumed that both the content (e.g., embedded medical information) and the process (e.g., opportunity to ask freely formulated questions) of history taking can be trained. After the development phase, an evaluation was conducted with N = 146 third-year medical students in winter term 2021/2022 (application number by the local Institutional Review Board: 34/8/21). The study addressed usability and user experience using the System Usability Scale (SUS, Brooke, 1996) and the User Experience Questionnaire (UEQ, Laugwitz et al., 2008), respectively. Although another SG was included in the study, it was not central to the research focus. To maintain conciseness during the evaluation, only half of the participants were invited to complete the UEQ for the newly developed SG. The evaluable data set for the SUS consisted of n = 127 questionnaires, while the data set for the UEQ consisted of n = 76 questionnaires. In terms of usability, the results (M = 59.19) were below average indicating a marginal acceptability (Bangor, 2009). The UEQ is accompanied by an excel-based analysis tool, which is available online (Schrepp et al., n.d., ueq-online.org) and provides interpretation guidelines as well as benchmark values. The UEQ is divided in six subscales, when compared to benchmarks retrieved from commercial products only the scale

novelty reached good results, while the other scales received results below average (i.e., attractiveness, perspicuity, simulation) or even bad results (i.e., efficiency, dependability). The excel data analysis provides the interpretation that results between -0.8 and 0.8 can be regarded as a neutral evaluation, and results >0.8 as a positive evaluation. Following these classifications, the scales efficiency (*M* = 0.05, *SD* = 0.98, *95% CI* [-0.18, 0.27]), and dependability (*M* = 0.73, *SD* = 0.86, *95% CI* [0.54, 0.93]) can be viewed as being evaluated neutrally. On the contrary, the scales attractiveness (*M* = 0.94, *SD* = 1.29, *95% CI* [0.65, 1.23]), perspicuity (*M* = 1.12, *SD* = 1.02, *95% CI* [0.89, 1.34]), stimulation (*M* = 0.82, *SD* = 1.32, *95% CI* [0.52, 1.11]), novelty (*M* = 1.25, *SD* = 0.88, *95% CI* [1.06, 1.45]) can be regarded as being evaluated positively. Conclusively, the user experience was evaluated positively in most scales even though performing worse compared to the benchmarks. Additionally, students recommended within open feedback questions to further improve the chatbot.

7.2 Strengths and limitations of Study 1

One fundamental strength was the development of the SG by a multidisciplinary team consisting of experts from different professions and medical students. Hence, a comprehensive approach to the development was followed since including prospective users yields promising aspects in terms of including their perceptions of user experience (Crossley et al., 2016). Including their perspectives was not only guaranteed throughout the development process but especially by the conclusive evaluation as the gathered insights were incorporated into further iterations of the games' development. However, Study 1 came along with some limitations. As mentioned before, the UEQ was completed only by one-half of the cohort for the sake of conciseness during the evaluation. For this reason, comparing the results of the SUS and the UEQ may not lead to a sound comparison due to the different dataset sizes. Another point of criticism refers to the UEQ, as this measurement instrument was primarily designed for the use on commercial products and not games for learning. Therefore, just relying on the benchmark comparisons may lead to a false impression about the evaluation. However, since commercial games and SGs have common aspects as discussed in earlier

chapters, the UEQ can be considered as a valid tool for evaluating SGs. Nonetheless, the evaluation must not be concluded with Study 1 but should be maintained in an iterative manner over a prolonged time span.

7.3 Contribution of Study 1 to the framework

The first study laid the foundation for the following studies by developing a SG illustrating a virtual emergency department. Developing the SG first was especially important since the GDE "chatbot" should be examined within the context of a SG and not standalone. Therefore, Study 1 introduced the context by providing an overarching setting for subsequent examinations. In line with the findings in Chapter 3.2 that SGs frequently use emergency departments as settings for their contents, the SG was also conceptualised to represent an emergency department. This is crucially relevant since competencies needed for an emergency department profit from being trained in a safe environment. This study did not only develop the SG but also the inherent chatbot. For a chatbot to be used as a useful GDE for history taking it has to fulfil two aspects: It should provide sufficient content about the VP and it should rely on a theoretical basis evoking intrinsic motivation in the users. Therefore, the following study examined the theoretical basis of the chatbot.

8 Study 2 – Examining the theoretical foundation of a chatbot in a serious game – Does it elicit autonomy?

Aster, A., Lotz, A., & Raupach, T. (2025). Theoretical background of the game design element "chatbot" in serious games for medical education. *Advances in Simulation, 10*(1), Article 10. https://doi.org/10.1186/s41077-025-00341-7

8.1 Summary of Study 2

Following the line of thoughts presented in this thesis, Study 2 aimed at assessing the chatbots theoretical foundation. As previously shown, the SDT is a frequently used theory for SGs as well as for GDEs. Therefore, Study 2 focused on evaluating whether the need for autonomy

stemming from the SDT can be regarded as a theoretical foundation of chatbots for history taking. This assumption is reasonable since a chatbot in which free entries are possible may lead to a more pronounced exploratory behaviour which in turn is associated with an satisfied need for autonomy (Schutte & Malouff, 2019). For examining this assumption, a study was conducted comparing two different chat systems. One being the chatbot set up in Study 1 and the other being a more constrained chat system embedded in a SG already used in previous studies (Middeke et al., 2020; Middeke et al., 2018). In the constrained chat system students entered only keywords, in parts or completely, of their questions and the chat system proposed all questions containing the keyword. Thus, students could select from a list of predefined questions and did not have to word their own questions. In terms of the previously defined chatbot characteristics, this chat system can be considered keyword-based. Therefore, the research followed three hypotheses. First, significantly more questions are asked within a free-entry chatbot. Second, significantly more irrelevant questions are asked within a free-entry chatbot. Third, significantly more subjective feelings of autonomy are reported after using a free-entry chatbot for history taking.

For the analysis of these hypotheses, students' autonomy was operationalised two-fold. On the one hand, students answered a conclusive questionnaire regarding their subjective feelings of autonomy, and on the other hand a more sophisticated approach was chosen: To assess students' autonomy not only after taking the history in a chatbot but also meanwhile, students' entries or chosen questions in the chat systems were evaluated. Alternatively, students could have answered questions regarding their autonomy during the game play. However, in order to avoid impairing students' play experience and hindering a natural gaming flow it is recommended to assess process data in terms of game analytics (Qian & Clark, 2016). Another factor that led to this approach is that this could best illustrate students' exploratory behaviour during history taking, which is especially important when assessing how students conduct a medical history. The study consisted of two groups of fourth-year undergraduate medical students from Göttingen medical school playing the two SGs with the embedded chat

systems during the winter term 2023/2024 (application number by the local Institutional Review Board: 8/9/23). Both SGs covered VPs suffering from cardiac diseases. A checklist was developed in collaboration with a cardiologist to rate students' history entries. All history takings were evaluated blindly, resulting in a history score for each student. Applying the checklist meant that only students' questions were rated irrespective of the chatbots' answers. Moreover, each question was rated only once, even if it was later reformulated during the course of the history-taking process. The checklist was applied for all diseases with questions receiving one or two points depending on their medical importance. The checklist can be retrieved as supplementary material from the journal's website. Students' subjective feelings of autonomy were assessed with the subscale for autonomy "perceived choice" of the *Intrinsic Motivation Inventory* (IMI, McAuley et al., 1989). Moreover, self-efficacy was assessed as an exploratory variable using the *General Self-Efficacy Short Scale* (German: Allgemeine Selbstwirksamkeit Kurzskala, ASKU, Beierlein et al., 2013). For the analysis of the hypotheses, Mann-Whitney U-tests and a polynomial regression analysis were conducted. In line with the first hypothesis, the results showed that the free-entry chatbot led to significantly more questions being asked than in the constrained chat system. Sequential analyses were conducted to examine the second hypotheses and conclusively showed that more irrelevant questions were asked in a free-entry chatbot than in a constrained chat system. In line with the theoretical derivations, this led to the assumption that students showed more objective exploratory behaviour in a free-entry chatbot what can be regarded as a measure for autonomy. Each history taking could be scored with a maximum of 49 points, however the average history taking scores were *Mdn* = 14.5 (29.6%) for the free-entry chatbot and *Mdn* = 14 (28.6%) for the constrained chat system. Detailed information about the statistical procedures can be found in Aster, Lotz and Raupach (2025). Students reported on average a moderate subjectively experienced autonomy, while no significant difference was found between the different chat systems. Moreover, no significant correlation between self-efficacy and autonomy appeared nor were differences in self-efficacy between the chat systems found.

In conclusion, Study 2 showed that autonomy is objectively addressed by a free-entry chatbot while the subjectively experienced autonomy did not differ between the chat systems.

8.2 Strengths and limitations of Study 2

Study 2 offered an innovative approach for assessing students' autonomy during virtual history taking in a chatbot by operationalising their exploratory behaviour through process data in terms of their entries. One further advantage of the study is that two different chat systems were compared in terms of how autonomy was addressed. As mentioned above, it is also essential for testing a chatbot as a GDE to be embedded in a SG. This prerequisite was given in the study and two different chat systems could be compared as GDEs with the SGs not essentially differing in their basic building blocks except from their visual appearance and chat systems. However, future studies should examine different chat systems within the same SG, but since this requires more resources, Study 2 provides first insights in the theoretical foundations of different chat systems as GDEs.

Although the chosen approach provided advantages, the use of process data was simultaneously a limitation in terms of their quantification. A checklist for the scoring of the individual history taking data was developed and used consistently across all diseases. Since not all diseases presented with all symptoms (especially pain), this led to the problem of not being able to adequately assess the students' history taking in all cases. While using one checklist for all diseases enabled a facilitated scoring procedure, it possibly biased the history scores. A general question that arose from the aforementioned aspects of history taking is whether and how these can be trained and depicted during a virtual history taking within a chatbot keeping in mind the different aspects it comprises. This also gives rise to the question how checklists for assessing virtual history taking can best capture all shown aspects. The checklist developed for Study 2 was based on the SAMPLER/OPQRST scheme (Kegel et al., 2022) frequently applied in emergency medicine, but since already validated assessment scales for history taking exist (other scales can exemplarily be found in Bachmann et al., 2017; Keifenheim et al., 2015), future studies should consider refining the checklist in order to allow

for a specified scoring for different diseases while allowing an efficient scoring procedure or testing other validated assessment scales. Moreover, using the IMI as a questionnaire for assessing needs in SGs is fraught with problems since it is not a validated measurement instrument in this field. The subjective data could not be correlated with the objective data as a result of the different data quantity and hence the data could not be linked with each other. Since validated questionnaires for the explicit measurement of addressed needs by GDEs in SGs do not exist yet, future research should concentrate on the development and validation of such a measurement instrument.

The assumption that a chatbot with more degrees of freedom (i.e., allowing free entries) should address autonomy in terms of exploratory behaviour more than a chatbot with fewer degrees of freedom (i.e., choosing from predefined questions) was derived from the literature and can be viewed as a deductive approach. However, the need for autonomy is only one of three psychological needs proposed by the SDT. It is possible that a chatbot might also address other needs which is why further studies should follow a more inductive approach by assessing the role of all needs and hence deriving a theoretical foundation. Since the chatbot is not meant to address needs in terms of communication with other users, it is unlikely that a chatbot for history taking might address the need for social relatedness while the need for competence might be more likely addressed. The cognitive evaluation theory (Deci & Ryan, 2012) substantiates this assumption as it states that autonomy and competence are intertwined and that intrinsic motivation can be best promoted when autonomy and competence are addressed simultaneously (Deci & Ryan, 2012). An additional analysis conducted exclusively for this thesis revealed that subjective autonomy and perceived competence had a weak nearly moderate but statistically significant correlation on the significance level of .05 (r = .295; p = .011; N = 74). In this sense, it is conceivable that a retrieval-based chatbot rather addresses the need for competence than the need for autonomy since questions compatible to the chatbots' rules have to be queried while a generative chatbot may address autonomy due to

its open knowledge domain. Further studies should examine chatbots based on generative models.

8.3 Contribution of Study 2 to the framework

Study 2 made an initial step to fill the theoretical aspect of the GDE chatbot within the GATE framework. Moreover, analysing the process data of students' history taking can be regarded as an approximation to effectiveness testing since the process data allow for an evaluation of the quality of history taking and how effective the chatbot could be used. It is already proposed by the literature to use internal game parameters to evaluate the users' performance but also to validate the SG by means of the collected data (Graafland et al., 2012). Nonetheless, this was not the focus of the study and further research should investigate the chatbots effectiveness.

As defined, feelings of autonomy arise when one feels that their own behaviour can be executed in a self-determined and volitional manner (Niemiec & Ryan, 2009). Compared with the findings of Study 1 that the subscale "dependability" (i.e., feeling of being able to control an interaction) of the UEQ yielded poor results, it can be assumed that the user experience may be one factor that hindered students from subjectively feeling autonomous although the objective results offered hints about students autonomous behaviour.

The relatively low history scores achieved by students raised the question of whether chatbots should be equipped with additional instructional content on history taking. Building on this, the following study examined the impact of presenting additional instructional content regarding history taking between two sessions on students' history taking, using the same chatbots as in Study 2.

9 Study 3 - Testing the effectiveness of self-directed learning in a chatbot for history taking – Do history taking skills profit from studying a customized guideline for history taking?

Aster, A., Lotz, A., Laupichler, M.C., & Raupach, T. (2025). Impact of providing a customized guideline on virtual medical history taking in two serious games for medical education. *Medical Education Online, 30*(1), Article 2527175.

https://doi.org/10.1080/10872981.2025.2527175

9.1 Summary of Study 3

Although the focus of Study 2 was to examine students' exploratory behaviour and therefore their autonomy during history taking, the findings also revealed that students achieved less than one-third of all possible points during history taking. This finding has to be viewed in line with students attending a module explicitly covering history taking before the module in which the study was conducted. Thus, the assumption remained that solely using the chatbot is not sufficient for training history taking. This raised the question of whether providing material for additional self-study would improve the history taking scores between two sessions. It was assumed that material for self-directed learning would help students recall their existing knowledge. Self-directed learning is understood as process in which the learning process can be executed self-planned (Knowles, 1975; Loyens et al., 2008). In this sense it was conceived that students should build an internal scheme for their history taking by studying the provided guideline. This assumption was derived from the literature stating that cognitive representations are viewed as parts of clinical reasoning (Young et al., 2018). Due to clinical reasoning being closely linked to history taking (Furstenberg et al., 2020), it was assumed that these representations in form of an internal scheme can be developed for history taking. Building on the findings of Study 2, which showed that a chatbot allowing free-text entries encouraged more exploratory behaviour during history taking, these insights were combined with the assumption of developing an internal scheme. Based on this integration, it was

hypothesised that self-directed learning would improve history taking in a free-entry chatbot, but not in a constrained chatbot. For a baseline evaluation, it was hypothesised that a chatbot being keyword-based in terms of providing a long menu format would outperform a chatbot relying on free entries. This assumption is based on the constructs of cued and free recall. In this sense, a free-entry chatbot allows free recall, while a keyword-based chatbot enables cued recall. It is known from earlier research that cued recall (i.e., providing cues as a retrieval aid) outperforms free recall (i.e., recall without cues) (Higham & Guzel, 2012). In particular, two hypotheses were examined. First, it was hypothesised that the keyword-based chatbot leads to significantly higher history scores in session 1 compared to the first and the second session of the free-entry chatbot. Second, it was hypothesised that the guideline would lead to higher history scores in the free-entry chatbot but not in the keyword-based chatbot when comparing session 1 and session 2.

For examining these assumptions, the study design of Study 2 was adapted accordingly. Before the study was conducted in summer term 2024, the local Institutional Review Board in Göttingen approved the study (application number: 21/3/24). The two SGs previously described were used and N = 159 fourth-year students from Göttingen medical school were randomized to one of the games. This resulted in n = 79 students using the keyword-based chatbot and n = 80 students using the free-entry chatbot after providing their informed consent. Both groups used the same chatbot over two consecutive weeks, each session presenting several cardiac diseases. As a workaround for presenting the additional instructional content, students received a guideline covering relevant aspects of history taking previously developed by the two study authors AA and AL. The guideline was presented between the two sessions and students were encouraged to use the guideline only between both sessions. The guideline served as material for self-directed learning. Students' process data regarding the history taking of both sessions was analysed similarly to Study 2. Since the checklist used in Study 2 suffered from some limitations, the checklist was refined for the purpose of this study. Please refer to (Aster, Lotz, Laupichler, et al., 2025) for a detailed description of the guideline

development and the checklist refinement. Both, the guideline and the checklist can be found at the journals homepage. Additionally, students self-assessed their learning outcomes after the second session by answering questions regarding their history taking skills in a retrospective and in a post manner. The comparative self-assessment (CSA) gain method was used for analysis (Raupach et al., 2011). Although the assumption of normal distribution was violated, paired and unpaired *t*-tests as well as a mixed ANOVA could be conducted due to the large sample size (Glass et al., 1972; Rasch & Guiard, 2004).

As was the case in Study 2, this study also found that students averagely achieved about only one-third of all possible points for their history taking. Unpaired *t*-tests revealed no significant differences neither between the first keyword-based chatbot session with the first nor with the second free-entry chatbot session. However, in both comparisons the average score achieved in both free-entry chatbot sessions slightly outperformed the first keyword-based chatbot session. For the analysis of the second hypothesis, paired *t*-tests were conducted. No significant difference emerged for the comparison between the first and the second free-entry chatbot session, although on average slightly higher scores were achieved in the second than in the first session. Contrary to the postulated hypothesis, a significant difference occurred between the first and the second keyword-based chatbot session with a higher score being achieved in the second session. A proxy variable was derived to perform an ANOVA (please refer for the concrete derivation to Aster, Lotz, Laupichler, et al., 2025), but neither an interaction effect for session and SG nor a main effect for the session were found. Conclusively, the analysis of the CSA gain showed that the SGs did not differ significantly across individual items although using the keyword-based chatbot led to slightly better results albeit being slight improvements.

The findings of Study 3 indicated that history taking conducted in a keyword-based chatbot profited more from interposed material for self-directed learning than history taking conducted in a free-entry chatbot. A possible explanation refers back to previous findings related to cued and free recall showing that cued recall evoked better results (Higham & Guzel, 2012; Tulving

& Pearlstone, 1966). Thus, it has to be discussed further if chatbots should provide the user with additional cues in order to focus on training the *what* of history taking (Kleinsmith et al., 2015).

9.2 Additional analyses

In addition, the questionnaires IMI, SUS, and UEQ (please refer to the description of Study 1 and 2 for a detailed explanation) were administered during the study but were not reported in the original publication. Thus, they were analysed for the purpose of this thesis. The ethical approval obtained for Study 3 also encompassed the data used in the additional analyses.

Overall, n = 69 questionnaires were completed by students who played the free-entry chatbot and n = 53 questionnaires were completed by students who used the keyword-based chatbot. The assumption of normal distribution was also violated in these data but due to the sample size t-tests can be assumed to be robust and can nonetheless be conducted (Rasch & Guiard, 2004).

Due to the omission of two data series, the final dataset for the IMI consisted of n = 51 data for the keyword-based chatbot and n = 67 for the free-entry chatbot. Levene's test revealed homogeneity of variances (p > .05). An unpaired t-test revealed a significant difference between both SGs in terms of the perceived autonomy $t(116) = -6.110$, $p < .001$, $d = -1.135$ with a higher perceived autonomy in the keyword-based chatbot (M = 3.96, SD = 0.94) than in the free-entry chatbot (M = 2.93, SD = 0.88). Although a significant difference appeared, both average ratings represent only a medium autonomy score compared to the scale ranging from 1 to 7.

Additionally, students provided ratings regarding the user experience and the usability. Regarding the user experience for the SG including the free-entry chatbot, two complete data sets were missing. Therefore, n = 66 data sets were used for the analyses while one data point was present for novelty this subscale consisted of n = 67 data sets. Contrasting Study 1 with Study 3, the SG including the free-entry chatbot received a more positive evaluation on all

scales (efficiency (*M* = 1.20, *SD* = 1.04, *95% CI* [0.95, 1.45]); dependability (*M* = 1.17, *SD* = 0.92, *95% CI* [0.95, 1.39]); attractiveness (*M* = 1.75, *SD* = 1.07, *95% CI* [1.49, 2.01]); perspicuity (*M* = 1.56, *SD* = 1.14, *95% CI* [1.29, 1.83]); stimulation (*M* = 1.95, *SD* = 1.03, *95% CI* [1.70, 2.19]); novelty (*M* = 1.97, *SD* = 0.99, *95% CI* [1.73, 2.21])). An unpaired *t*-test was conducted for the comparison of the usability ratings of Study 1 with those of Study 3. All data of Study 1 (*n* = 127) and Study 3 (*n* = 69) found entrance in the analysis. Levene's test did not prove homogeneity of variances ($p < .001$), which is why the result of the Welch test is reported. A significant improvement in the usability could be observed between the studies (*M* = 59.19, *SD* = 9.82; *M* = 74.86, *SD* = 16.08), $t(96.227) = -7.379$, $p < .001$, $d = -1.104$. Conclusively, it can be stated that the SG including the free-entry chatbot improved in user experience and usability due to continuous development of the interface.

9.3 Strengths and limitations of Study 3

Study 3 compared the impact of material for self-directed learning on history taking in two distinct chatbots embedded in SGs. Moreover, the study drew on a substantial amount of data not only in terms of the student cohort but also in terms of the assessed data points.

Nonetheless, the study came along with several limitations. Firstly, due to the diverse assessment settings divided into an online and an on-site setting, it was difficult to assess whether students have used the guideline during the second session. A control item in the consecutive questionnaire assessed students' use of the guideline. It can be assumed that students did not use the guideline in the second session, as their scores would likely have been higher otherwise. The second limitation concerns the applied scoring checklist. Although the checklist was refined for Study 3, students again did not reach more percent than in Study 2. It was aimed to guarantee a good cost-benefit ratio for the raters. However, it has to be discussed if the checklist might have led to the low history scores since it is plausible that not all necessary questions might have been covered. Thirdly, the free-entry chatbot sometimes provided mismatched answers to the posed questions and students had to reformulate their questions. Regarding the additional analyses, the same limiting factors as in Study 2 can be

applied for the IMI. Although the additional analyses allowed for a comparison of usability and user experience measures between Study 1 and Study 3, it should be noted that the analyses were based on different sample sizes.

9.4 Contribution of Study 3 to the framework

Study 3 built upon the finding of Study 2 that students achieved on average less than one-third of all possible points for their history taking. From this finding the idea was derived whether a chatbot alone is not sufficient enough for training history taking and needs to give additional assistance to students. Although students attended a course covering history taking several semesters earlier, they showed an averagely low amount of achieved points. When acquiring new skills within a chatbot, the chatbot should assist and guide learners during their learning procedure in terms of scaffolding (Wollny et al., 2021). Scaffolding was originally defined as additional assistance provided by experts or more capable peers (Wood et al., 1976). However, it has already been shown that chatbots are also capable to scaffold students during their learning process (Sikström et al., 2024). In line with this findings, the third study added material for self-directed learning in the sense of scaffolding learners for developing their internal scheme for history taking and compared it between two distinct chatbots. Contrary to the assumption, students who played the SG with the keyword-based chatbot profited more from the guideline. This has to be viewed in line with the findings that cued recall outperforms free recall. Therefore, it has to considered whether a keyword-based chatbot is better suitable for teaching the “what” of history taking. Building upon this assumption, the fourth study was developed and took the thought further to assess whether a LLM is better suitable for training the “how” of history taking. In line with the previous findings, it is assumable that a chatbot for history taking needs more than just providing the contents and the opportunity for conducting a conversation with a VP to be regarded as a GDE. At this stage, it can be concluded that a free-entry chatbot allows more exploratory behaviour and leads to slightly higher performance scores. However, a keyword-based chatbot profits more from material for self-directed learning, what was not only reflected in the improvements of the process data but also in

students' self-assessment of their learning outcome. Therefore, more factors have to be researched that contribute to a chatbot being a GDE.

Moreover, the additional analyses regarding students' subjective autonomy revealed a significant difference between both chatbots with the keyword-based chatbot eliciting more autonomy. Compared to the second study, where no statistically significant difference appeared between the chatbots, it can be concluded that the material for self-directed learning did not only lead to better objective results but also to higher subjective autonomy. Future studies should further evaluate whether a scaffolding chatbot is better suitable for training history taking, in terms of objectively achieved scores as well as in terms of students' satisfied need for autonomy.

History taking cannot only be considered as a GDE but must also be considered as an essential part of clinical encounters. Therefore, and in line with the previously mentioned aspects and functions of history taking, it is crucial to not only focus on what was asked but also on how the history taking was conducted. Training with SPs allows to train all aspects of history taking. Hence, the question arose whether training with a chatbot is suitable to train empathy. Study 2 also showed that retrieval-based chatbots addressed objective autonomy but not subjective autonomy and it is therefore of interest whether generative chatbots may also lead to higher subjective autonomy.

10 Study 4 – Is ChatGPT an alternative? Testing ChatGPT as a virtual patient for empathic history taking

Aster, A., Ragaller, S.V., Raupach, T., & Marx, A. (2025). ChatGPT as a virtual patient: Written empathic expressions during medical history taking. *Medical Science Educator, 35*(3), 1513-1522. https://doi.org/10.1007/s40670-025-02342-7

10.1 Summary of Study 4

Study 4 aimed at assessing students' written empathy during history takings with ChatGPT as a VP and their subjectively experienced autonomy afterwards. Empathy is known to be an essential part of communication and can be conveyed by verbal, non-verbal or paraverbal aspects. The study assessed whether empathic history training can be conducted with ChatGPT as a VP, therefore the focus was on verbal written aspects. Third-year undergraduate medical students (N = 35) participated in the study and took medical histories within their own ChatGPT 3.5 accounts. Students entered a previously developed (by the first author) prompt into their ChatGPT interface and took the medical history until they felt they had collected all necessary information. The prompt consisted of three parts, a first part introduced ChatGPT how to act as a SP, the second part contained the medical information about endocarditis or heart failure (only one disease was chosen per prompt), and the third part contained information about the following history taking procedure. Detailed information about the crafting of the prompt can be found in Aster, Ragaller, et al. (2025). Moreover, the original publication contains information about response temperature adjustments in the prompt but since these did not have a high impact on the results, they are not discussed in this thesis. Detailed information can be found in the original publication (Aster, Ragaller, et al., 2025). After they took the history, students made their chats available in an anonymous form by sharing the export link. Following this, they completed a questionnaire about their subjectively experienced autonomy during the history taking. Due to the limitations of the questionnaire of Study 2 and 3, a different questionnaire was used in this study. Six questions

regarding students' autonomy were extracted from a questionnaire developed by Sailer (2016) and were rephrased to fit the context of history taking. This questionnaire was chosen since it was developed in the broader context of gamification and it can be assumed that it is more appropriate for assessing needs in SGs than the IMI used in Study 2 and 3. For analysing students' written empathy during the history takings, the Empathic Communication Coding System (ECCS, Bylund & Makoul, 2002, 2005) was applied by two independent raters. The ECCS provided a rating scale for students' responses but also for ChatGPT's statements, which is why two separate variables were examined. The local ethics committee approved the study beforehand (application number: 2024–96-BO). Overall, 28 chat protocols were analysed and the results showed that students were able to exchange 659 general interactions with ChatGPT, although only 14% could be scored as being empathic. Even though this appears to be a small number, the high number of interactions has to be highlighted meaning that ChatGPT was able to maintain its role as SP throughout a prolonged time. Regarding ChatGPT's statements it was found that it mostly provided emotion statements expressing feelings or challenge statements explaining negative consequences of problems on their quality of life. Students mostly responded with either implicitly recognising ChatGPT's perspective or completely denying empathic opportunities given by ChatGPT. In terms of subjectively experienced autonomy, students reported high levels of autonomy. Even in comparison with the developmental study of the used questionnaire of which the statistical values can be used as references, students reported higher levels of autonomy in the present study. Conclusively, the study yielded preliminary results that ChatGPT might be useful for training empathic history taking. Moreover, in comparison with Study 2, this study showed that a chatbot being generative-based leads to higher experiences of autonomy than a retrieval-based chatbot. This might be elicited by the different knowledge domains the chatbots rely on. While the chatbots compared within Study 2 had closed knowledge domains to the medical diseases, ChatGPT relies on an open knowledge base as it can retrieve information to more than the medical disease. A possible explanation is that this has led to a more satisfying human-like conversation which in turn made students feel more autonomous.

10.2 Strengths and limitations of Study 4

When it comes to training history taking, it is not only important to shed light on the students' perspective in terms of learning outcomes but also to examine the learning environment. It is crucial that the learning environment provides empathic opportunities for students to react empathically. Study 4 covered both aspects by employing the ECCS as the tool for analyses. Large language models such as ChatGPT or other more medicine-related applications are increasingly used. Therefore, it is important to assess how students use these tools for their learning. It can be regarded both as a strength and as a limitation simultaneously that students used their own ChatGPT accounts. While its use enhances external validity by allowing use both in medical school and at home, the learning process is associated with less regimentation by teachers. Therefore, it is conceivable that generative chatbots can be used for training history taking but further studies should incorporate them in a way that allows for more regimentation.

A major limitation that restrains the study from providing more than preliminary results is the limited sample size. Since the present results show the potential of ChatGPT to be implemented in empathy training, it is likely that future studies with larger samples may support the results. The ECCS is already validated for evaluating empathy in human conversations, but not yet for communication between humans and LLMs and thus might not be sufficient in a chatbot. Future studies should use other scales extending the focus of empathy as exemplarily done in Bachmann et al. (2017) or should validate the use of the ECCS in these contexts. However, previous research has already used the ECCS to assess the empathy of third-year medical students obtaining the history of VPs and found that students showed empathy on a high level (Kleinsmith et al., 2015). Another limitation of chatbots used for empathy training in a virtual environment is the absence of non-verbal and paraverbal communication aspects. Future research should examine to which extent this is relevant by comparing the performance in trainings with a chatbot with SP trainings.

10.3 Contribution of Study 4 to the framework

Alongside Study 2, this study examined the theoretical foundation of a chatbot. However, this study added value to the findings so far regarding retrieval-based chatbots by specifically focusing on a generative chatbot. This is crucial since the use of generative chatbots is increasing steadily as already shown in chapter 5. Additionally, this study shed light on the question whether ChatGPT can be used for training empathic history taking. Empathy is an important cornerstone in medical history taking and should therefore be extensively trained in medical education (Vogel et al., 2018). Hence it is obvious that when a chatbot is used for history taking, these competencies should also be trainable within it. Previous research implied that a learning environment supporting students' autonomy led to more empathy but only in students with generally better satisfied psychological needs (Neufeld & Malin, 2024). However, since empathy comprises different aspects and a virtual learning environment is still artificial for training empathy in history taking, future research is needed to compare it with real SP contacts and prove its effectiveness. In line with Study 2, it should also be examined to which extent a generative chatbot also addresses the need for competence. This is possible since the human-like conversation may lead to the feeling that students' questions have an effect on the progress of the history taking. The next logical step would be to integrate a generative LLM into a SG to evaluate it as a GDE and Study 4 yielded initial results that this might advance the respective SG. Referring to the SG presented in Study 1 and students' reactions to the chatbot, it is likely that this SG might benefit from being equipped with a generative chatbot.

11 Merging and discussing the findings

The following chapter synthesises the gained findings from Study 1 to Study 4 to answer the question of whether the need for autonomy can be regarded as the theoretical foundation of a chatbot used for history taking. The chapter concludes by elaborating on the characteristics that qualify a chatbot as a GDE and discusses which type of chatbot is most suitable for training history taking in medical education.

11.1 Theory underlying a chatbot for history taking

The thesis project aimed at examining the underlying theoretical foundation of chatbots for training medical history taking while assessing the characteristics that qualify a chatbot as a GDE. It was hypothesised that the need for autonomy stemming from the SDT can be regarded as the theoretical foundation.

Merging the findings of Study 2 and Study 4 indicated that the more degrees of freedom a chatbot had, the more subjective autonomy was elicited with high subjective autonomy being experienced in a generative chatbot. However, the results on subjective autonomy are yet difficult to compare between these studies due to the different questionnaires used. The additional analyses conducted for Study 3 allow for a comparison with Study 2 as well. Both studies used the same questionnaire and it became apparent that, descriptively, students reported moderate subjective experiences of autonomy in both studies. Nevertheless, using the keyword-based chatbot in Study 3 led to a significantly higher perceived autonomy than using the free-entry chatbot. Due to the interposed additional instructional content regarding history taking, it can be assumed that the keyword-based chatbot eased students' application of their knowledge and therefore enhanced their subjective autonomy. Although the subjective results in Study 2 suggested otherwise, the objective data revealed that students showed more exploratory behaviour, and thus autonomy, during history taking when interacting with a free-entry chatbot compared to a keyword-based chatbot.

The following order of autonomy-supportive chatbots is likely (from eliciting less to eliciting more autonomy): 1. Keyword-based, retrieval-based, closed domain knowledge chatbot; 2. Free-entry based, retrieval-based, closed domain knowledge chatbot; 3. Free-entry based, generative-based closed domain knowledge chatbot (albeit not being explicitly studied within the thesis); 4. Free-entry based, generative-based, open domain knowledge chatbot. These findings underline the hypothesis made by Huang et al. (2025) that generative chatbots might lead to higher experienced levels of autonomy due to the possibility to conduct free conversations. Although the assumption that the need for autonomy may be addressed was

derived deductively from existing theories (see Chapter 2), future studies should also focus on the need for competence as these needs are closely intertwined for evoking intrinsic motivation (Deci & Ryan, 2012). Besides this approach, it is also possible to approximate the theoretical foundation inductively and derive the theoretical groundwork from empirical findings. As became apparent during the design and conduction of the studies, sound valid and reliable measurement instruments for the assessment of addressed needs by GDEs in SGs are missing. Therefore, a preceding step should be to develop such an instrument before evaluating the theoretical foundation. However, earlier studies also relied on results deducted from the applied questionnaire (Sailer et al., 2017).

Although the conducted studies provided insights into students' autonomous feelings and behaviour, the chatbots effectiveness was not directly measured and examined. It can be stated that the theory section of the GATE framework has been initially addressed, but further research is needed to populate its effectiveness section in relation to chatbots. As previously explained, to evaluate any element as a GDE within a SG, it consequently has to be embedded in a SG. While this was the case for Study 1, 2 and 3, Study 4 concentrated on ChatGPT standalone what can be regarded as a disadvantage of the study. ChatGPT is a widely available tool that is not limited to the use at medical schools but is also available in students' private life. Hence, this approach enhanced the external validity and it can be hypothesised that students would produce similar results when using ChatGPT at home for training history taking. Nonetheless, at least initially, students should be guided in its use by medical educators, for instance by being provided with a previously tested prompt. Moreover, at least initially, it should be followed by a debriefing or discussion session in medical school. To explicitly analyse ChatGPT or other generative chatbots as GDEs, future studies should repeat the study with them being embedded in SGs.

Conclusively, the studies presented in this thesis yield insights into the theoretical basis of chatbots used for medical history taking, but following the GATE framework the entire research of chatbots is not yet finalised since insights into their effectiveness are still lacking. Assessing

the theory is one component of the GATE framework, but it represents only one step, especially when considering which aspects are relevant for a tool designed to train history taking. Addressing the psychological needs leads to intrinsic motivation but future studies should investigate whether intrinsic motivation is enough for a chatbot to substitute other training methods involving humans (e.g., interacting with patients, SPs, or peers). Therefore, future research on this is needed, but it is more likely that chatbots may be used as a supplement to existing teaching methods. However, it is difficult to train interpersonal interactions or recognising content “between the lines” within a chatbot. But since chatbots provide a safe learning environment, future studies should continue to evaluate to which extent chatbots can be and should be used for training history taking in medical education. Study 4 shows that chatbots can be used for training empathy but should only be a supplement for the time being. Combined with the findings from Study 2, this suggests that chatbots can be better used for training the *what* of history taking (i.e., factual knowledge and procedure) while it is to date hardly possible to train non-verbal or paraverbal communication aspects within a chatbot. This is also conceivable combined with the results of Study 3, as the results suggest that a keyword-based chatbot would benefit from providing additional cues for history taking thus training the *what* of history taking. Referring back to the proposed functions of history taking and combining them with the discovered findings, it is likely that it depends on the characteristics of the chatbot whether all functions can be trained.

To advance the GATE framework and evaluate its effectiveness, the question arises: how can such effectiveness testing be carried out? In principle, effectiveness can be assessed either during the use through process data or retrospectively after the chatbot has been used. This later can be viewed more nuanced. One important aspect of effectiveness is whether the transfer from a chatbot to real-life is as good as the transfer from learning with a SP to real-life. For these assessments, it should be referred back to the evaluation levels proposed by Kirkpatrick (Kirkpatrick & Kirkpatrick, 2016). Acquired communication skills can be either directly measured in physician-patient encounters or in SP encounters via validated

assessment scales (Gärtner et al., 2022) or in OSCEs, although these have to be considered carefully (Setyonugroho et al., 2015). It is also possible to equip the chatbot with these measurement methods in order to provide students with feedback during or after their history taking and to receive an evaluation directly afterwards without executing a specific tests. Previous research has demonstrated the importance of feedback on medical students learning of history taking but while feedback generated by humans may be subject to biases (Wagner-Menghin et al., 2020) and chatbots have been shown to be able to provide feedback (Holderried, Stegemann-Philipps, Herrmann-Werner, et al., 2024), it is possible to use chatbots to provide students with more standardised feedback. In terms of equipping the chatbot with further characteristics, it is also conceivable to train it in accordance with communication strategies to adjust it to various learning settings (Huang et al., 2025).

11.2 Which chatbot to use in serious games for training medical history taking?

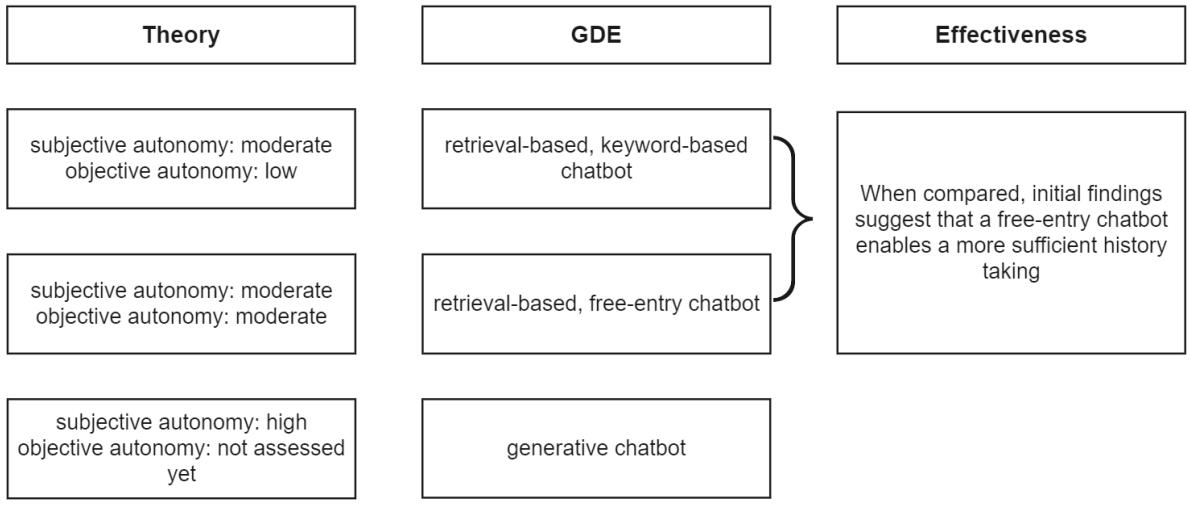
When answering the question which chatbot to use as a GDE within a SG to train medical history taking, the characteristics explained in Chapter 5 need to be taken into account. Following them, retrieval-based chatbots with free-entries as well as with keyword-based, and a generative-based chatbot were compared regarding students' autonomy. Therefore, based on the obtained results, the question can only be addressed within the context of autonomy. When focusing on students' autonomy, a generative-based chatbot would be recommended. This remains valid in light of preliminary results demonstrating that other dimensions of history taking (i.e., empathy) are trainable. However, both types of chatbots have their own strengths. On the one hand, retrieval-based chatbots are better capable of depositing the knowledge base, making them more standardisable. On the other hand, generative-based chatbots provide more human-like conversation and are based on an open knowledge base. Additionally, further aspects should be considered with which chatbots can be equipped. Since students using the keyword-based chatbot profited more from the additional instructional content, it should be considered whether chatbots for history taking should contain additional cues. However, students achieved scores did never exceed the 50% threshold. Further

research should explore potential explanations and investigate the necessary characteristics for chatbots to be used as a GDE for history taking.

Conclusively, it is necessary to undertake further research on chatbots for history taking before choosing one depending on the purpose. It is possible to separate the overarching GDE 'chatbot' within the GATE framework in more specified subcategories (e.g., a chatbot for training the procedure of history taking or factual knowledge acquisition vs. a chatbot for training empathic history taking). Figure 5 illustrates a version of the GATE framework adapted to the studies' findings regarding autonomy. Additionally, the setting the SG provides should be critically evaluated. Since SGs in medical education, and also in the presented studies, focussed on emergency medicine, this might not be the best suited environment for training history taking and other settings, such as family medicine should be considered (Ziebarth et al., 2014). Reportedly, medical students seem to be more critical towards chatbots and rarely use them for their learning process (Stöhr et al., 2024), although they simultaneously valued the use of ChatGPT as a VP and favour it over traditional training (Rädel-Ablass et al., 2025).

Figure 5

GATE framework adapted to the studies' findings



12 Conclusion

This thesis aimed at assessing under which circumstances a chatbot can be regarded as a game design element and aimed at closing the research gap regarding its theoretical foundation. Therefore, three different types of chatbots were analysed and compared. Initially, a SG was designed with a retrieval-based free-entry chatbot for history taking as it is important to examine a GDE within its supposed SG. In a following step, the newly build chatbot was compared with a previously existing more restricted chatbot also embedded in a SG in regard of elicited autonomy. For this comparison, two variables were compared, students' autonomous behaviour in terms of their exploratory behaviour within the chatbot during the history taking as well as their subjectively experienced autonomy afterwards. Although the subjective data did not reveal high levels of autonomy and no differences between the chatbots, the objective data led to the assumption that a free-entry although retrieval-based chatbot allowed more exploration and therefore autonomous behaviour. In a subsequent step it was assessed whether students might profit from additional instructional material before using the chatbot. The findings indicated that students using a keyword-based chatbot profited more from this material suggesting that a chatbot for history taking needs more than just providing a VP. Through the launch of ChatGPT, new opportunities opened up and subjective autonomy experiences were conclusively tested for history taking with ChatGPT as a VP. As hypothesised, students showed high levels of subjective autonomy after taking the history within a generative-based chatbot. However, this chatbot was tested standalone and not embedded within a SG making it difficult to generalise the findings for the use as a GDE within a SG. This needs to be further investigated to produce a stronger data base. In conclusion, students' subjective autonomy was best addressed by a generative chatbot, whereas a free-entry chatbot also encouraged exploratory behaviour in students.

Disclosures

I was employed as a research assistant at the Institute of Medical Education of the University Hospital Bonn throughout the process of developing the original studies and preparing the dissertation. No additional financial benefits or other support were granted for the conduction of the original studies referred to in this thesis. Study 1, Study 2, and Study 3 were conducted at Göttingen medical school while Study 4 was conducted at the University of Bonn. All studies received approval from the local Ethics Committees of which the respective file numbers can be found in each of the publications.

I assure that all external material was referenced accordingly by citation and entry in the reference list and confirm that I have completed this dissertation on my own. All used aids were mentioned in the respective studies or in the dissertations' main text. In rare instances the following tools were used to refine phrases: DeepL (DeepL SE, https://www.deepl.com/de/translator) and ChatGPT (OpenAI, Version 3.5, https://chat.openai.com). I declare that I have not made use of generative AI for the generation of ideas, or text that was not originally written by myself or the analysis of data at any stage of the entire dissertation project.

References

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, *132*, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Abt, C. C. (1970). *Serious games*. The Viking Press.

Adamopoulou, E., & Moussiades, L. (2020). *An overview of chatbot technology* Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, Cham.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*, 179-211.

Akl, E. A., Pretorius, R. W., Sackett, K., Erdley, W. S., Bhoopathi, P. S., Alfarah, Z., & Schunemann, H. J. (2010). The effect of educational games on medical students' learning outcomes: A systematic review: BEME Guide No 14. *Medical Teacher*, *32*(1), 16-27. https://doi.org/10.3109/01421590903473969

Alexiou, A., & Schippers, M. C. (2018). Digital game elements, user experience and learning: A conceptual framework. *Education and Information Technologies*, *23*(6), 2545-2567. https://doi.org/10.1007/s10639-018-9730-6

Alyami, H., Alawami, M., Lyndon, M., Alyami, M., Coomarasamy, C., Henning, M., Hill, A., & Sundram, F. (2019). Impact of using a 3D visual metaphor serious game to teach history-taking content to medical students: Longitudinal mixed methods pilot study. *JMIR Serious Games*, *7*(3), e13748. https://doi.org/10.2196/13748

Aparicio, A. F., Vela, F. L. G., Sánchez, J. L. G., & Montes, J. L. I. (2012). *Analysis and application of gamification* Proceedings of the 13th international conference on interacción persona-ordenador,

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., de Freitas, S., Louchart, S., Suttie, N., Berta, R., & De Gloria, A. (2014). Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology*, *46*(2), 391-411. https://doi.org/10.1111/bjet.12113

Aspegren, K. (1999). BEME Guide No. 2: Teaching and learning communication skills in medicine-a review with quality grading of articles. *Medical Teacher*, *21*(6), 563-570. https://doi.org/10.1080/01421599978979

Aster, A., Hütt, C., Morton, C., Flitton, M., Laupichler, M. C., & Raupach, T. (2024). Development and evaluation of an emergency department serious game for undergraduate medical students. *BMC Medical Education*, *24*(1), Article 1061. https://doi.org/10.1186/s12909-024-06056-z

Aster, A., Laupichler, M. C., Zimmer, S., & Raupach, T. (2024). Game design elements of serious games in the education of medical and healthcare professions: A mixed-methods systematic review of underlying theories and teaching effectiveness. *Advances in Health Sciences Education*, *29*(5), 1825-1848. https://doi.org/10.1007/s10459-024-10327-1

Aster, A., Lotz, A., Laupichler, M. C., & Raupach, T. (2025). Impact of providing a customized guideline on virtual medical history taking in two serious games for medical

education. *Medical Education Online*, *30*(1), Article 2527175. https://doi.org/10.1080/10872981.2025.2527175

Aster, A., Lotz, A., & Raupach, T. (2025). Theoretical background of the game design element “chatbot” in serious games for medical education. *Advances in Simulation*, *10*(1), Article 10. https://doi.org/10.1186/s41077-025-00341-7

Aster, A., Ragaller, S. V., Raupach, T., & Marx, A. (2025). ChatGPT as a virtual patient: Written empathic expressions during medical history taking. *Medical Science Educator*, *35*(3), 1513-1522. https://doi.org/10.1007/s40670-025-02342-7

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*(6), 589-596. https://doi.org/10.1001/jamainternmed.2023.1838

Bachmann, C., Barzel, A., Roschlaub, S., Ehrhardt, M., & Scherer, M. (2013). Can a brief two-hour interdisciplinary communication skills training be successful in undergraduate medical education? *Patient Education and Counseling*, *93*(2), 298-305. https://doi.org/10.1016/j.pec.2013.05.019

Bachmann, C., Roschlaub, S., Harendza, S., Keim, R., & Scherer, M. (2017). Medical students' communication skills in clinical education: Results from a cohort study. *Patient Education and Counseling*, *100*(10), 1874-1881. https://doi.org/10.1016/j.pec.2017.05.030

Bangor, A. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114-123.

Bartle, R. (1996). Hearts, clubs, diamonds, spades: Players who suit MUDs. *Journal of MUD research*, *1*(1), 19.

Beierlein, C., Kemper, C. J., Kovaleva, A., & Rammstedt, B. (2013). Kurzskala zur Erfassung allgemeiner Selbstwirksamkeitserwartungen (ASKU). *Methoden, Daten, Analysen (mda)*, *7*(2), 251-278. https://doi.org/10.12758/mda.2013.014

Benfatah, M., Marfak, A., Saad, E., Hilali, A., Nejjari, C., & Youlyouz-Marfak, I. (2024). Assessing the efficacy of ChatGPT as a virtual patient in nursing simulation training: A study on nursing students' experience. *Teaching and Learning in Nursing*, *19*(3), e486-e493. https://doi.org/10.1016/j.teln.2024.02.005

Bird, J., & Cohen-Cole, S. A. (1990). The three-function model of the medical interview. An educational device. In M. S. Hale (Ed.), *Methods in teaching consultation-liasion psychiatry* (Vol. 20, pp. 65-88). Karger.

Blohm, I., & Leimeister, J. M. (2013). Gamification: Design of IT-based enhancing services for motivational support and behavioral change. *Business & Information Systems Engineering*, *5*(4), 275-278. https://doi.org/10.1007/s12599-013-0273-5

Botella, C., Breton-López, J., Quero, S., Baños, R. M., García-Palacios, A., Zaragoza, I., & Alcaniz, M. (2011). Treating cockroach phobia using a serious game on a mobile phone and augmented reality exposure: A single case study. *Computers in Human Behavior*, *27*(1), 217-227. https://doi.org/10.1016/j.chb.2010.07.043

Brodahl, K. O., Storoy, H. E., Finset, A., & Pedersen, R. (2022). Medical students' experiences when empathizing with patients' emotional issues during a medical

interview - a qualitative study. *BMC Medical Education*, *22*(1), 145. https://doi.org/10.1186/s12909-022-03199-9

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. A. Thomas, I. L. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189 - 194). Taylor & Francis.

Bylund, C. L., & Makoul, G. (2002). Empathic communication and gender in the physician-patient encounter. *Patient Education and Counseling*, *48*(3), 207-216.

Bylund, C. L., & Makoul, G. (2005). Examining empathy in medical encounters: An observational study using the empathic communication coding system. *Health Communication*, *18*(2), 123-140.

Calik, A., Cakmak, B., Kapucu, S., & Inkaya, B. (2022). The effectiveness of serious games designed for infection prevention and promotion of safe behaviors of senior nursing students during the COVID-19 pandemic. *American Journal of Infection Control*, *50*(12), 1360-1367. https://doi.org/10.1016/j.ajic.2022.02.025

Cambridge Dictionary. (2025). Chatbot. In *Cambridge Dictionary*.

Cleland, J. A., Abe, K., & Rethans, J. J. (2009). The use of simulated patients in medical education: AMEE Guide No 42. *Medical Teacher*, *31*(6), 477-486. https://doi.org/10.1080/01421590903002821

Continisio, G. I., Serra, N., Guillari, A., Simeone, S., Lucchese, R., Gargiulo, G., Toscano, S., Capo, M., Capuano, A., Sarracino, F., Esposito, M. R., & Rea, T. (2021). Evaluation of soft skills among italian healthcare rehabilitators: A cross sectional study. *Journal of Public Health Research*, *10*(3). https://doi.org/10.4081/jphr.2021.2002

Crossley, C., Fanfarelli, J. R., & McDaniel, R. (2016). User experience design considerations for healthcare games and applications. 2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH),

Csikszentmihalyi, M. (1975). *Beyond boredeom and anxiety: Experiencing flow in work and play*. Jossey-Bass.

Dankbaar, M. E., Richters, O., Kalkman, C. J., Prins, G., Ten Cate, O. T., van Merrienboer, J. J., & Schuit, S. C. (2017). Comparative effectiveness of a serious game and an e-module to support patient safety knowledge and awareness. *BMC Medical Education*, *17*(1), 30. https://doi.org/10.1186/s12909-016-0836-5

Dankbaar, M. E., Roozeboom, M. B., Oprins, E. A., Rutten, F., van Merrienboer, J. J., van Saase, J. L., & Schuit, S. C. (2017). Preparing residents effectively in emergency skills training with a serious game. *Simulation in Healthcare*, *12*(1), 9-16. https://doi.org/10.1097/SIH.0000000000000194

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, *11*(4), 227-268.

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol. 1, pp. 416-436). SAGE.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). *From game design elements to gamefulness: Defining "gamification"* Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments,

Dörner, R., S., G., Effelsberg, W., & Wiemeyer, J. (2016). Introduction. In R. Dörner, S. Göbel, W. Effelsberg, & J. Wiemeyer (Eds.), *Serious Games*. Springer. https://doi.org/https://doi.org/10.1007/978-3-319-40612-1_1

Edwards, S. L., Zarandi, A., Cosimini, M., Chan, T. M., Abudukebier, M., & Stiver, M. L. (2025). Analog Serious Games for Medical Education: A Scoping Review. *Academic Medicine*, *100*(3), 375-387. https://doi.org/10.1097/ACM.0000000000005911

Emerson, A., Cloude, E. B., Azevedo, R., & Lester, J. (2020). Multimodal learning analytics for game-based learning. *British Journal of Educational Technology*, *51*(5), 1505-1526. https://doi.org/10.1111/bjet.12992

Furstenberg, S., Helm, T., Prediger, S., Kadmon, M., Berberat, P. O., & Harendza, S. (2020). Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. *BMC Medical Education*, *20*(1), 368. https://doi.org/10.1186/s12909-020-02260-9

Gärtner, J., Bussenius, L., Schick, K., Prediger, S., Kadmon, M., Berberat, P. O., & Harendza, S. (2022). Validation of the ComCare index for rater-based assessment of medical communication and interpersonal skills. *Patient Education and Counseling*, *105*(4), 1004-1008. https://doi.org/10.1016/j.pec.2021.07.051

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. Palgrave / Macmillan.

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-Learning*, *2*(1), 5-16.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237-288.

Graafland, M., Bemelman, W. A., & Schijven, M. P. (2017). Game-based training improves the surgeon's situational awareness in the operation room: A randomized controlled trial. *Surgical Endoscopy*, *31*(10), 4093-4101. https://doi.org/10.1007/s00464-017-5456-6

Graafland, M., Schraagen, J. M., & Schijven, M. P. (2012). Systematic review of serious games for medical education and surgical skills training. *Journal of British Surgery*, *99*(10), 1322-1330. https://doi.org/10.1002/bjs.8819

Habicht, J., Viswanathan, S., Carrington, B., Hauser, T. U., Harper, R., & Rollwage, M. (2024). Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nature Medicine*, *30*(2), 595-602. https://doi.org/10.1038/s41591-023-02766-x

Hamari, J., & Keronen, L. (2017). Why do people play games? A meta-analysis. *International Journal of Information Management*, *37*(3), 125-141. https://doi.org/10.1016/j.ijinfomgt.2017.01.006

Hampton, J. R., Harrison, M. J. G., Mitchell, J. R. A., Prichard, J. S., & Seymour, C. (1975). Relative contributions of history-taking, physical examination, and laboratory

investigation to diagnosis and management of medical outpatients. *British Medical Journal*, *2*(5969), 486-489.

Hartt, M., Hosseini, H., & Mostafapour, M. (2020). Game on: Exploring the effectiveness of game-based learning. *Planning Practice & Research*, *35*(5), 589-604. https://doi.org/10.1080/02697459.2020.1778859

Higham, P. A., & Guzel, M. A. (2012). Cued Recall. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning*. Springer. https://doi.org/10.1007/978-1-4419-1428-6

Holderried, F., Stegemann-Philipps, C., Herrmann-Werner, A., Festl-Wietek, T., Holderried, M., Eickhoff, C., & Mahling, M. (2024). A Language Model-powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Medical Education*, *10*, e59213. https://doi.org/10.2196/59213

Holderried, F., Stegemann-Philipps, C., Herschbach, L., Moldt, J. A., Nevins, A., Griewatz, J., Holderried, M., Herrmann-Werner, A., Festl-Wietek, T., & Mahling, M. (2024). A Generative Pretrained Transformer (GPT)-powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR Medical Education*, *10*, e53961. https://doi.org/10.2196/53961

Hsu, C.-L. (2022). Applying cognitive evaluation theory to analyze the impact of gamification mechanics on user engagement in resource recycling. *Information & Management*, *59*(2). https://doi.org/10.1016/j.im.2022.103602

Huang, W., Jiang, J., King, R. B., & Fryer, L. K. (2025). Chatbots and student motivation: A scoping review. *International Journal of Educational Technology in Higher Education*, *22*(1). https://doi.org/10.1186/s41239-025-00524-2

Hwang, G.-J., & Chang, C.-Y. (2021). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, *31*(7), 4099-4112. https://doi.org/10.1080/10494820.2021.1952615

Isaza-Restrepo, A., Gomez, M. T., Cifuentes, G., & Arguello, A. (2018). The virtual patient as a learning tool: A mixed quantitative qualitative study. *BMC Medical Education*, *18*(1), 297. https://doi.org/10.1186/s12909-018-1395-8

Jonassen, D. (1999). Designing constructivist learning environments. In C. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory* (pp. 215-239).

Kaplonyi, J., Bowles, K. A., Nestel, D., Kiegaldie, D., Maloney, S., Haines, T., & Williams, C. (2017). Understanding the impact of simulated patients on health care learners' communication skills: A systematic review. *Medical Education*, *51*(12), 1209-1219. https://doi.org/10.1111/medu.13387

Kassirer, J. P. (2010). Teaching clinical reasoning: Case-based and coached. *Academic Medicine*, *85*(7), 1118-1124.

Kayyali, R., Wells, J., Rahmtullah, N., Tahsin, A., Gafoor, A., Harrap, N., & Nabhani-Gebara, S. (2021). Development and evaluation of a serious game to support learning among pharmacy and nursing students. *Currents in Pharmacy Teaching and Learning*, *13*(8), 998-1009. https://doi.org/10.1016/j.cptl.2021.06.023

Kegel, M., Klee, O., Herrmann, T., & Dietz-Wittstock, M. (2022). Beobachtung und Beurteilung von Patienten in der Notaufnahme. In M. Dietz-Wittstock, M. Kegel, P.

Glien, & M. Pin (Eds.), *Notfallpflege - Fachweiterbildung und Praxis*. Springer, Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-662-63461-5_6

Keifenheim, K. E., Teufel, M., Ip, J., Speiser, N., Leehr, E. J., Zipfel, S., & Herrmann-Werner, A. (2015). Teaching history taking to medical students: A systematic review. *BMC Medical Education*, *15*, 159. https://doi.org/10.1186/s12909-015-0443-x

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of training evaluation*. Association for Talent Development.

Klauß, H., Kunkel, A., Mussgens, D., Haaker, J., & Bingel, U. (2024). Learning by observing: A systematic exploration of modulatory factors and the impact of observationally induced placebo and nocebo effects on treatment outcomes. *Frontiers in Psychology*, *15*, 1293975. https://doi.org/10.3389/fpsyg.2024.1293975

Klein, R., Julian, K. A., Koch, J., Snyder, E. D., Jassal, S., Simon, W., Millard, A., Uthlaut, B., Burnett-Bowie, S. M., Ufere, N. N., Alba-Nguyen, S., Volerman, A., Thompson, V., Kumar, A., White, B. A., Park, Y. S., Palamara, K., & Gender Equity in Medicine, W. (2024). Gender differences in clinical performance assessment of internal medicine residents: A longitudinal analysis of the influence of faculty and trainee gender. *Academic Medicine*, *99*(12), 1413-1422. https://doi.org/10.1097/ACM.0000000000005884

Kleinsmith, A., Rivera-Gutierrez, D., Finney, G., Cendan, J., & Lok, B. (2015). Understanding empathy training with virtual patients. *Computers in Human Behavior*, *52*, 151-158. https://doi.org/10.1016/j.chb.2015.05.033

Knowles, M. S. (1975). *Self-directed learning: A guide for learners and teachers*. Association Press.

Koelewijn, G., Hennus, M. P., Kort, H. S. M., Frenkel, J., & van Houwelingen, T. (2024). Games to support teaching clinical reasoning in health professions education: A scoping review. *Medical Education Online*, *29*(1), 2316971. https://doi.org/10.1080/10872981.2024.2316971

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice Hall.

Kowalewski, K. F., Hendrie, J. D., Schmidt, M. W., Proctor, T., Paul, S., Garrow, C. R., Kenngott, H. G., Muller-Stich, B. P., & Nickel, F. (2017). Validation of the mobile serious game application Touch Surgery for cognitive training and assessment of laparoscopic cholecystectomy. *Surgical Endoscopy*, *31*(10), 4058-4066. https://doi.org/10.1007/s00464-017-5452-x

Krath, J., Schürmann, L., & von Korflesch, H. F. O. (2021). Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior*, *125*. https://doi.org/10.1016/j.chb.2021.106963

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2022). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973-1018. https://doi.org/10.1007/s10639-022-11177-3

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education*, *20*(1). https://doi.org/10.1186/s41239-023-00426-1

Landers, R. N. (2015). Developing a theory of gamified learning. *Simulation & Gaming*, *45*(6), 752-768. https://doi.org/10.1177/1046878114563660

Laugwitz, B., Held, T., & Schrepp, M. (2008, 2008//). Construction and evaluation of a user experience questionnaire. HCI and Usability for Education and Work, Berlin, Heidelberg.

Lee, M., Shin, S., Lee, M., & Hong, E. (2024). Educational outcomes of digital serious games in nursing education: A systematic review and meta-analysis of randomized controlled trials. *BMC Medical Education*, *24*(1), 1458. https://doi.org/10.1186/s12909-024-06464-1

Liaw, S. Y., Tan, J. Z., Lim, S., Zhou, W., Yap, J., Ratan, R., Ooi, S. L., Wong, S. J., Seah, B., & Chua, W. L. (2023). Artificial intelligence in virtual reality simulation for interprofessional communication training: Mixed method study. *Nurse Education Today*, *122*, 105718. https://doi.org/10.1016/j.nedt.2023.105718

Lippitsch, A., Steglich, J., Ludwig, C., Kellner, J., Hempel, L., Stoevesandt, D., & Thews, O. (2024). Development and evaluation of a software system for medical students to teach and practice anamnestic interviews with virtual patient avatars. *Comput Methods and Programs in Biomedicine*, *244*, 107964. https://doi.org/10.1016/j.cmpb.2023.107964

Loyens, S. M. M., Magda, J., & Rikers, R. M. J. P. (2008). Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational Psychology Review*, *20*(4), 411-427. https://doi.org/10.1007/s10648-008-9082-7

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quaterly for exercise and sport*, *60*(1), 48-58.

Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, *71*, 525-534. https://doi.org/10.1016/j.chb.2015.08.048

Michael, D. R., & Chen, S. L. (2005). *Serious Games: Games That Educate, Train, and Inform*. Muska & Lipman/Premier-Trade.

Middeke, A., Anders, S., Raupach, T., & Schuelper, N. (2020). Transfer of clinical reasoning trained with a serious game to comparable clinical problems: A prospective randomized study. *Simulation in Healthcare*, *15*(2), 75-81. https://doi.org/10.1097/SIH.0000000000000407

Middeke, A., Anders, S., Schuelper, M., Raupach, T., & Schuelper, N. (2018). Training of clinical reasoning with a Serious Game versus small-group problem-based learning: A prospective study. *PLoS One*, *13*(9), e0203851. https://doi.org/10.1371/journal.pone.0203851

Molloy, M. A., Holt, J., Charnetski, M., & Rossler, K. (2021). Healthcare Simulation Standards of Best PracticeTM Simulation Glossary. *Clinical Simulation in Nursing*, *58*, 57-65. https://doi.org/10.1016/j.ecns.2021.08.017

Moore, J. (2011). Behaviorism. *Psychological Record*, *61*(3), 449-464.

Nardone, D. A., Reuler, J. B., & Girard, D. E. (1980). Teaching history-taking: Where are we? *The Yale Journal of Biology and Medicine*, *53*, 233-250.

Nayak, A., Alkaitis, M. S., Nayak, K., Nikolov, M., Weinfurt, K. P., & Schulman, K. (2023). Comparison of history of present illness summaries generated by a chatbot and senior internal medicine residents. *JAMA Internal Medicine*, *183*(9), 1026-1027.

Neufeld, A., & Malin, G. (2024). Cultivating physician empathy: A person-centered study based in self-determination theory. *Medical Education Online*, *29*(1), 2335739. https://doi.org/10.1080/10872981.2024.2335739

Niemiec, C. P., & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom. *Theory and Research in Education*, *7*(2), 133-144. https://doi.org/10.1177/1477878509104318

Olszewski, A. E., & Wolbrink, T. A. (2017). Serious gaming in medical education: A proposed structured framework for game development. *Simulation in Healthcare*, *12*(4), 240-253. https://doi.org/10.1097/SIH.0000000000000212

OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/index/chatgpt/

Orland, B., Ram, N., Lang, D., Houser, K., Kling, N., & Coccia, M. (2014). Saving energy in an office environment: A serious game intervention. *Energy and Buildings*, *74*, 43-52. https://doi.org/10.1016/j.enbuild.2014.01.036

Peterson, M. C., Holbrook, J. H., Von Hales, D. E., Smith, N. L., & Staker, L. V. (1992). Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. *Western Journal of Medicine*, *156*(2), 163-165.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, *50*(4), 258-283. https://doi.org/10.1080/00461520.2015.1122533

Prensky, M. (2001). *Digital game-based learning*. McGraw-Hill.

Qian, M., & Clark, K. R. (2016). Game-based learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, *63*, 50-58. https://doi.org/10.1016/j.chb.2016.05.023

Quiroga Pérez, J., Daradoumis, T., & Marquès Puig, J. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549-1565. https://doi.org/10.1002/cae.22326

Rädel-Ablass, K., Schliz, K., Schlick, C., Meindl, B., Pahr-Hosbach, S., Schwendemann, H., Rupp, S., Roddewig, M., & Miersch, C. (2025). Teaching opportunities for anamnesis interviews through AI based teaching role plays: a survey with online learning students from health study programs. *BMC Medical Education*, *25*(1), 259. https://doi.org/10.1186/s12909-025-06756-0

Rasch, D., & Guiard, V. (2004). The Robustness of Parametric Statistical Methods. *Psychology Science*, *46*, 175-208.

Raupach, T., Munscher, C., Beissbarth, T., Burckhardt, G., & Pukrop, T. (2011). Towards outcome-based programme evaluation: Using student comparative self-assessments to determine teaching effectiveness. *Medical Teacher*, *33*(8), e446-453. https://doi.org/10.3109/0142159X.2011.586751

Rutter, D. R., & Maguire, G. P. (1976). History-taking for medical students - II - Evaluation of a training programm. *The Lancet*, *308*(7985), 558 - 560.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68-78.

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Publications.

Sailer, M. (2016). *Die Wirkung von Gamification auf Motivation und Leistung: Empirische Studien im Kontext manueller Arbeitsprozesse* (1 ed.). Springer. https://doi.org/https://doi.org/10.1007/978-3-658-14309-1

Sailer, M., Hense, J., Mandl, H., & Klevers, M. (2013). Psychological perspectives on motivation through gamification. *Interaction Design and Architecture(s) Journal - IxD&A*, *19*, 28-37.

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371-380. https://doi.org/10.1016/j.chb.2016.12.033

Savazzi, F., Isernia, S., Jonsdottir, J., Di Tella, S., Pazzi, S., & Baglio, F. (2018). Engaged in learning neurorehabilitation: Development and validation of a serious game with user-centered design. *Computers & Education*, *125*, 53-61. https://doi.org/10.1016/j.compedu.2018.06.001

Schrepp, M., Thomaschewski, J., & Hinderks, A. (n.d.). *UEQ - User Experience Questionnaire*. ueq-online.org

Schutte, N. S., & Malouff, J. M. (2019). Increasing curiosity through autonomy of choice. *Motivation and Emotion*, *43*(4), 563-570. https://doi.org/10.1007/s11031-019-09758-w

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14-31. https://doi.org/10.1016/j.ijhcs.2014.09.006

Setyonugroho, W., Kennedy, K. M., & Kropmans, T. J. (2015). Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Education and Counseling*. https://doi.org/10.1016/j.pec.2015.06.004

Sezer, B., Sezer, T. A., Teker, G. T., & Elcin, M. (2023). Developing a virtual patient: Design, usability, and learning effect in communication skills training. *BMC Medical Education*, *23*(1), 891. https://doi.org/10.1186/s12909-023-04860-7

Sikström, P., Valentini, C., Sivunen, A., & Kärkkäinen, T. (2024). Pedagogical agents communicating and scaffolding students' learning: High school teachers' and students' perspectives. *Computers & Education*, *222*. https://doi.org/10.1016/j.compedu.2024.105140

Skinner, B. F. (1953). *Science and human behavior*. Pearson Education, Inc.

Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education*, *151*. https://doi.org/10.1016/j.compedu.2020.103862

Stevens, A., Hernandez, J., Johnsen, K., Dickerson, R., Raij, A., Harrison, C., DiPietro, M., Allen, B., Ferdig, R., Foti, S., Jackson, J., Shin, M., Cendan, J., Watson, R., Duerson, M., Lok, B., Cohen, M., Wagner, P., & Lind, D. S. (2006). The use of virtual patients to

teach medical students history taking and communication skills. *The American Journal of Surgery*, *191*(6), 806-811. https://doi.org/10.1016/j.amjsurg.2006.03.002

Stöhr, C., Ou, A. W., & Malmström, H. (2024). Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study. *Computers and Education: Artificial Intelligence*, *7*. https://doi.org/10.1016/j.caeai.2024.100259

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games: An overview. https://www.diva-portal.org/smash/get/diva2:2416/FULLTEXT01.pdf

Sweller, J. (2010). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Thomas, T. H., Bender, C., Donovan, H. S., Murray, P. J., Taylor, S., Rosenzweig, M., Sereika, S. M., Brufsky, A., & Schenker, Y. (2023). The feasibility, acceptability, and preliminary efficacy of a self-advocacy serious game for women with advanced breast or gynecologic cancer. *Cancer*, *129*(19), 3034-3043. https://doi.org/10.1002/cncr.34887

Tsoy, D., Sneath, P., Rempel, J., Huang, S., Bodnariuc, N., Mercuri, M., Pardhan, A., & Chan, T. M. (2019). Creating GridlockED: A serious game for teaching about multipatient environments. *Academic Medicine*, *94*(1), 66-70. https://doi.org/10.1097/ACM.0000000000002340

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of verbal learning and verbal behavior*, *5*(4), 381-191.

Vogel, D., Meyer, M., & Harendza, S. (2018). Verbal and non-verbal communication skills including empathy during history taking of undergraduate medical students. *BMC Medical Education*, *18*(1), 157. https://doi.org/10.1186/s12909-018-1260-9

Wagner-Menghin, M., de Bruin, A. B. H., & van Merrienboer, J. J. G. (2020). Communication skills supervisors' monitoring of history-taking performance: An observational study on how doctors and non-doctors use cues to prepare feedback. *BMC Medical Education*, *20*(1), 36. https://doi.org/10.1186/s12909-019-1920-4

Whittaker, L., Russell-Bennett, R., & Mulcahy, R. (2021). Reward-based or meaningful gaming? A field study on game mechanics and serious games for sustainability. *Psychology & Marketing*, *38*(6), 981-1000. https://doi.org/10.1002/mar.21476

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? - A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, *4*, 654924. https://doi.org/10.3389/frai.2021.654924

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89-100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wu, J. H., Du, J. K., & Lee, C. Y. (2021). Development and questionnaire-based evaluation of virtual dental clinic: A serious game for training dental students. *Medical Education Online*, *26*(1), 1983927. https://doi.org/10.1080/10872981.2021.1983927

Wu, R., & Yu, Z. (2023). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology*, *55*(1), 10-33. https://doi.org/10.1111/bjet.13334

Wu, W.-H., Chiou, W.-B., Kao, H.-Y., Alex Hu, C.-H., & Huang, S.-H. (2012). Re-exploring game-assisted learning research: The perspective of learning theoretical bases. *Computers & Education*, *59*(4), 1153-1161. https://doi.org/10.1016/j.compedu.2012.05.003

Wu, W. H., Hsiao, H. C., Wu, P. L., Lin, C. H., & Huang, S. H. (2011). Investigating the learning-theory foundations of game-based learning: A meta-analysis. *Journal of Computer Assisted Learning*, *28*(3), 265-279. https://doi.org/10.1111/j.1365-2729.2011.00437.x

Young, M., Thomas, A., Lubarsky, S., Ballard, T., Gordon, D., Gruppen, L. D., Holmboe, E., Ratcliffe, T., Rencic, J., Schuwirth, L., & Durning, S. J. (2018). Drawing boundaries: The difficulty in defining clinical reasoning. *Academic Medicine*, *93*(7), 990-995. https://doi.org/10.1097/ACM.0000000000002142

Ziebarth, S., Kizina, A., Hoppe, H. U., & Dini, L. (2014). *A Serious Game for Training Patient-Centered Medical Interviews* 2014 IEEE 14th International Conference on Advanced Learning Technologies,

Original publications

The original publications listed under “List of references for Studies 1 to 4” can be found in the following. The publications were not altered in any way and are listed the way they can be retrieved from the journal’s website. All studies were published open access.

Aster *et al. BMC Medical Education* (2024) 24:1061
https://doi.org/10.1186/s12909-024-06056-z

BMC Medical Education

RESEARCH Open Access

Development and evaluation of an emergency department serious game for undergraduate medical students

Alexandra Aster[1*], Christopher Hütt[1,2], Caroline Morton[3], Maxwell Flitton[3], Matthias Carl Laupichler[1] and Tobias Raupach[1]

Abstract

Background Serious games are risk-free environments training various medical competencies, such as clinical reasoning, without endangering patients' safety. Furthermore, serious games provide a context for training situations with unpredictable outcomes. Training these competencies is particularly important for healthcare professionals in emergency medicine.

Methods Based on these considerations, we designed, implemented, and evaluated a serious game in form of an emergency department, containing the features of a virtual patient generator, a chatbot for medical history taking with self-formulated questions, artificially generated faces based on an artificial intelligence algorithm, and feedback for students. The development process was based on an already existing framework resulting in an iterative procedure between development and evaluation. The serious game was evaluated using the System Usability Scale and the User Experience Questionnaire.

Results The System Usability Scale provided a substantial result for the usability. In terms of the user experience, four scales yielded positive results, whereas two scales yielded neutral results.

Conclusion The evaluation of both usability and user experience yielded overall positive results, while simultaneously identifying potential areas for improvement. Further studies will address the implementation of additional game design elements, and testing student learning outcome.

Keywords Serious game, Gamification, Medical education, Clinical reasoning, Game development, Usability, User experience

*Correspondence:
Alexandra Aster
alexandra.aster@ukbonn.de
[1]Institute of Medical Education, University Hospital Bonn, Venusberg-Campus 1, 53127 Bonn, Germany
[2]Department of Anesthesiology and Intensive Care Medicine, University Hospital Bonn, Bonn, Germany
[3]Yellow Bird Consulting Ltd, London, UK

Background

Serious games are known to be safe, cost and time effective learning environments, which are used in various application areas, including the healthcare sector [1–3]. Besides, there is an urgent need to create digital learning environments [4] to motivate learners, as more and more users are digital natives and it is assumed that serious games are particularly helpful for enhancing the learning outcomes of this user group [5]. In the healthcare sector, serious games provide a risk-free environment for tasks that might put patients at risk or have an unpredictable outcome, while maintaining a high fidelity [6, 7]. Particularly noteworthy is the psychological fidelity, which reflects the extent to which the experience of psychological factors (e.g., stress) are simulated similarly to the real environment [8]. Hence, serious games seem to have a significant effect on learning outcomes with regard to patient safety, and help students feel more confident and eased in real-life situations afterwards [9].

For contexts such as medicine, healthcare professions, or even patient education the use of serious games proves to be an effective teaching method in terms of learning outcomes [6, 10]. According to the frequently referenced definition by Michael and Chen [11, p. 21] serious games are defined as "games that do not have entertainment, enjoyment, or fun as their primary purpose". Therefore, the fundamental difference is that serious games primarily focus on learning objectives and simultaneously contain entertaining elements, whereas the sole goal of entertaining games is to elicit amusement in the players [12] without having a primary learning objective. Concerning digital serious games, Laamarti, Eid [12, p. 4] provide the refined definition of "serious games as an application with three components: experience, entertainment, and multimedia […]". Conclusively, serious games have to be contrasted with the related concepts 'gamification' and 'game-based learning'. Contrary to the fully fledged serious game, 'gamification' only implicates the addition of game elements to non-game contexts [13]. 'Game-based learning' instead can be understood as the pedagogical approach of incorporating games into curricula with serious games being its operationalization [14].

In the particular context of medical education, serious games offer the opportunity to gather knowledge as well as abilities in a "safe space" without the risk of endangering the health of real patients [6, 15]. Moreover, serious games allow for fostering and strengthening non-technical skills (e.g. communication or coping with stress) or knowledge about patient safety in medical students [9, 16]. Strengthening these and other non-technical skills (e.g. teamwork) is of great importance for successful work in dynamic environments, such as those found in an emergency department [17]. Serious games are already used to strengthen and enhance teamwork between disciplines such as medical and nursing undergraduates in the emergency room [18]. Besides, the development of distinct and crucial clinical reasoning competencies (i.e. proposing a suspected diagnosis as well as initiating necessary investigations and appropriate therapies, [19]) is essential for physicians in emergency departments. Clinical reasoning competencies comprise a holistic view of the patient, including the surrounding factors, as well as the adaptation to altering circumstances [20]. Since medical and healthcare professionals are confronted with difficulties and biases during clinical decision-making, it is necessary to further teach and train those competencies [20]. It has already been shown, that serious games can work as an effective method for training clinical reasoning in medical students [21] as well as in other healthcare professions education such as nurse education [22].

It is essential to find an appropriate learning environment for teaching highly relevant skills to medical students in order to prepare them for working in an emergency department. The implementation of a valid emergency department simulation in face-to-face teaching (e.g. with simulated patients) is hardly viable, hence the idea to create a suitable serious game arose. Fostering learning achievements while simultaneously evoking fun and entertainment depends fundamentally on a structured development and evaluation process of the serious game [23]. Therefore, Olszewski and Wolbrink [24] proposed a structured framework for serious game development in medical education, which we applied for the development of our presented serious game. The framework consists of three iterative phases, namely Preparation & Design, Development, and Formative Evaluation. The current development stage of our serious game is in the transition between phase two and three, more precisely in iteration loops between usability testing and the ongoing development. Usability is therefore understood as defined in ISO 9241-11 as "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [25, p. 269]. User Experience is also understood according to the definition by ISO 9241−210 as "a person's perceptions and responses that result from the use and/or anticipated use of a product, system or service." [26, p. 1]. The following section describes the endeavor and the current development status in further detail. Finally, an evaluation process with medical students and its results are presented.

Development and construction of the serious game

In this paper, we summarize the development and evaluation process of the serious game "DIVINA" and provide a prospect on further steps and possible studies. The term "DIVINA" is an acronym of the German 'DIgitale

VIrtuelle NotAufnahme', which translates to 'digital virtual emergency department'.

Design and development

The development stage started in 2020 with a design phase. A design team was put together with a software developer who is also a physician (from the commercial company leading the development) and two additional physicians. The concept here was to use co-design where stakeholders (in this case medical educators and doctors) were involved from the beginning in the design of the product. Up to three medical students assisted the design team and helped enter disease-specific data (see below). Since medical students are supposed to be the prospective users, there is a growing claim to include them in a participatory way in the development process [27]. It is recommended to involve end-users in many steps in the development process to take into account user expectations and facilitate a proper user experience [28]. In addition, students can support game development with regard to adherence to design principles (e.g. goals, feedback, rewards, as well as more general narratives and aesthetics) from a user perspective [29]. Thus, the combined perceptions of creators and end-users yield a holistic approach.

Later during the ongoing process, a psychologist joined the design team to supervise the design process and provide support regarding psychological background knowledge. The focus was on design elements to promote learning processes and outcomes. Our interdisciplinary approach to game development (i.e. software developer, physicians, students, psychologists) was in line with current recommendations as it holds a number of advantages [16]. One of the main advantages of our team is the software developer simultaneously being a physician. Hence it is unlikely that information or expert knowledge gets lost due to communication difficulties or misconceptions [30].

Educational content and learning objectives

Clinical reasoning, according to Kassirer [19], covers on the one hand the competencies of formulating a suspected diagnosis based on the patients' medical history and contrasting it against differential diagnoses. On the other hand, it is about initiating necessary investigations to confirm the suspected diagnosis and initiating an appropriate initial therapy. Fostering the competencies of clinical reasoning represents one of the main learning objectives of the serious game presented here. Another important non-technical skill is coping with stress. Nevertheless, the right decisions regarding the urgency of patients' symptoms and consequently the order of treatments should be made. Thus, a further learning objective is prioritizing patients and initiating the necessary medical procedures under time pressure.

The player is placed in the situation of being a physician currently working at an emergency department. The game starts with a variety of patients appearing on a dashboard (representing the waiting room or arrivals via ambulance). Subsequently, students have to admit patients according to the perceived urgency of the situation. Once a patient has been assigned to a treatment room, students can perform the following actions: taking a medical history via a chatbot, measuring vital signs, arranging diagnostic tests including interpretation of findings, performing a physical examination, ordering laboratory tests, prescribing medication and further measures. Before patients can be discharged or transferred (e.g. to their home, a normal ward or to an intensive / intermediate care unit), students have to complete discharge notes and choose a diagnosis. After discharging or transferring a patient, students are provided with static feedback on the specific disease of that patient. Constantly incoming new patients in the dashboard with different presenting complaints and symptoms create time pressure.

Diseases are not restricted to one particular area of medicine (e.g., cardiology, gastroenterology or gynecology). The structure of the database facilitates the addition of diseases related to all specialties lending themselves to be included in a virtual accident and emergency department. This leads to a current number of 50 implemented diseases.

Target group

Undergraduate medical students are the primary target group of this digital teaching resource. However, the game is not solely intended for asynchronous self-directed learning. Instead, gaming 'sessions' containing specific diseases need to be created by teachers and made available to students during synchronous learning sessions. Depending on curricular requirements and opportunities, actual gaming sessions may be accompanied by teachers via online communication services (lends itself for larger groups) or in a small-group setting.

Structure of the game

The serious game was programmed in Python, Rust and React, with a data pipeline set up via GitHub. The serious game is playable in every common browser. In the interest of maximal flexibility for users, and in order to avoid disruptions due to server overload, the design team decided to use a 2D- rather than a 3D-Design. Currently, the serious game is available to all German medical students via a DFN (Deutsches Forschungsnetz) login that recognizes students at all German medical schools. For a presentation of the game interface, see Fig. 1.

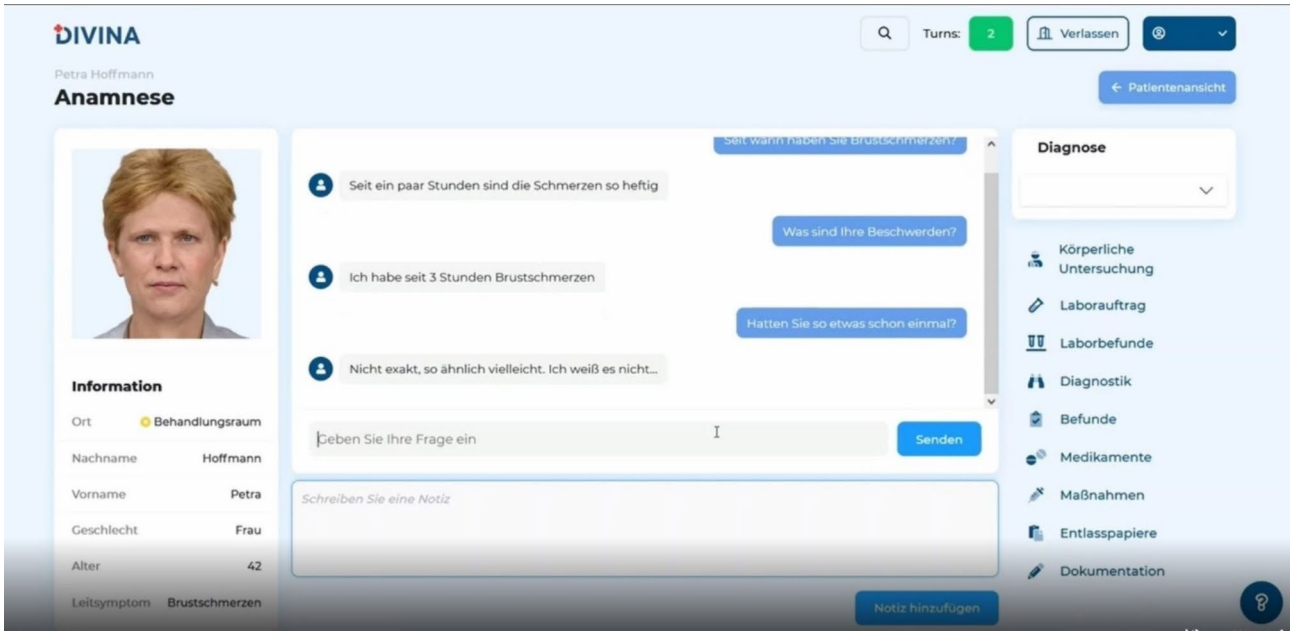


Fig. 1 DIVINA interface in German. On the left side, information about the patient, including given name, name, age, and leading symptom as well as the AI generated face is depicted. The central element is the chatbot. On the right side, all other necessary actions can be found. The information contained in the picture is meant for demonstration purposes only. Neither the image, nor the depicted information represent real patient data

The first innovative feature of this digital resource is the way in which virtual patients (VP) are generated. To build a VP suffering from a specific disease an algorithm refers to the deposited epidemiological data. Due to the disease generation stemming from epidemiological probabilities, each VP is unique that prevents easy identification of the disease. Thus, it is secured that the learning process is not only driven by recognition, as students cannot easily recall a VP from an earlier gaming session. At the same time, the VPs provide recurring learning scenarios and offer the possibility for regular knowledge repetition as it is recommended for learning environments in general, but especially for simulated ones [31]. In addition, artificial faces generated by a style-based generator using an artificial intelligence (AI) algorithm [32] complement the VPs and illustrate the disease by matching the symptoms. Faces appear either as healthy, or pale, reddish, cyanotic, or jaundiced. This enhances the psychological fidelity of the environment as it allows replicating the real emergency department environment to a certain extent [8]. It is also noteworthy, that the usage of AI-generated faces prevents data privacy problems accompanying the usage of real patient faces.

Another innovative feature is the chatbot that offers students the opportunity to enter questions related to the patients' anamnesis on which the chatbot provides automatically generated answers. Whereas most other solutions offer a question menu to choose from, the chatbot in DIVINA allows for freely formulated questions. This feature sets DIVINA apart, as students have the opportunity to learn history taking by considering questions and formulating them on their own. The questions asked by students can relate to the presenting complaint, associated symptoms, history of presenting complaint, past-medical history, as well as drug and social history. Furthermore, the chatbot offers the chance for teachers to analyze how students approach the history taking, yielding potential for further teaching on this topic. Due to the necessity of asking self-generated questions and considering all relevant topics, the chatbot aims to foster clinical reasoning competencies.

Evaluation

Methods

In order to test the usability and user experience of the proposed serious game, we collected data during a study in summer term 2022 with third-year medical students (N=146) participating. The local Institutional Review Board reviewed the study protocol in winter term 2021/2022 (application number 34/8/21). The study was embedded in a six-week cardiology and pneumology module at Goettingen Medical School. Students participating in the module played the game in two sessions lasting 90 min each. The content of the respective game sessions were taught in the formal teaching sessions the week before. The evaluation of usability and user experience formed the conclusion of the study, for which students gave their informed consent beforehand.

Usability was assessed using the 5-point Likert-scaled System Usability Scale (SUS, [33]) and user experience

was evaluated via the User Experience Questionnaire (UEQ, [34]). The SUS offers a result scale between 0 and 100 with an average score set around 68 [35]. Regarding the UEQ, two opposing extreme response options were rated on a seven-stage scale ranging from −3 to +3, with −3 being the negative extreme, 0 neutral, and +3 the positive extreme [36]. Taken together the UEQ comprises six different scales, namely Attractiveness (i.e. overall impression), Perspicuity (i.e. easy familiarization), Efficiency (i.e. efficient task solving), Dependability (i.e. feeling of control regarding interaction), Stimulation (i.e. exciting and motivating usage), and Novelty (i.e. innovation of the product) [36]. All participants answered the System Usability Scale, whereas, for the sake of brevity, only one-half of the cohort answered the UEQ for DIVINA. The second half were asked to comment on a different learning resource that was not the focus of this paper. In addition, free text questions were provided in which participants could indicate what they already liked or disliked about the game and where they see further room for improvement.

Results

According to Brooke [33] the inverted items were recoded manually. As items could be omitted, only a total number of 127 complete datasets were assessed for the SUS and revealed a mean of $M=59.19$. Hence, the response rate for the SUS was 85%. The result of the SUS falls within the first quartile spanning from 30.0 to 62.6 and is below the average score of 68 [35].

The UEQ comes along with an excel data analysis tool, which was used for the evaluation of the data. Since missing values could be included in the appraisal, 76 datasets were used for the evaluation of the UEQ, leading to a response rate of 100%. Broken down according to the scales, the results were as follows: Attractiveness ($M=0.94$, $SD=1.29$, *95% CI* [0.65, 1.23]), Perspicuity ($M=1.12$, $SD=1.02$, *95% CI* [0.89, 1.34]), Efficiency ($M=0.05$, $SD=0.98$,

95% CI [-0.18, 0.27]), Dependability ($M=0.73$, $SD=0.86$, *95% CI* [0.54, 0.93]), Stimulation ($M=0.82$, $SD=1.32$, *95% CI* [0.52, 1.11]), Novelty ($M=1.25$, $SD=0.88$, *95% CI* [1.06, 1.45]). The excel data analysis tool also provided benchmarks that concerned different commercial products. Compared to these benchmarks the results regarding DIVINA were bad for the scales efficiency and dependability, below average for the scales attractiveness, perspicuity and stimulation, and good for the scale novelty.

Since students are supposed to be the end-users of the serious game, we invited them to provide us with additional free text feedback in order to conclude further room for improvement. Most of all, students suggested to further improve the server capacity as well as improvements of the chatbot. Besides suggestions for improvement, students also emphasized the opportunity to apply knowledge in a safe learning environment without the risk of endangering patients' lives. Furthermore, students appreciated receiving feedback that we already implemented based on a pilot iteration with a different cohort of medical students.

General discussion

In this overview of the development and evaluation process, we co-operated with a software company to introduce a serious game representing a schematic simulation of an emergency department. Besides the introduction of the serious game, a first usability and user experience testing was conducted. To support the achievement of the learning outcomes, the serious game is already equipped with some supporting features, namely feedback, AI-generated faces, and a free text chatbot.

The evaluation of the current stage revealed promising preliminary results while simultaneously highlighting areas for further improvement. The usability score can be interpreted as substantial and falls in the first quartile of the SUS ratings, which can be set between 30.0 and 62.6 according to Bangor [35]. In other words, the result at hand represents a marginally low acceptability. Since usability defines how systems can be used effectively, efficiently and satisfactorily to achieve predefined goals, it can also be understood as the operability of a system. The low usability score might be due to the user interface or inadequate server capacity, as students remarked some issues. The user interface might have been confusing for students in terms of the arrangement of elements, as they sometimes reported problems finding the needed elements. Another problem area that might have led to the low usability rating became apparent in the users' feedback texts. Students perceived the chatbot as an inadequate tool for taking a medical history, as it sometimes answered incoherently. Although the chatbot answered incoherently and students did not always felt like getting satisfactory answers, it is nevertheless an innovative feature that enhances the psychological fidelity of the serious game. It can be argued that the chatbot enhances the psychological fidelity, as the students have to ask self-formulated questions, just like in a real-life emergency department. In addition to the chatbot, the AI-generated faces enhance the psychological fidelity and the game's educational value, as it can be assumed that students remember the patients better by their faces compared to their leading symptom. However, future studies have to investigate these assumptions.

In terms of the user experience, the scales attractiveness, perspicuity, stimulation, and novelty yielded positive ratings, whereas the scales efficiency, and dependability received neutral ratings. The positive

ratings of attractiveness and novelty showed that participants liked the use of the serious game, and perceived the design as innovative. In addition, perspicuity and stimulation were positively evaluated, as the use of the serious game was easy to learn, but also exciting and motivating. Efficiency received a neutral rating, possibly due to the additional effort while solving the tasks. Another scale, which received a neutral score, was dependability, as participants might not have felt the game to be predictable and therefore did not feel like having control over the interaction with the game. While both scales were rated with a neutral rather than a negative score, they indicate potential opportunities for improvement. These usability and user experience ratings were to be expected based on the user interface and playability of the game, but should be still used to improve the game in further iterations.

We finished the first stage of the development framework proposed by Olszewski and Wolbrink [24] since we already assembled a suitable interdisciplinary team, transferred the medical concepts, produced the essential content, and mapped the learner experience. From now on, the serious game will pass a continuous loop between development and formative evaluation. Based on evidence, all relevant evaluative findings will be implemented in further development iterations, which will in turn be evaluated again. The next step in the development process is the implementation of effective game design elements based on psychological learning theories and accompanying the development with further studies. Additionally to the implementation of further game design elements, already existing elements like the chatbot as well as the general user interface will consistently be improved. At each stage, student feedback is heeded, and helpful suggestions will be incorporated into the game evidence-based. Since this is a dynamic project, we will assist the software company to further implement new disease data to meet the demands of undergraduate medical students and their teachers. It is also conceivable to realize the opportunity of asynchronous self-directed learning with this serious game. Furthermore, it is also conceivable to extend the game in the long term to postgraduate medical students, residents, or other healthcare professions working in an emergency department.

Limitations

To ensure a concise evaluation following the game session, only half of the students completed the UEQ for DIVINA. Consequently, the interpretability of UEQ data in comparison to SUS data may be constrained due to the unequal population size. It is important to note that the UEQ, designed primarily for the commercial market, compares results of the individual scales to benchmark values oriented towards such products. Additionally, the study occurred at a single time of measurement, introducing potential limitations related to situational events like server capacity issues, as already reported by students. Replicating the evaluation at several times may circumvent these issues. The study, being part of a mandatory event, implies that participation was largely driven by extrinsic factors, potentially influencing the evaluation results. Repeating the evaluation in a voluntary study setting, where participation is likely driven by intrinsic motivation, could offer different insights. However, whether participation is influenced by intrinsic or extrinsic motivational factors should be a focus of future studies.

Conclusion

Serious games provide a risk-free learning environment that is highly valuable in the context of medical education. Therefore, a serious game presenting a virtual emergency department was developed and evaluated in terms of its usability and user experience. Overall, the evaluation yielded positive results and identified potential areas for further improvement. The results of the evaluation will be integrated into the consistent development of the serious game to offer medical students a valuable learning source for education in the field of emergency care.

Acknowledgements

We would like to thank all medical students and student assistants who helped realizing the project.

Author contributions

AA conceived the evaluation, analyzed the data, and wrote the manuscript. CH contributed to the development of the serious game in terms of equipping it with medical data. CM and MF developed the software of the serious game. MCL revised the manuscript. TR conceived of the study, developed its design, conducted the evaluation, and revised the manuscript. All authors have approved the final version of the manuscript and agreed with its submission.

Funding

The development of the serious game was funded by the German Federal Ministry of Health and was part of the overarching project "Nationale Lernplattformen für digitales Patienten-bezogenes Lernen im Medizinstudium - DigiPaL" with the funding reference number ZMVI1-2520COR200. Open Access funding enabled and organized by Projekt DEAL.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

The study protocol was reviewed by the local Institutional Review Board at Goettingen medical school in winter term 2021/2022 (application number 34/8/21).

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Consent for publication

Not applicable.

Non-financial interests
None.

Competing interests
Financial interests: Authors CM, MF and TR hold shares in the company Yellowbird Consulting LTD that has developed the serious game referred to in this article. No other author has competing interests.Non-financial interests: none.

Received: 18 August 2023 / Accepted: 19 September 2024
Published online: 27 September 2024

References

1. Susi T, Johannesson M, Backlund P. Serious games: An overview. 2007.
2. Gorbanev I, Agudelo-Londono S, González RA, Cortes A, Pomares A, Delgadillo V, et al. A systematic review of serious games in medical education: quality of evidence and pedagogical strategy. Med Educ Online. 2018;23(1):1438718.
3. Maheu-Cadotte MA, Cossette S, Dube V, Fontaine G, Lavallee A, Lavoie P, et al. Efficacy of Serious games in Healthcare Professions Education: a systematic review and Meta-analysis. Simul Healthc. 2021;16(3):199–212.
4. Sobolewska P, Pinet Peralta LM. Use of the educational mobile applications by emergency medical services personnel. Crit Care Innovations. 2019;2(2):25–31.
5. Min A, Min H, Kim S. Effectiveness of serious games in nurse education: a systematic review. Nurse Educ Today. 2022;108:105178.
6. Sharifzadeh N, Kharrazi H, Nazari E, Tabesh H, Edalati Khodabandeh M, Heidari S, et al. Health Education Serious Games Targeting Health Care Providers, patients, and Public Health Users: scoping review. JMIR Serious Games. 2020;8(1):e13459.
7. Oblinger D. The next generation of educational engagement. J Interact Media Educ. 2004;2004(1).
8. Dankbaar ME, Alsma J, Jansen EE, van Merrienboer JJ, van Saase JL, Schuit SC. An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. Adv Health Sci Educ Theory Pract. 2016;21(3):505–21.
9. Dankbaar ME, Richters O, Kalkman CJ, Prins G, Ten Cate OT, van Merrienboer JJ, et al. Comparative effectiveness of a serious game and an e-module to support patient safety knowledge and awareness. BMC Med Educ. 2017;17(1):30.
10. Haoran G, Bazakidi E, Zary N. Serious games in Health professions Education: review of trends and Learning Efficacy. Yearb Med Inf. 2019;28(1):240–8.
11. Michael D, Chen S. Serious games: games that educate, train and inform. Boston, MA: Thomson Course Technology; 2006.
12. Laamarti F, Eid M, El Saddik A. An overview of Serious games. Int J Comput Games Technol. 2014;2014:1–15.
13. Deterding S, Dixon D, Khaled R, Nacke L. From Game Design Elements to Gamefulness: Defining Gamification. Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments; Tampere, Finland2011.
14. Anastasiadis T, Lampropoulos G, Siakas K. Digital Game-based Learning and Serious games in Education. Int J Adv Sci Res Eng. 2018;4(12):139–44.
15. Graafland M, Schraagen JM, Schijven MP. Systematic review of serious games for medical education and surgical skills training. Br J Surg. 2012;99(10):1322–30.
16. Ward M, Ni She E, De Brun A, Korpos C, Hamza M, Burke E, et al. The co-design, implementation and evaluation of a serious board game 'PlayDecide patient safety' to educate junior doctors about patient safety and the importance of reporting safety concerns. BMC Med Educ. 2019;19(1):232.
17. Morey JC, Simon R, Jay GD, Wears RL, Salisbury M, Dukes KA, et al. Error reduction and performance improvement in the emergency department through formal teamwork training: evaluation results of the MedTeams project. Health Serv Res. 2002;37(6):1553–81.
18. Wong JY-H, Ko J, Nam S, Kwok T, Lam S, Cheuk J, et al. Virtual ER, a serious game for interprofessional education to enhance teamwork in medical and nursing undergraduates: development and evaluation study. JMIR Serious Games. 2022;10(3):e35269.
19. Kassirer JP. Teaching clinical reasoning: case-based and coached. Acad Med. 2010;85(7):1118–24.
20. Andersson U, Maurin Soderholm H, Wireklint Sundstrom B, Andersson Hagiwara M, Andersson H. Clinical reasoning in the emergency medical services: an integrative review. Scand J Trauma Resusc Emerg Med. 2019;27(1):76.
21. Middeke A, Anders S, Schuelper M, Raupach T, Schuelper N. Training of clinical reasoning with a serious game versus small-group problem-based learning: a prospective study. PLoS ONE. 2018;13(9):e0203851.
22. Johnsen HM, Fossum M, Vivekananda-Schmidt P, Fruhling A, Slettebo A. Teaching clinical reasoning and decision-making skills to nursing students: design, development, and usability evaluation of a serious game. Int J Med Inf. 2016;94:39–48.
23. Wang R, DeMaria S Jr., Goldberg A, Katz D. A systematic review of Serious games in Training Health Care professionals. Simul Healthc. 2016;11(1):41–51.
24. Olszewski AE, Wolbrink TA. Serious gaming in Medical Education: a proposed structured Framework for Game Development. Simul Healthc. 2017;12(4):240–53.
25. Bevan N, Carter J, Earthy J, Geis T, Harker S. New ISO Standards for Usability, Usability Reports and Usability Measures. Human-Computer Interaction Theory, Design, Development and Practice. Lecture Notes in Computer Science2016. pp. 268 – 78.
26. Bevan N. What is the difference between the purpose of usability and user experience evaluation methods? Proceedings of the Workshop UXEM. 2009.
27. Maheu-Cadotte MA, Dube V, Cossette S, Lapierre A, Fontaine G, Deschenes MF, et al. Involvement of end users in the development of Serious games for Health Care Professions Education: systematic descriptive review. JMIR Serious Games. 2021;9(3):e28650.
28. Crossley C, Fanfarelli JR, McDaniel R. User experience design considerations for healthcare games and applications. 2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH); 11–13 May 2016; Orlando, FL, USA: IEEE; 2016. pp. 1–8.
29. Alexiou A, Schippers MC. Digital game elements, user experience and learning: a conceptual framework. Educ Inform Technol. 2018;23(6):2545–67.
30. Escribano BB, A, Del Blanco, Torrente J, Mate JB, Manjon BF. Educational Game Development Approach to a particular case: the donor's evaluation. Transpl Proc. 2015;47(1):13–8.
31. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach. 2005;27(1):10–28.
32. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; Long Beach, California, USA2019. pp. 4401-10.
33. Brooke J. SUS-A quick and dirty usability scale. Usability Evaluation Ind. 1996;189(194):4–7.
34. Laugwitz B, Held T, Schrepp M, editors. Construction and evaluation of a user experience questionnaire. HCI and Usability for Education and work; 2008 2008; Berlin, Heidelberg: Springer Berlin Heidelberg.
35. Bangor A. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud. 2009;4(3):114–23.
36. Schrepp M. User experience questionnaire handbook. All you need to know to apply the UEQ successfully in your project. 2015.

Publisher's note

Aster *et al. Advances in Simulation* (2025) 10:10
https://doi.org/10.1186/s41077-025-00341-7

Advances in Simulation

RESEARCH

Open Access

Theoretical background of the game design element "chatbot" in serious games for medical education

Alexandra Aster[1*], Arietta Lotz[1] and Tobias Raupach[1]

Abstract

Background The use of virtual patients enables learning medical history taking in a safe environment without endangering patients' safety. The use of a chatbot embedded in serious games provides one way to interact with virtual patients. In this sense, the chatbot can be understood as a game design element, whose implementation should be theory driven and evidence based. Since not all game design elements are already connected to theories, this study aimed to evaluate whether the game design element chatbot addresses the need for autonomy rooted in the self-determination theory.

Method A cross-sectional study was conducted to compare two distinct chat systems integrated in serious games with one system being an open chatbot and the other system being a constrained chat system. Two randomized groups of medical students at a German medical school played one of two serious games each representing an emergency ward. The data collected included both objective data in terms of students' question entries and subjective data on perceived autonomy.

Results Students using the open chatbot generally asked significantly more questions and diagnosed significantly more patient cases correctly compared to students using a constrained chat system. However, they also asked more questions not directly related to the specific patient case. Subjective autonomy did not significantly differ between both chat systems.

Conclusion The results suggest that an open chatbot encourages students' free exploration. Increased exploration aligns with the need for autonomy, as students experience freedom of choice during the activity in terms of posing their own questions. Nevertheless, the students did not necessarily interpret the opportunity to explore freely as autonomy since their subjectively experienced autonomy did not differ between both systems.

Keywords Self-determination theory, Game design elements, Chatbot, Serious game, Medical education

Background

Serious games are increasingly used in medical education [1–3], but the optimal design of these games has not been completely established yet. In particular, the link between theoretical underpinnings and game design elements is not clear yet, although being fundamental [4]. This gap means that serious games may be less effective in achieving learning goals [4]. This study aimed to examine two different types of chatbots embedded in serious games, considered through the lens of self-determination theory, to determine the impact on learning behavior and students' perceived autonomy.

*Correspondence:
Alexandra Aster
alexandra.aster@ukbonn.de
[1] Institute of Medical Education, University Hospital Bonn, Bonn, Germany

Training of history-taking skills in medical education

History taking contributes about 76% to the final medical diagnoses made by physicians [5]. In a study conducted with medical students, 43 out of 60 first-year medical students (71.7%) who diagnosed a simulated patient correctly made the correct diagnosis directly after taking the medical history [6]. Thus, it appears that teaching history taking to medical students is of particular importance. A systematic review revealed that there is a plethora of interventions used to teach history taking [7]. The types of intervention ranged from instructional approaches (i.e., focus scripts, video tape, or online courses) to more sophisticated approaches (i.e., small-group workshops with role-play, simulated patients, real patients, or virtual patients) [7]. While different interventions are applied for teaching history taking, simulated patients (SPs) are still used most frequently [7, 8]. SPs provide a risk-free learning environment for students to improve their communication skills [8]. Since training SPs requires resources and they are also a limited resource themselves [9], another efficient method of standardized training has emerged with virtual patients (VP) [10, 11]. VPs are a secure, reliable, and valid learning resource offering the opportunity of repeated exposure to the same potentially complex scenarios difficult to replicate in real life [11]. The authenticity of VPs depends on three aspects: the learner's perception of the story surrounding the VP, the format, and the quality in which the VP is presented [12]. It is already shown that VPs can be developed emotionally responsive [13], which is relevant for training history taking. One way to access VPs during history taking is through chatbots [13, 14], which appear to be a learning resource largely equally effective as controls [14, 15]. A chatbot can be best defined as a computer program imitating human conversation when addressed through written or spoken language [16].

Theoretical underpinnings of serious games as training environments

Serious games, defined as games whose primary aim is reaching a learning goal and not solely inducing fun or enjoyment [17], offer a learning environment in which VPs can be usefully integrated. An abundance of serious games already found entrance into health professions education and seem to be as effective as or more effective than not only control conditions such as traditional or digital learning formats but also other types of serious games or gamification [1]. Besides the effectiveness in terms of improved learning outcomes, serious games can also enhance the motivation and engagement of the players [2, 3] and should therefore be designed according to motivational theories. One such motivational theory frequently used in the design of serious games is the *self-determination theory* (SDT) [18]. According to the SDT, the three basic psychological needs for autonomy, competence, and relatedness have to be addressed to lead to intrinsic motivation [18]. The need for autonomy relates to the feeling of acting volitional according to one's own will with perceived decision freedom enabling the choice between different kinds of action [19, 20]. The need for competence is addressed by the feeling of being capable to meet a goal based on the effective execution of one's own behavior [20], while the need for relatedness is addressed by a sense of belonging to a reference group [19]. In serious games, these needs are addressed by inherent game design elements, which are essential for games to be characterized as such [21]. Existing literature already examined the importance for game design elements to be based on theoretical underpinnings [4]. Some game design elements were already linked with the addressed need, for instance points or badges refer to the need for competence, avatars or meaningful stories refer to the need for autonomy, and teammates refer to the need for relatedness [19]. However, a considerable number of game design elements have not yet been matched to the SDT.

Chatbots as a game design element

In this sense, a chatbot embedded in a serious game can also be understood as a game design element with unclear theoretical background. It is long known that an autonomy fostering learning environment during medical education not only enhances students' autonomous motivation but also positively influences their perseverance as well as their interaction with patients [22]. Previous research has shown that an addressed need for autonomy is aligned with greater experienced curiosity in terms of exploration [23]. It can be assumed that providing users with the opportunity to freely select or enter their queries may address their subjective feelings of autonomy, as reflected in exploratory behavior. Following this line of thought, the game design element "chatbot" might be assigned to the need for autonomy from the SDT, especially under the included aspect of decision freedom [19]. Since internal game analytics should be evaluated to not compromise the players' flow during the game experience [24], it is reasonable to use the chatbot entries as an operationalization for assessing exploration and therefore autonomy. For the purpose of this study, a chatbot in which questions can be formulated freely via free entries is referred to as open. Vice versa, a chatbot in which questions can be selected from a set of predefined questions is referred to as constrained.

Research aim

This research's overarching aim was to assess whether the need for autonomy stemming from the SDT can be linked to the serious games' game design element "chatbot" and whether this association depends on the type of chatbot used. Therefore, two serious games presenting different history-taking systems with different degrees of freedom were compared. The need for autonomy was operationalized through medical students' free exploration during history taking. It is assumed that an open chatbot that mimics a real-world situation by requiring self-formulated questions addresses students' autonomy due to offering a free environment with the opportunity of decision freedom expressed through free exploration.

- H1: Students ask significantly more questions in an open chatbot compared to a constrained chat system.
- H2: Students ask significantly more irrelevant questions in an open chatbot compared to a constrained chat system.
- H3: Students report significantly more subjective feelings of autonomy in an open chatbot compared to a constrained chat system.

Methods

The local Institutional Review Board at Göttingen Medical School approved this study in winter term 2023/2024 (application number: 8/9/23). All participants gave written informed consent beforehand.

Study procedure

The study was conducted in a mandatory module for fourth-year undergraduate medical students covering the areas cardiology and pneumology at Göttingen medical school in winter term 2023/2024. All students attending the module were invited to voluntarily participate in the study, but participation was not mandatory. The module comprised four sessions each lasting 90 minutes. However, only the data collected during the first session were relevant for this study, as it was the first time students interacted with the serious games, ensuring that the data were not biased by familiarity with the game. Students were randomly assigned to one of two study groups. One group engaged in on-site gameplay of the serious game EMERGE [25], representing the constrained chat system, while the other group simultaneously played the serious game DIVINA [26] online, representing the open chatbot. Both serious games provided the students with the diseases ST-segment elevation myocardial infarction (STEMI), non-ST-segment elevation myocardial infarction (NSTEMI), musculoskeletal chest pain, and hypertensive crisis, while DIVINA additionally provided the disease congestive heart failure. At the end of the first session, students were invited to participate in an evaluation.

Serious game environments

Both serious games represent emergency departments with similar procedures within the games, although differing in their visual design as well as in their game structure. In both games, players take a patients' medical history, order investigations, initiate treatments, and finally discharge the patient. For the present study, the focus is only on the manner how the medical history taking takes place. In the serious game EMERGE, players use the constrained chat system by choosing from a long menu of 70 predefined questions. Precisely, students enter specific letters or words included in their sought question to which the long menu proposes suitable questions including the entered letters or words. Please refer to Middeke, Anders [25] for further information on the design of EMERGE. Contrary to EMERGE, the serious game DIVINA does not provide predefined questions for medical history taking, but students have to phrase questions themselves in an open chatbot. The chatbot refers to a script-based system and provides answers based on a system that draws on information about the specific virtual patient and their symptoms. Please refer to Aster, Hütt [26] for further information on the design of DIVINA. In both serious games, students are not limited in the number of questions for taking a sufficient medical history.

Data collection and preparation

History data

All qualitative history data gathered in both games were quantified first. To do so, a checklist was developed in collaboration between a physician specialized in the field of cardiology and a psychologist. The physician contributed medical expertise and ensured content accuracy, while the psychologist focused on assessing psychometric properties of the checklist. These two authors used the checklist to independently and blindly score all history-taking data for both serious games. For the sake of uniformity, the same checklist was used for all diseases. The data were quantified in the way that all questions were scored irrespectively of the received answer, and each question was rated once regardless of reformulations. More precisely, it was irrelevant whether students received a sufficient and satisfactory answer; the questions were evaluated independently of the received answers. Depending on the medical relevance, questions were scored with 1 or 2 points. Overall, a total of 49 points could be achieved. The checklist oriented towards the SAMPLER/OPQRST scheme [27] and contained the following areas: "basic patient-related data", "current

reason for consultation", "specific somatic anamnesis" (subdivided in "current complaints and development" and "focused pain anamnesis"), "general somatic anamnesis" (containing "past medical history", "vegetative anamnesis", "risk factors", in particular "cardiovascular risk factors"), as well as "family and social anamnesis", and "orienting psychiatric anamnesis". The complete checklist can be found in Supplementary 1.

Questionnaires

The evaluation consisted of the subscale for "perceived choice" from the *Intrinsic Motivation Inventory* (IMI) [28] and the *General Self-Efficacy Short Scale* (German: Allgemeine Selbstwirksamkeit Kurzskala, ASKU) [29]. Both questionnaires are reliable and validated measuring instruments [28, 29]. The IMI was chosen to measure the intrinsic motivation of students, while the ASKU was chosen to measure self-efficacy, which can be considered as related to the SDT. According to Bandura [30], self-efficacy implies that a person has the belief to successfully master a situation by performing the necessary behavior. Moreover, self-efficacy has already been found to be an underlying construct for gamification [31].

Data analysis

History data

The history data were analyzed according to the following procedure: In a first step, the interrater reliability for both authors was assessed. All analyses were conducted utilizing a mean rating score derived from the assessments provided by both reviewers for each serious game, hereinafter referred to as "history score". Since normal distribution of the data was not given, nonparametric statistical methods or methods that are not affected by violation of this assumption were chosen. For each statistical procedure, the corresponding effect size was conducted and reported.

Prior to hypothesis 1, descriptive statistics about the serious games were evaluated, and a chi-squared test was conducted to assess differences in the number of correctly diagnosed cases between the serious games.

For hypothesis 1, which states that students asked significantly more history questions in an open chatbot (i.e., DIVINA) than in a constrained chat system (i.e., EMERGE), a Mann–Whitney *U*-test comparing the absolute number of questions between the two groups was conducted.

A hierarchical sequence of steps was followed to evaluate hypothesis 2, which states that students significantly asked more irrelevant questions in an open chatbot compared to a constrained chat system. Firstly, a Mann–Whitney *U*-test for comparing the achieved history scores between the two serious games was performed. Following this, a regression analysis to examine the relation between the number of questions asked and the achieved history scores for each serious game was performed. The irrelevance was defined as a ratio of the achieved history score and the number of questions asked for each chat protocol separately. A ratio < 1 represents that more questions were asked than points were achieved implying the presence of more irrelevant questions. Vice versa, a ratio > 1 implies less irrelevant questions since a higher history score was achieved with less questions asked, in a sense that the history score exceeds the number of questions. For examining the hypothesis, a final Mann–Whitney *U*-test comparing the ratios between the two groups was conducted.

Questionnaires

The questionnaire data were analyzed for answering hypothesis 3 stating that students reported significantly higher subjective autonomy feelings after playing DIVINA compared to EMERGE. Both questionnaires were analyzed according to its guidelines before a mean value comparison was conducted. Since these data were not normally distributed, Mann–Whitney *U*-tests were carried out.

Results

History data

$N = 154$ fourth-year medical students consented to have their data entries in the serious games analyzed. Since all data were recorded anonymously, no further statements and conclusions about the population could be made except them being fourth-year students at a German medical school. Interrater reliability was computed for both serious games using the intraclass correlation coefficient (ICC), resulting in an ICC of 0.890 for DIVINA and an ICC of 0.939 for EMERGE. According to Cicchetti [32], both coefficients can be interpreted as very good agreements.

Only chat protocols containing at least one question were deemed valid. This led to 249 valid chat protocols stemming from DIVINA (4 of 254 initial chat protocols had to be excluded) and 456 valid chat protocols stemming from EMERGE (62 of 518 initial chat protocols had to be excluded). Students correctly diagnosed 162 patient cases (65%) in DIVINA and 236 patient cases (52%) in EMERGE ($\chi^2(1) = 13.025$, $p < 001$). Generally, the number of questions asked per chat protocol in DIVINA ranged from 3 to 57 ($Mdn = 13$) and in EMERGE from 1 to 40 ($Mdn = 9$). Students asked significantly more questions in DIVINA than in EMERGE, $U = 37{,}980.000$, $p < 0.001$, $r = 0.27$, although with a weak effect size [33].

For evaluating whether students asked a higher number of irrelevant questions in an open chatbot, several

analyses were conducted. In a first step, it was found that the achieved history scores did not differ significantly between DIVINA ($Mdn = 14.5$) and EMERGE ($Mdn = 14$), $U = 51{,}766.00$, $p = 0.053$. In a next step, it was examined whether the number of questions asked was related to the achieved history score. A polynomial regression was conducted for each serious game, since the assumption of linearity required for performing a linear regression was not met. The models were significant for both serious games, DIVINA ($F(2248) = 307.44$, $p < 0.001$) and EMERGE ($F(2455) = 1508.84$, $p < 0.001$). All specific parameters for both serious games can be found in Table 1. The scatterplot of the polynomial regressions for the relation between the number of questions asked and the achieved history score can be found in Fig. 1.

The subsequent Mann–Whitney *U*-test using the ratio showed that significantly more points were achieved by asking less questions in EMERGE ($Mdn = 1.5$) than in DIVINA ($Mdn = 1.13$), $U = 28{,}367.000$, $p < 0.001$, $r = 0.41$ with a moderate effect size [33]. The indicated medians refer to the abovementioned ratio of which the distribution of frequencies can be found in Fig. 2. Supporting the hypothesis, the lower ratio indicates a tendency to ask more irrelevant or not expedient questions in DIVINA.

Subjective autonomy measures

Overall, $N = 81$ ($n = 44$ DIVINA, $n = 37$ EMERGE) students participated in the questionnaire of which $n = 41$ data sets could be analyzed for DIVINA and $n = 35$ for EMERGE. Considering the subjectively experienced autonomy during the play of both games, the autonomy scale of the IMI showed no significant differences between the experienced autonomy in DIVINA ($Mdn = 4.29$) and EMERGE ($Mdn = 4.43$), $U = 654.00$, $p = 0.507$. An explorative analysis addressed the relationships between the autonomy scale and the ASKU, since the SDT and self-efficacy are already jointly used constructs in the design of serious games [34]. Across the two serious games as well as broken down for each serious game, no significant correlation was found. Moreover, no significant group difference between EMERGE and DIVINA was found for the ASKU, $U = 627.00$, $p = 0.266$.

Discussion

General discussion

This study aimed to determine whether the need for autonomy stemming from the self-determination theory can be understood as a theoretical basis for the game design element "chatbot". For the purpose of this study, autonomy was operationalized by students' free

Table 1 Overview of the parameters of the polynomial regression

		R^2	Adjusted R^2	β	*SE*	*t*	*p*
DIVINA		0.714	0.712				
	Linear term			1.008	0.068	14.72	< .001
	Squared linear term			−0.011	0.002	−7.07	< .001
EMERGE		0.869	0.869				
	Linear term			1.593	0.048	32.91	< .001
	Squared linear term			−0.025	0.002	15.63	< .001

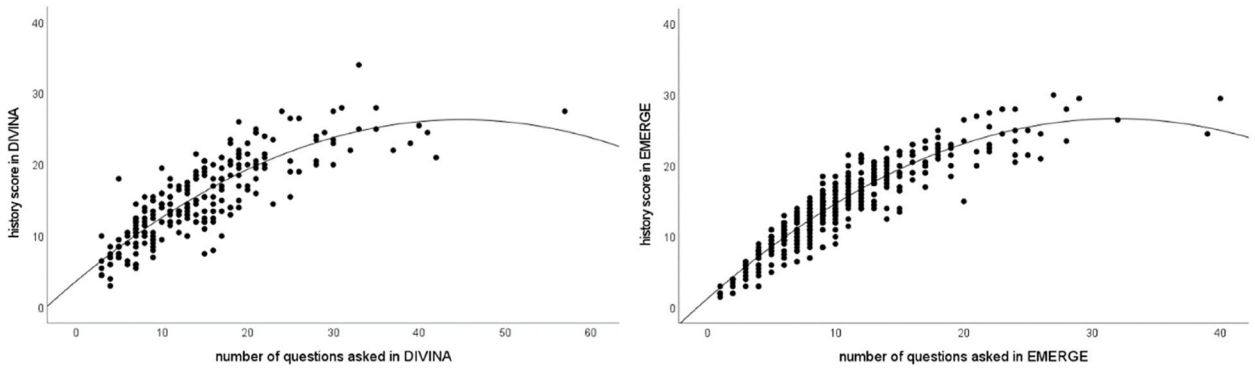


Fig. 1 Scatterplot of the polynomial regression for both serious games

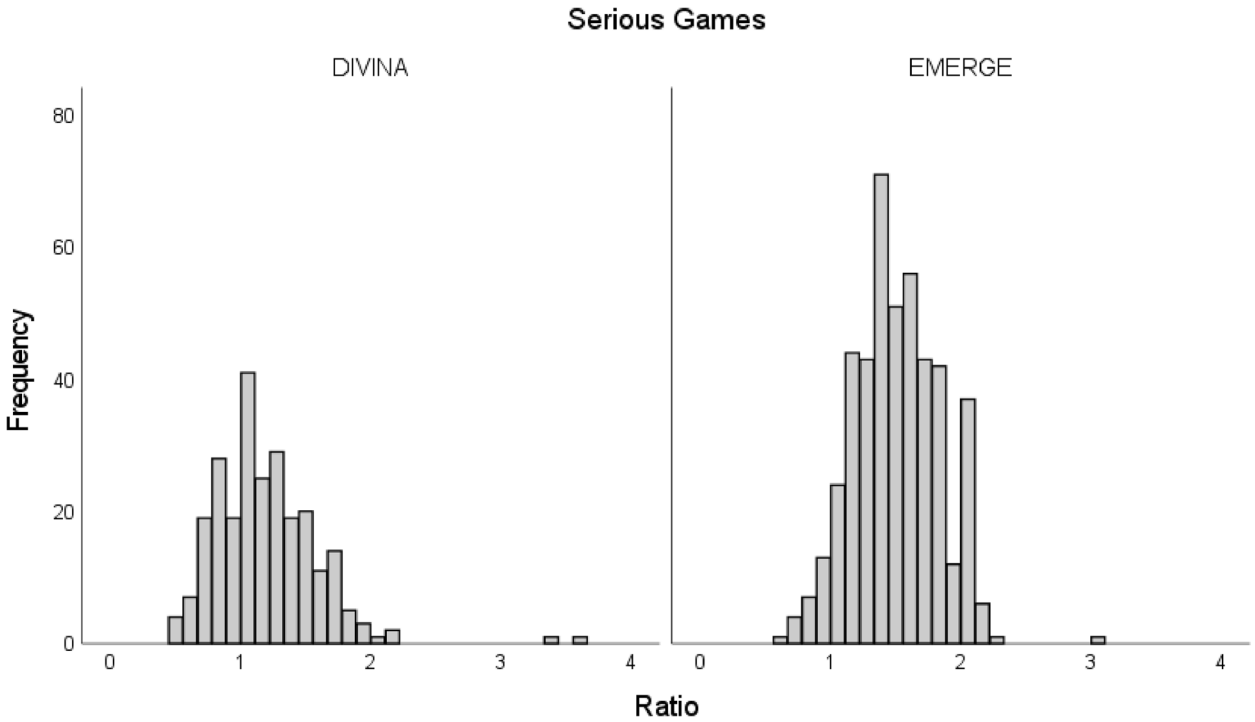


Fig. 2 Distribution of the frequencies of the ratios

exploration during history taking as well as by subjective autonomy ratings.

Overall, students showed better objective results (i.e., correctly diagnosed patient cases) in DIVINA (open chatbot) than in EMERGE (constrained chat system) and gained a slightly higher history score without there being a significant difference. The serious games were similar in content and action possibilities; however, they differed in their specific appearance. Since our analyses focused solely on interactions within the chatbot rather than the entire serious game, we assume that these differences in appearance did not affect the results. The first hypothesis that more history-taking questions were asked in an open chatbot compared to a constrained chat system showed a significant difference albeit with a weak effect size. Thus, it can be assumed that students explored more during history taking when provided with the opportunity to ask self-developed questions. The results support the assumption that the opportunity to formulate questions on their own fosters students' exploration as operationalized by the amount of questions. Nevertheless, the amount of questions asked was relatively small for both serious games. One possible explanation could be the internal setting of the serious games (i.e., an emergency department), which may prompted students to prioritize further investigations over asking additional history taking questions.

Based on the idea of hypothesis 1, the next logical step was to examine whether students did not only ask more questions but also asked more irrelevant questions in an open chatbot compared to a constrained chat system. Therefore, hypothesis 2 was driven by the assumption that more irrelevant questions were asked as a result of increased exploration. In this context, irrelevant questions are queries, statements, or nonconstructive inputs that are not directly focusing on the central aspects of the patient case in terms of furthering the medical treatment. Nevertheless, the questions do not necessarily have to be irrelevant in the medical sense. In line with the hypothesis, the analysis revealed a significant difference with a moderate effect, indicating that students tend to ask more irrelevant questions in an open chatbot. Since the open chatbot was sometimes unable to usefully reply to the initial question, students tried to handle it by reformulating their entry. This increased the number of questions but did not affect the score, as these questions were only scored in their initial version. It is already known that script- or rule-based chatbots show difficulties in understanding the input, demonstrated by a virtual patient mismatching approximately 40% of students' entries with the appropriate response [11]. An upcoming area of interest is the attempt to use large language models (LLMs) for the simulation of virtual patients [35, 36]. Further studies could consider using LLMs and thereby

assessing students' perceived autonomy using sound research designs.

An explanation for the difference in the amount of irrelevant questions derives from the manner of how questions were asked. While students needed to formulate their own questions in accordance with the rule-based open chatbot, the constrained chat system presented all possible requested questions, from which students only needed to select. Moreover, the generally limited number of available questions could have led to less irrelevant questions in the constrained chat system. At the same time, in this scenario, opportunities for students to pursue their own line of inquiry were very limited. The moderate effect size suggests that students nevertheless did not chose the perfect amount of questions in the constrained chat system, although the long menu format already disclosed potential questions. Although the use of in-game analytics is a recommended approach in serious game research [24], it is worth noting that students' actions are difficult to interpret without considering their intent. Future research should aim to capture students' intent and merge these insights. By means of the ratio, results showed that neither in the open chatbot nor in the constrained chat system one question led to one point, which may have been also caused by the amount of irrelevant questions or entries. Generally, an explanation for the relatively low history scores might be that students are possibly not familiar enough with history taking. Further studies should address this idea by adding an intervention to the study design. Furthermore, it would be intriguing to calculate the number of questions required to reach a diagnosis and examine its accuracy. Doing so, it could be tested whether the statement that up to three-quarters of the diagnoses are already correct after taking a history also applies for history taking with VPs [5].

Besides the objective data, the subjective data gave important insights on the experienced autonomy during each history taking. The subjectively experienced autonomy did not significantly differ between both serious games. Together with the results of hypothesis 2, it can be concluded that although students did not subjectively feel more autonomous in an open chatbot than in a constrained chat system, they still asked more questions and subsequently got more diagnoses correctly in the open chatbot. It is assumable that the discussed limitations associated with an open albeit script-based chatbot may have negatively influenced students' feelings of autonomy. Consequently, students felt rather forced than autonomous during the interaction with the script-based chatbot given the necessary reformulation of their questions. These assumptions are based on questionnaire data, and although questionnaires are a frequently used measuring instrument for assessing autonomy, this particular questionnaire might not have been a sufficient instrument for the present study. Future research should consider alternative approaches, such as focus groups, which may yield different insights. However, a meta-analysis on gamification found that in most of the included studies, taking part in gamified classes enhanced students' perception of autonomy [37]. Nonetheless, in line with our results, the authors found studies where gamification did not lead to enhanced perceptions of autonomy [37].

Limitations

The limitations primarily concern the generalizability of the results due to the used game environments as well as the used data analysis instrument. Both serious games simulate emergency departments, raising the question whether this setting with its time pressure is adequate for studying students' history taking. Moreover, some among the studied diseases might have required more, and some perhaps needed less history taking due to the risk of serious deterioration or even life-threatening complications. It has to be considered whether other settings, such as a general practitioner's practice, an outpatient clinic, or a normal ward, are more suitable for examining students' history taking. Future studies could possibly examine these different settings and contextualize the medical history within the framework of other conducted investigations to clarify the role of history taking in order to provide better generalizability.

The predefined checklist used to rate the history data constitutes another limitation. The checklist was oriented towards the SAMPLER/OPQRST scheme [27] that is commonly used in emergency management and includes a section specifically related to pain. Not all included diseases manifested with pain; however, due to the structure of the serious games' outputs, it was not possible to control for whether the virtual patient presented with pain. As a result, the entire checklist was applied across all patient cases to provide comparability and to do justice to students specifically asking for pain. Second, the checklist was used for all diseases without being specialized for some diseases. While this procedure enhanced the simplicity of the data preparation, it may have also led to biased history scores. Future research should use disease-specific checklists tailored to the presented symptoms and count redundant questions.

Due to the different amount of subjective and objective data, drawing any conclusions on possible correlations between them was not possible. Moreover, due to the lack of identifying data, it was not possible to match questionnaire answers with the respective objective data.

Conclusion

Our research focused on the theoretical underpinning of the game design element "chatbot". Two chatbot systems were compared to determine whether the need for autonomy stemming from the self-determination theory is addressed when using a chatbot. We observed more exploratory behavior favoring autonomy in student history taking with an open chatbot, but our measures of subjective student experience did not reflect that. Even though measuring instruments require reconsideration to confirm this assumption, our study yields initial proof that an open chatbot may address the need for autonomy as operationalized by students' exploration behavior. In conclusion, open chatbots can be considered valuable tools for medical students to practice history taking. However, further research is needed to identify the specific characteristics of chatbots that contribute to fostering autonomy during their use.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41077-025-00341-7.

Supplementary Material 1. Checklist.

Acknowledgements
We would like to thank Matthias Carl Laupichler and Johanna Flora Rother for their valuable support, feedback, and help during the research.

Authors' contributions
A.A. conceptualized the methodology, conducted the investigation and formal analysis, analyzed the final data, and wrote the original draft of the manuscript. A.L. conducted the formal analysis, and reviewed and edited the final draft of the manuscript. T.R. conducted the investigation, reviewed and edited the final draft of the manuscript, and supervised the study. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The local Institutional Review Board at Göttingen Medical School approved this study in winter term 2023/2024 (application number: 8/9/23). All participants gave written informed consent beforehand.

Consent for publication
Not applicable.

Competing interests
The author TR declares a financial conflict of interest as he holds shares in the company Yellowbird Consulting LTD that has developed the serious game DIVINA referred to in this article. No other author has competing interests.

Received: 13 November 2024 Accepted: 28 February 2025
Published online: 12 March 2025

References

1. Gentry S, Gauthier A, L'Estrade Ehrstrom B, Wortley D, Lilienthal A, Tudor Car L, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res. 2019;21(3):e12994.
2. Dankbaar MEW, Roozeboom MB, Oprins EAPB, Rutten F, van Merrienboer JJG, van Saase JLCM, Schuit SCE. Preparing residents effectively in emergency skills training with a serious game. Simulation in Healthcare. 2017;12(1):9–16.
3. Zairi I, Ben Dhiab M, Mzoughi K, Ben MI. The effect of serious games on medical students' motivation, flow and learning. Simul Gaming. 2022;53(6):581–601.
4. Aster A, Laupichler MC, Zimmer S, Raupach T. Game design elements of serious games in the education of medical and healthcare professions: a mixed-methods systematic review of underlying theories and teaching effectiveness. Advances in Health Sciences Education. 2024.
5. Peterson MC, Holbrook JH, Von Hales DE, Smith NL, Staker LV. Contributions of the history, physical examination, and laoratory investigation in making medical diagnoses. West J Med. 1992;156(2):163.
6. Tsukamoto T, Ohira Y, Noda K, Takada T, Ikusaka M. The contribution of the medical history for the diagnosis of simulated cases by medical students. Int J Med Educ. 2012;3:78–82.
7. Keifenheim KE, Teufel M, Ip J, Speiser N, Leehr EJ, Zipfel S, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ. 2015;15:159.
8. Kaplonyi J, Bowles KA, Nestel D, Kiegaldie D, Maloney S, Haines T, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ. 2017;51(12):1209–19.
9. Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach. 2009;31(6):477–86.
10. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ. 2020;54(9):786–95.
11. Stevens A, Hernandez J, Johnsen K, Dickerson R, Raij A, Harrison C, et al. The use of virtual patients to teach medical students history taking and communication skills. Am J Surg. 2006;191(6):806–11.
12. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. Acad Med. 2010;85(10):1589–602.
13. Xu J, Yang L, Guo M. Designing and evaluating an emotionally responsive virtual patient simulation. Simul Healthc. 2024;19(3):196–203.
14. Lippitsch A, Steglich J, Ludwig C, Kellner J, Hempel L, Stoevesandt D, et al. Development and evaluation of a software system for medical students to teach and practice anamnestic interviews with virtual patient avatars. Comput Methods Programs Biomed. 2024;244:107964.
15. Frangoudes F, Hadjiaros M, Schiza EC, Matsangidou M, Tsivitanidou O, Neokleous K. An overview of the use of chatbots in medical and healthcare education. International Conference on Human-Computer Interaction, vol. 12785. Cham: Springer International Publishing; 2021. pp. 170–84.
16. Adamopoulou E, Moussiades L, editors. An overview of chatbot technology. IFIP international conference on artificial intelligence applications and innovations. Cham: Springer; 2020.
17. Michael DR, Chen SL. Serious games: games that educate, train, and inform. Boston: Muska & Lipman/Premier-Trade; 2005.
18. Deci EL, Ryan RM. The general causality orientations scale: self-determination in personality. J Res Pers. 1985;19(2):109–34.
19. Sailer M, Hense JU, Mayr SK, Mandl H. How gamification motivates: an experimental study of the effects of specific game design elements on psychological need satisfaction. Comput Hum Behav. 2017;69:371–80.
20. Niemiec CP, Ryan RM. Autonomy, competence, and relatedness in the classroom. Theory Res Educ. 2009;7(2):133–44.
21. Deterding S, Dixon D, Khaled R, Nacke L. From game design elements to gamefulness: defining" gamification". In Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments. 2011. pp. 9-15.

22. Williams GC, Saizow RB, Ryan RM. The importance for self-determination theory for medical education. Acad Med. 1999;74(9):992–5.
23. Schutte NS, Malouff JM. Increasing curiosity through autonomy of choice. Motiv Emot. 2019;43(4):563–70.
24. Qian M, Clark KR. Game-based learning and 21st century skills: a review of recent research. Comput Hum Behav. 2016;63:50–8.
25. Middeke A, Anders S, Raupach T, Schuelper N. Transfer of clinical reasoning trained with a serious game to comparable clinical problems: a prospective randomized study. Simul Healthc. 2020;15(2):75–81.
26. Aster A, Hütt C, Morton C, Flitton M, Laupichler MC, Raupach T. Development and evaluation of an emergency department serious game for undergraduate medical students. BMC Med Educ. 2024;24(1):1061.
27. Kegel M, Klee O, Herrmann T, Dietz-Wittstock M. Beobachtung und beurteilung von patienten in der notaufnahme. In: Dietz-Wittstock M, Kegel M, Glien P, Pin M, editors. Notfallpflege - Fachweiterbildung und Praxis. Berlin, Heidelberg: Springer; 2022.
28. McAuley E, Duncan T, Tammen VV. Psychometric properties of the Intrinsic Motivation Inventory in a competitve sport setting: a confirmatory factor analysis. Research quaterly for exercise and sport. 1989;60(1):48–58.
29. Beierlein C, Kemper CJ, Kovaleva A, Rammstedt B. Kurzskala zur erfassung allgemeiner Selbstwirksamkeitserwartungen (ASKU). Methoden, Daten, Anal. 2013;7(2):251–78.
30. Bandura A. Self-efficacy mechanism in human agency. Am Psychol. 1982;37(2): 122.
31. Krath J, Schürmann L, von Korflesch HFO. Revealing the theoretical basis of gamification: a systematic review and analysis of theory in research on gamification, serious games and game-based learning. Comput Human Behavior. 2021;125:106963.
32. Cicchetti D. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284.
33. Cohen J. Statistical power analysis. Curr Dir Psychol Sci. 1992;1(3):98–101.
34. Jamshidifarsani H, Tamayo-Serrano P, Garbaya S, Lim T, Blazevic P. Integrating self-determination and self-efficacy in game design. In International Conference on Games and Learning Alliance . Cham: Springer International Publishing; 2018. p. 178-190.
35. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023;9:e48291.
36. Potter L, Jefferies C. Enhancing communication and clinical reasoning in medical education: Building virtual patients with generative AI. Future Healthcare Journal. 2024;11:100043.
37. Li L, Hew KF, Du J. Gamification enhances student intrinsic motivation, perceptions of autonomy and relatedness, but minimal impact on competency: a meta-analysis and systematic review. Educ Technol Res Develop. 2024;72(2):765–96.

Publisher's Note

MEDICAL EDUCATION ONLINE
2025, VOL. 30, 2527175
https://doi.org/10.1080/10872981.2025.2527175

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Impact of providing a customized guideline on virtual medical history taking in two serious games for medical education

Alexandra Aster, Arietta Lotz, Matthias Carl Laupichler and Tobias Raupach

Institute of Medical Education, University Hospital Bonn, Bonn, Germany

ABSTRACT

Background: Serious games are known as safe learning environments, allowing medical students to train their skills without endangering patients' safety. By integrating virtual patients via chatbots, serious games provide the opportunity to practice history taking. The study investigated the impact of self-directed learning by means of a customized guideline on history taking in two distinct chatbot systems embedded in serious games. Methods: Fourth-year medical students ($N = 159$) were randomized to one of two serious games, each representing an emergency department and simulating different clinical scenarios. Students played the serious games at two measurement points and received a guideline between both sessions. The chatbots differed in the manner of query entry, with one requiring students to formulate history taking questions themselves, while the other provided a long menu of selectable questions. The dependent variables analyzed included the history taking data entered into the chatbots, represented as a quantified history score, as well as students' comparative self-assessments of their learning outcomes. Results: Comparing only the first measurement point, students achieved higher scores in the free-entry chatbot (85.2 ± 27.7) compared to the long menu chatbot (78.8 ± 35.7). Students achieved significantly higher scores in the second than in the first session in the long menu chatbot ($t(315) = -2.918$, $p = .004$, $d = -0.229$) but not in the free-entry chatbot after receiving the guideline. In terms of students' self-assessment, no significant difference between both serious games was found. Discussion: The results suggest that history taking benefits from self-directed learning in a long menu format relying on cued recall but not in a free-entry chatbot relying on free recall. Since serious games are partially artificial learning environments for training history taking, future studies should examine the extent to which students can transfer their learning in and out of serious games.

ARTICLE HISTORY
Received 27 November 2024
Revised 3 June 2025
Accepted 25 June 2025

KEYWORDS
Chatbot; serious game; medical education; history taking; self-directed learning

Introduction

Medical history taking and clinical reasoning are intertwined [1], reflected in the fact that history taking contributes to a majority of correct differential diagnoses given directly afterwards [2]. The vast majority of history taking deals with gathering information [3] and building the informational basis for arriving at a differential diagnosis (e.g., gathering information about a patient's alcohol intake [4]). A comprehensive history taken by the physician and sufficient information shared by the patient are essential for a sufficiently accurate initial diagnosis, which should be as precise as possible to avoid diagnostic errors [5]. Especially in the context of emergency medicine, gathering information precisely and comprehensively through history taking is essential [6]. Since taking a medical history plays an important role, it must be given special attention in the training of medical students.

Literature review

One training method for history taking is the use of simulated patients (SPs) [7]. SPs offer a safe, nevertheless resource intensive setting for learning history taking [8,9], therefore virtual patients (VPs) can be used as a more cost-effective alternative. The use of VPs is a proven method by which medical students train their history taking in a safe and consistent environment [10]. VPs can exemplarily be used embedded in serious games expanding the VP with an accompanying storyline and more context (e.g., conducting further examinations, initiate treatments). Serious games are already being used for teaching the contents of history taking [11]. A way to realistically train history taking with VPs is through the use of a chatbot, that is generally definable as a software designed to allow written or spoken interactions imitating a human conversation [12].

To date, few studies have used chatbots for teaching medical history taking either embedded in a serious

CONTACT Alexandra Aster alexandra.aster@ukbonn.de Institute of Medical Education, University Hospital Bonn, Venusberg-Campus 1, Bonn, 53127, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10872981.2025.2527175.

game [13] or as stand-alone chatbots [14–16]. For instance, training with a stand-alone chatbot, led to improved OSCE scores as well as to improved subjective competence and confidence ratings [17]. In this sense, a chatbot serves as a serious games' inherent game design element and should be selected theory-based to support not only students' intrinsic motivation but also to enable a sound evaluation [18]. For serious games in general, the *Self-Determination Theory* (SDT) proposed by Deci and Ryan [19] as well as the *Flow-Theory* proposed by Csikszentmihalyi [20] are frequently referenced as underlying theories [21]. Serious games allow repeated practice in a safe environment independent of resource or time constraints [22] and are already designed and used as learning environments enabling self-directed learning [23]. Self-directed learning is a process of self-planning and executing one's own learning process [24,25]. It is frequently regarded as an inherent feature of a learning environment that promotes students' freedom in the learning process [25]. Freedom is closely correlated with autonomy [26] which is why it can be assumed that self-directed learning addresses students' feelings of autonomy and therefore enhances their intrinsic motivation [27,28]. One further advantage of serious games is their external validity, which is ensured through a realistic design and enables transfer to real-life situations [29]. A previous study on a serious game for trauma triage decision-making demonstrated that physicians' behavior in the serious game mirrored real-life actions [30]. Irrespective of serious games, studies showed that the transfer of communication skills from theory to practice is challenging [31,32]. It is conceivable that serious games may facilitate the transfer of history taking skills due to their external validity.

Research aim & hypotheses

Combining the aspects of the importance of training history taking and the opportunity to train it in a safe environment enabling self-directed learning, the question occurs whether specific material for self-directed learning improves history taking between two game play sessions and whether this depends on the type of chatbot used. Therefore, the study compared the impact of a history taking guideline (i.e., specific material for self-directed learning) on the medical history taking in two distinct serious games containing a chatbot (i.e., chat system using a long menu format vs. chatbot with entry of free text questions).

In one serious game, EMERGE, students chose questions from a long menu format, while in the other serious game, DIVINA, students had to pose and formulate their own questions. The long menu provided students with all questions containing the word students entered, making it easier to recognize further relevant questions. Both chatbots can be associated with the two constructs, cued recall and free recall, stemming from the area of cognitive psychology. Cued recall denotes memory retrieval aided by internal or external cues, whereas free recall occurs without any cues [33]. In this sense, the long menu format refers to cued recall and the free-entry chatbot refers to free recall. Since it is widely acknowledged that cued recall outperforms free recall [33], it was hypothesized that students achieve more points, in terms of a higher history score, in session 1 in EMERGE than in both DIVINA sessions.

H1: Students achieve a significantly higher history score in EMERGE session 1 than in DIVINA session 1 and session 2.

It is conceivable that a guideline providing additional material for self-directed learning might reverse the effect by affecting the history taking in a free-entry chatbot but not in chat system using a long menu. The guideline not only reactivated previous knowledge but also allowed students to establish an internal scheme of history taking that can be retrieved during session two. Establishing a cognitive representation can be understood as an element of clinical reasoning [34] and since clinical reasoning and history taking are intertwined [1], it can be assumed that creating an internal scheme also applies for history taking. DIVINA allows free application of this internal scheme due to the entry format. Students can enter self-formulated questions without any restrictions, giving them the opportunity to explore thoroughly during their game play and achieve a higher history score. In contrast, a long menu, limits the amount of questions that can be asked. Previous research has already shown that students show a greater exploration behavior and ask a higher number of questions in an open chatbot compared to a constrained chat system [35]. Combining the exploration behavior with the internal scheme, it is hypothesized that studying a guideline leads to an improvement in the history taking score in DIVINA but not in EMERGE.

H2: After the individual studying of a history taking guideline, history taking will result in higher history scores in DIVINA but not in EMERGE over the course of the two sessions.

Materials and methods

Participants

The study took part during a mandatory cardiopulmonary module at Göttingen medical school in summer

term 2024 and was approved by the local Institutional Review Board beforehand (application number: 21/3/24). Although the study was conducted during a mandatory module, student consent to participate in the study (i.e., have their data analyzed) was voluntarily. Overall, a total of $N = 159$ ($n = 79$EMERGE, $n = 80$ DIVINA) fourth-year students gave their informed consent. Two semesters before the respective module, students attended a module covering history taking with a conclusive summative exam.

Study design

The module, in which the study was conducted, comprised a total of four mandatory serious gaming sessions each lasting 90 min. However, only the first two were relevant for the below mentioned analyses. Students were assigned randomized to one of two groups using one of two serious games. One group attended the sessions on-site and used the serious game EMERGE, while the other group attended online and used the serious game DIVINA. The specific modalities of the serious games led to the distinction into on-site and online sessions and medical experts supervised both sessions. During the sessions, students were instructed to engage with the respective serious game and received no further instructions. In case of questions, students could contact the medical experts at any time. Between the first and second session, students were provided with the previously described guideline covering relevant aspects of medical history taking for allowing retrieval of previously learned content. Students were urged to use the guideline only for preparation between the sessions and not during the second session. Since students attended the session online, we decided upon adding a control item to the subsequent questionnaire asking students if they used the guideline during the second session. Moreover, another item assessed whether students used the guideline for preparation. Besides these control items, the questionnaire also contained questions regarding students' self-assessment of their learning outcomes regarding history taking (see Supplement 2). The learning outcomes were analyzed using the comparative self-assessment (CSA) gain method [36]. Therefore, students rated the questions at the end of session 2 in a retrospective version (i.e., assessing their skills before the first session) and a post version (i.e., assessing their skills after the second session). The questionnaire was provided online via evasys (evasys GmbH, version 10.0). For a depiction of the experimental setup and the presented diseases in both games, see Table 1.

Serious games

Two serious games were used, each representing an emergency department, requiring students to assume the role of a virtual doctor on shift. Detailed descriptions of the serious games can be found in Aster et al. [37] and in Middeke et al. [38], characterizing DIVINA and EMERGE, respectively. Both serious games enable users to perform various actions typically carried out in an emergency department, such as taking medical histories, conducting examinations, initiating treatments, and discharging virtual patients. To this end, both serious games employ virtual patients, whose medical histories can be obtained through chat systems. Although the contents and the possible actions within the game are similar, they differ in their visual representation and in the way in which questions can be asked in their chat systems. DIVINA provides a chatbot, where users can enter self-formulated questions, whereas EMERGE provides a chat system based on a long menu format. In this format, users are generally presented with an overview of all possible questions. Additionally, users can enter a part of the sought question and receive a drop-down menu with all possible questions containing this part, allowing the users to choose the most suitable formulation of the question. Examples of the chat systems can be found in Aster et al. [37] and Middeke et al. [39] or in the Figure 1. It can be assumed that DIVINA enables free recall since no references for recognition are provided, while EMERGE relies on cued recall due to the long menu format [33].

Development of the history taking guideline

The authors specifically developed a history taking guideline for the execution of the study based on

Table 1. Depiction of the study procedure.

Session	Group 1	Group 2	Diseases for both groups
1	DIVINA	EMERGE	STEMI, NSTEMI, musculoskeletal chest pain, hypertensive crisis
	Individual studying of the history taking guideline		
2	DIVINA	EMERGE	STEMI, NSTEMI, musculoskeletal chest pain, hypertensive crisis, *congestive heart failure, aortic stenosis*
	Online questionnaire		

Diseases that were presented only in session two are written in italics.

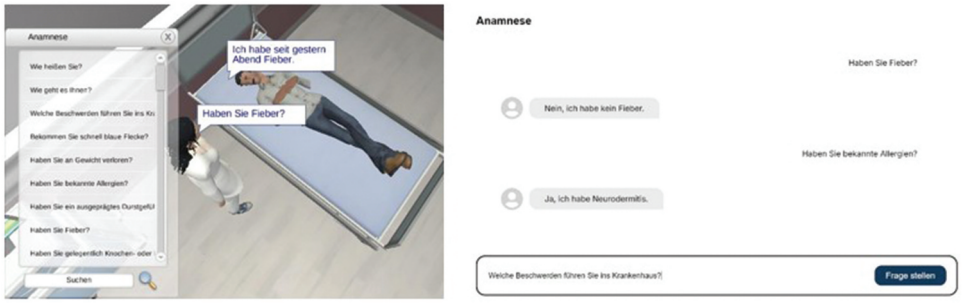


Figure 1. Screenshots of the chat systems in EMERGE and DIVINA.
The left screenshot represents the chat system in EMERGE, graphic by PatientZero Games GmbH. The right screenshot represents the chat system in DIVINA. The screenshots contain German questions, as the entire study was conducted in German.

the scoring checklist (see Supplement 1). Each checklists' category was also used as a category for the guideline resulting in five overarching sections (i.e., 1. current symptoms including a subcategory for a focused pain anamnesis; 2. pre-existing illnesses and past medical history including cardiovascular risk factors; 3. lifestyle and further risk factors; 4. medication anamnesis; 5. social anamnesis). Every category contained further specified sub-items representing relevant aspects that should be asked for the respective category during an ideal history taking (for the pain anamnesis in the first category: e.g., temporal occurrence, pain intensity). Moreover, every category contained two exemplary questions to which students could refer (e.g., for the second category pre-existing illnesses and past medical history: 'Do you have any pre-consisting illnesses?', 'Have you had an accident or fallen?'). Students used this guideline as a material for self-directed learning causing prior knowledge activation that could be applied during the second session. Additionally, students were instructed to use the guideline only for preparation and not during the second session.

Data analysis

To assess students' history taking, all entered qualitative questions were quantified. Two raters blindly evaluated the history data of each serious game and time point separately. In order to achieve a valid and reliable rating, a checklist for history taking data already developed for a previous study was used [35]. Some points of the checklist were refined in terms of ambiguous aspects as well as the achievable overall score and the respective points for each aspect. Furthermore, the checklist was refined for history taking in an emergency ward. Since each of the 26 questions could be scored with 10 points, a maximum history score of 260 was achievable. The checklist can be found in Supplement 1. These questions were just exemplarily and had not to be asked in the exact wording.

For the computation of the CSA gain values, first all missing values and the respective data pairs were excluded, before the mean self-assessment from the retrospective measurement was offset against the mean self-assessment from the post measurement using the formula provided in Raupach et al. [36]. Before analysis, data was prepared following consecutive steps. In a first step, all chats containing no queries were removed, resulting in a data set comprising only valid chats (i.e., chats including at least one query). Secondly, all confounding data was removed, including removing all data referring to the diseases occurring only in session 2 (i.e., congestive heart failure and aortic stenosis) as well as chats referring to students having attended only one session. The complete process of the derivation of the final data set is shown in Figure 2. All data was analyzed using IBM SPSS Statistics (Version 27) and Microsoft Excel. The assumption of normal distribution of the chat data was violated, except for EMERGE session 2, as well as for the proxy score of DIVINA session 1 used for the ANOVA analysis. Nevertheless, t-tests and ANOVAs are robust against this violation given a large sample size [40,41]. Since this requirement was given for the analysis of all hypotheses, an independent one-tailed t-test was conducted for H1, and paired one-tailed t-tests and a mixed ANOVA were conducted for H2.

Results

Descriptive data

The final data set consisted of 242 and 218 chats stemming from D1 and D2, respectively, as well as 402 and 316 chats stemming from E1 and E2. On average, students entered 3.51 valid chats in D1 and 3.16 chats in D2 as well as 5.66 chats in E1 and 4.45 in E2.

Regarding the control item whether students used the guideline during the second session, 89.9% participants from the DIVINA group ($n = 69$) and 80.8%

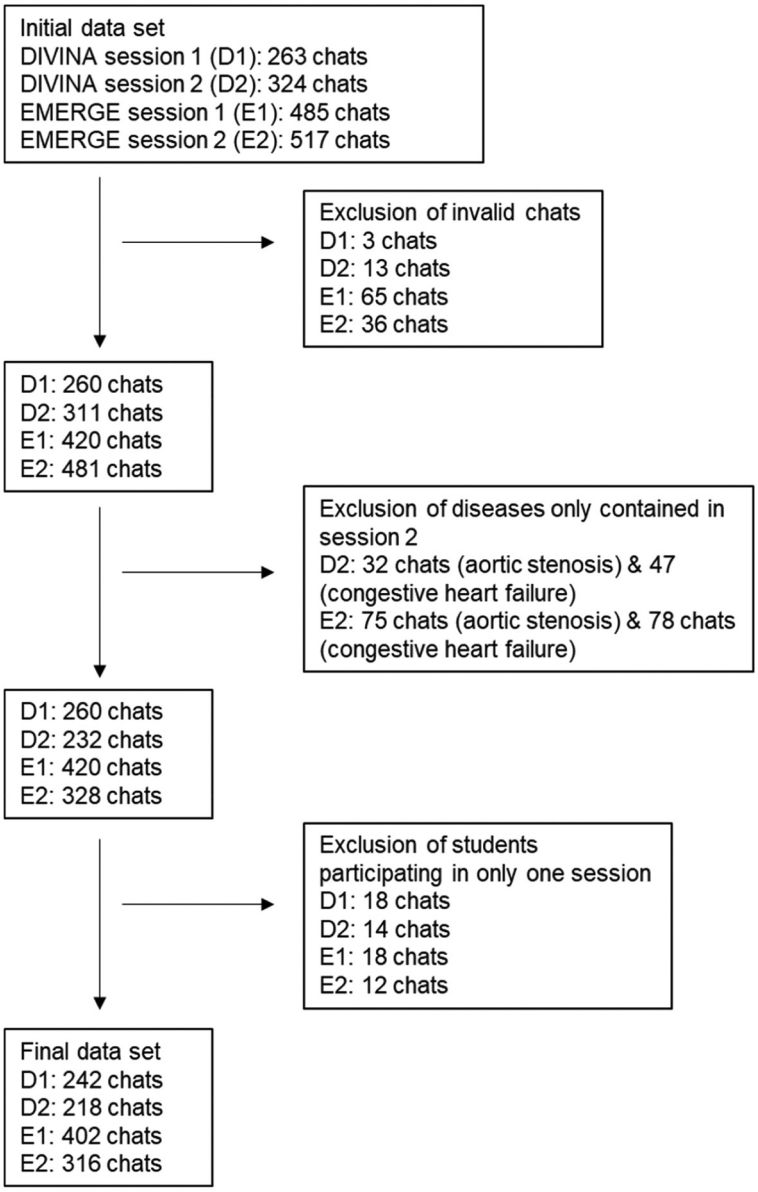


Figure 2. Process of the derivation of the final data set.
Each session is abbreviated with a capital letter referring to the serious game and a number referring to the respective session. Resulting in D1 for DIVINA session 1, D2 for DIVINA session 2, E1 for EMERGE session 1, and E2 for EMERGE session 2.

participants from the EMERGE group ($n = 52$) stated to have not used the guideline during the session. Only 60.9% of the DIVINA group and 47.2% of the EMERGE group ($n = 53$) have stated to have used the guideline in preparation.

History taking data

The analyses regarding the history data were conducted with a mean score of both raters for each chat. Both raters had a very good interrater reliability [42] as assessed by the intraclass correlation coefficient for each serious game session (D1: ICC .945; D2: ICC .960; E1: ICC .977; E2: ICC .965).

For the analysis of H1 whether students achieved a higher history score in E1 than in D1 as well as in D2, unpaired t-tests were conducted. Contrary to the assumption, the group comparison of E1 and D1 revealed no significant difference between both groups ($p > .05$) with on average slightly fewer points being achieved in E1 ($n = 402$; 78.8 ± 35.7) than in D1 ($n = 242$; 85.2 ± 27.7). The same pattern applies for the comparison between E1 and D2 revealing no significant difference ($p > .05$) with on average slightly fewer points being achieved in E1 (78.8 ± 35.7) than in D2 ($n = 218$; 88.3 ± 29.5).

The subsequent hypothesis 2 that students achieved significantly more points in D2 than in D1 but not in E2 compared to E1 was analyzed using two paired t-tests. The p-value was adjusted for the directed part of the hypothesis (D2 > D1) but not adjusted for the undirected part (E2 = E1). Even though data of students attending only one session were excluded from the analyses, each

student conducted a different number of chats in both sessions resulting in different amounts of chats between the sessions for each serious game. Therefore, it must be considered that the number of chats entering the analyses was adjusted for the execution of the paired t-tests. According to the assumption, students achieved a higher average score in D2 ($n = 218$; 88.3 ± 29.5) than in D1 ($n = 218$; 86.5 ± 27.9), although this difference was not significant, $t(217) = -0.622$, $p = .267$, $d = -0.062$. Contrary to the hypothesis that scores did not differ significantly between both EMERGE sessions, the difference between E1 ($n = 316$; 78.5 ± 34.6) and E2 ($n = 316$; 86.6 ± 35.0) was significant $t(315) = -2.918$, $p = .004$, $d = -0.229$. A possible explanation might be that the average amount of questions did not significantly differ between both DIVINA sessions (session D1: $M = 15.7$; session D2: $M = 15.3$), $t(216) = 0.509$, $p = .611$, $d = 0.052$, whereas the average amount significantly differed between both EMERGE sessions (session E1: $M = 10.1$; session E2: $M = 11$), $t(314) = -2.112$, $p = .035$, $d = -0.167$. Therefore, the significant improvement could be traced back to the increased amount of questions between the two EMERGE sessions.

A mixed ANOVA was conducted for analyzing possible interaction effects between the serious games and session. Therefore, a proxy variable was build calculating a mean value for every student over all conducted chats using the history score. The within subject factor had only two levels (i.e., session 1 and 2), which is why sphericity is automatically given. Levene's test showed that homogeneity of error variances was given ($p > .05$). No homogeneity of covariance was given, as assessed by Box's test ($p < .001$). Neither a significant interaction effect between session and group was found ($F(1, 138) = 1.051$, $p = .307$, partial $\eta^2 = .008$), nor a significant main effect for sessions ($F(1, 138) = 1.746$, $p = .189$, partial $\eta^2 = .012$) nor for group membership ($F(1, 138) = 0.172$, $p = .679$, partial $\eta^2 = .001$). The proxy variable was also used to create Figure 3 depicting the course of score improvements between the two sessions.

CSA gain

Additionally to the objective data, the subjective learning gain was assessed via a comparative self-assessment at the end of session 2. All subjective data was assessed anonymously, thus objective and subjective data could not be combined. On this account, the subjective data was not prepared like the objective data but all data fed into analysis. For the assessment of the subjective learning gain, all valid data pairs from retrospective and post assessment were included in the analysis. All CSA gain scores were compared question-wise between the serious games revealing no significant difference between the serious games on question level. As can be seen in Figure 4, attending EMERGE leads to an averagely higher increase in the self-assessment than attending DIVINA. Nevertheless, for both serious games only small improvements were found since CSA gain scores never exceeded 60% symbolizing a satisfactory learning success.

Discussion

General discussion

The study assessed whether providing students a guideline supporting self-directed learning improves history taking in a serious games' chatbot.

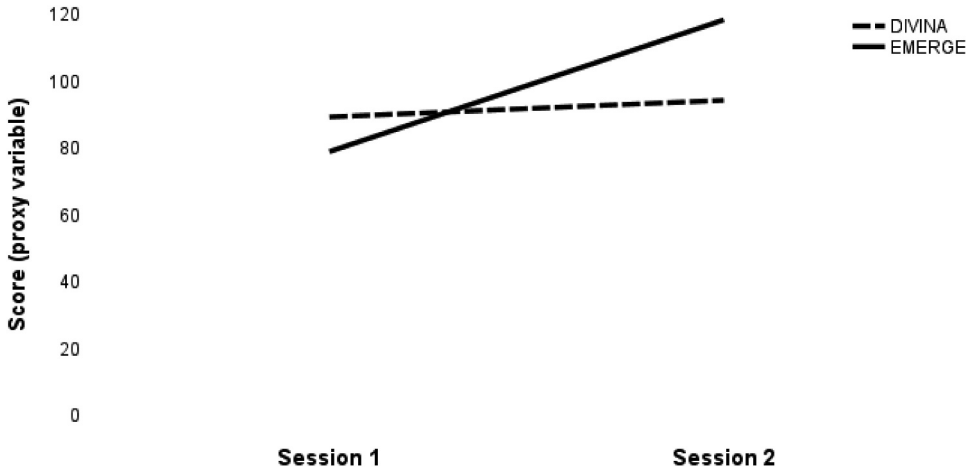


Figure 3. Course of the score improvements between the two sessions for each serious game.
The proxy variable build for the analysis of the mixed ANOVA was also used for the creation of the figure.

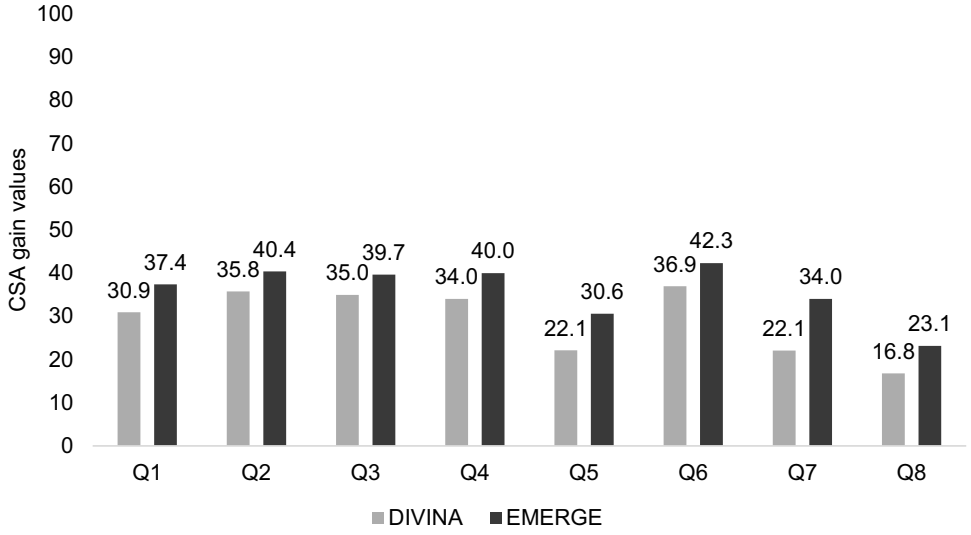


Figure 4. Mean CSA gain values for each question of both serious game groups.

Therefore, two different chatbots were compared, demonstrating that material for self-directed learning led to improved history taking in a long menu chatbot relying on cued recall but not in a free-entry chatbot relying on free recall.

The data did not justify rejection of the null hypothesis that there was no performance difference between the first EMERGE session and both DIVINA sessions. However, descriptively, students achieved higher average scores in both DIVINA sessions than in the first EMERGE session. Consequently, the assumption that the long menu format initially leads students to ask more questions has to be rejected.

The second hypothesis tested the effectiveness of supplementary material for self-directed learning, namely that students significantly improve their history taking through a guideline in DIVINA but not in EMERGE. According to the assumption, students obtained descriptively higher history scores after receiving the guideline in the second DIVINA session. Nevertheless, the comparison with the first DIVINA session was not significant. Based on our current results, our assumption must be reconsidered that the combination of a cognitive representation for history taking – induced by the guideline – and the possibility for exploration is beneficial in a free-entry chatbot. Contrary to the assumed hypothesis, the score significantly increased between the two EMERGE sessions. However, neither significant main effects of group and time nor an interaction effect was found. The significant improvement in EMERGE might apply to the assumption that cued recall due to questions being chosen from a long menu may be facilitated and students' retrieval after knowledge reactivation may be eased. Contrary to a previous study that found no improvement in history taking after attending EMERGE [13], our finding may be attributed to the interposed guideline, as all other conditions within the game remained identical. Although contrary to our assumed hypothesis, the result is in line with previous research demonstrating that cued recall leads to better results [33,43]. Since all students' queries were rated by means of a predefined checklist scoring questions relevant to history taking, it can be assumed that the found difference between the sessions is traceable to students asking more questions relevant for clinical reasoning. We assumed that using the guideline could have led to forming an internal representation of history taking as it was already discussed for clinical reasoning [34]. However, we analyzed the process data within the chatbot without explicitly questioning the students through focus groups or other qualitative measurements whether they developed an internal scheme. The question remains whether and why students were hindered to apply the reactivated knowledge in the free-entry chatbot during the second session or whether the free-entry chatbot itself hindered students to apply their reactivated knowledge. A potential drawback is the manner of question entry since students had to enter the entire question rather than only remembering certain keywords.

More than half of the students playing DIVINA declared to have used the guideline in preparation for the second session, while less than half of the students playing EMERGE stated to have used the guideline. However, since the individual questionnaire data could not be linked to the chatbot entries due to their anonymous assessment manner, no conclusion could be drawn regarding whether students who used the guideline improved more than those who reported to have not used it. Students' self-assessment of their learning improvement showed no significant difference between the serious games compared on question-level. Although the achieved CSA gain values did not reach the threshold symbolizing a satisfactory learning success, EMERGE tends to have led to

a higher increase in learning improvement than DIVINA. These subjective results align with the objective results, which showed that students' history taking improved in EMERGE but not in DIVINA. This suggests that the constrained chat system – with predefined questions – benefits more from an interposed guideline. Changes in the CSA gain allowed, to some extent, for inferences about whether students were able to reactivate and recall their knowledge. However, since no specific questions targeted the knowledge gained through use of the guideline, it is not possible to draw conclusions about individual differences in the benefit derived from it.

Irrespective of the significant changes between the sessions, it has to be mentioned that across both serious games and both sessions, students only achieved on average one-third of all possible points. Although they had previously attended a module on history taking, it can be assumed that applying or reactivating their knowledge is more challenging in a free recall chatbot compared to a cued recall chat system. Interestingly, more students who played the free-entry chatbot reported to have used the guideline for preparation than those who played the long menu chat system. Generally, serious games provide a safe learning environment and are frequently used in medical education [18], but it has to be scrutinized whether a serious game is suitable for training history taking for two reasons. First, a serious game is still an artificial learning environment and especially for history taking the human factor (e.g., non-verbal communication aspects, empathy) is missing. However, this might depend on the type of chatbot used, as previous research initially demonstrated that ChatGPT could be used for training empathic history taking [44]. Second, it is questionable whether history taking skills that are mostly acquired during interactions with real-life simulated patients can be transferred into an abstract setting like a serious game. The transfer performance is two-sided, the skills have not only to be transferred into the serious game but have also to be transferred from serious game to real life afterwards. Transfer performance may depend on the specific serious game and its inherent features, as previous research has shown mixed results – while transfer of in-game learned content is viable in some cases [29], it remains challenging in others [38,45]. Although, it could be shown that the transfer of learning after the use of a stand-alone chatbot was successful by leading to better results in a mock OSCE [17], more research is needed whether this also applies for chatbots embedded in serious games.

Limitations

The study design has several limitations. First, since an online study was conducted, it could not be guaranteed that students did not use any assistance during the sessions, especially the provided guideline during session two. Although the majority of students reported in the control item to have not used the guideline during the gaming session, this information could have been affected by biases such as the social desirability bias. Nevertheless, in case students had used the guideline during the session, it could be assumed that more points should have been achieved, thus it can be accepted that students indeed did not use the guideline during the session. During analysis, it became apparent that the chatbot in DIVINA sometimes mismatched questions and suitable answers and students therefore had to repeat questioning in different ways. This problem of natural language processing chatbots is already known in the literature but should best be avoided to provide a usable virtual patient [46]. Moreover, it could not be coded whether chats contained few questions because students did not ask any further questions or due to environmental conditions such as time limit of the session. This should be addressed in future studies by means of coding only chats of virtual patients that were discharged.

Second, during the development of the checklist, it was aimed to achieve a good cost-benefit ratio in terms of the time that accrues during rating the single chats. For this reason, it was decided to not score proceeding questions but only the first question of one rubric (e.g., 'do you smoke?'). Thus, not all relevant questions were covered, especially regarding attendant symptoms. It should be discussed whether more questions should be scored by a more detailed checklist or whether an efficient checklist is preferred. Simultaneously, a detailed checklist also needs students to ask more questions to receive a satisfactory score. Apart from a quantitative checklist, additional qualitative analysis of the history taking content should be considered.

Third, it is questionable whether the emergency department setting with its high urgency is a suitable learning environment for history taking. A precise history taking and arriving at the correct (differential) diagnosis is indeed needed in an emergency department; nevertheless, it might be impaired by the life threatening character of some diseases, the time pressure, as well as by the need to also conduct investigations in a low amount of time. Although this study focused solely on the chatbot, it cannot be ruled out that differences in the serious game's visual representations may have not influenced the history taking. Since such game design elements (i.e., chatbot) should be tested embedded in serious games and not standalone, it is always challenging to create consistent conditions.

Implications

Generally, serious games including a chatbot are useable for medical education since they provide a safe environment in which learning occurs without endangering patients' safety [22]. The study showed

that self-directed learning material in combination with a long menu format evocating cued recall is more suitable for learning history taking in a serious game than combined with a free-entry chatbot relying on free recall. Future studies should further investigate the differences between distinct types of chatbots and respective effect mechanisms. Moreover, it would be intriguing to answer the question whether it is more supportive for training history taking to orient towards prescribed questions from medical experts or to consider all relevant questions on one's own without a guidance but eventually learn in a more realistic way. It is conceivable that training history taking by means of a long menu chatbot is possibly more suitable for undergraduate medical education, while a free-entry chatbot is more suitable for postgraduate medical education. Future studies should compare students' objective exam data with the performance during a chatbot in order to assess its effectiveness. In line with this, future studies should investigate whether students are able to transfer learning from within a serious game to real exam settings or real-life situation, thus supporting a serious games' external validity.

Conclusion

This study assessed the influence of additional material for self-directed learning in the form of a history taking guideline on students' performance in history taking using chatbots embedded in serious games. The findings indicate that chatbots using a long menu format might benefit more from an interposed material for self-directed learning to increase students' history taking skills than chatbots using an open question entry format. Although students were initially better in a chatbot with free question entry, the provision of a written guideline only led to a significant improvement in the long menu chatbot. The study highlights the interplay between different types of chatbots and additional materials for self-directed learning in the teaching and training of medical history taking.

Acknowledgments

The author(s) used ChatGPT 3.5 in order to improve the readability and language of the manuscript. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the content.

Funding

The author (s) reported there is no funding associated with the work featured in this article.

Author contributions

AA: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft
AL: Formal analysis, Methodology, Writing – review & editing
MCL: Conceptualization, Writing – review & editing
TR: Investigation, Supervision, Writing – review & editing

Disclosure statement

The author TR declares a financial conflict of interest as he holds shares in the company Yellowbird Consulting LTD that has developed the serious game DIVINA referred to in this article. No other author has competing interests.

Data availability statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

ORCID

Alexandra Aster http://orcid.org/0000-0003-0407-3250
Matthias Carl Laupichler http://orcid.org/0000-0003-3104-1123
Tobias Raupach http://orcid.org/0000-0003-2555-8097

References

[1] Fürstenberg S, Helm T, Prediger S, et al. Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators. BMC Med Educ. 2020;20(1):368. doi: 10.1186/s12909-020-02260-9
[2] Peterson MC, Holbrook JC, Von Hales DE, et al. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med. 1992;156(2):163–165.
[3] Manalastas G, Noble LM, Viney R, et al. What does the structure of a medical consultation look like? A new method for visualising doctor-patient communication. Patient Educ Couns. 2021;104(6):1387–1397. doi: 10.1016/j.pec.2020.11.026
[4] Skotzko CE, Vrinceanu A, Krueger L, et al. Alcohol use and congestive heart failure: incidence, importance, and approaches to improved history taking. Heart Fail Rev. 2009;14(1):51–55. doi: 10.1007/s10741-007-9048-8
[5] Schwitzguebel AJ, Jeckelmann C, Gavinio R, et al. Differential diagnosis assessment in ambulatory care with an automated medical history-taking device: pilot randomized controlled trial. JMIR Med Inform. 2019;7(4):e14044. doi: 10.2196/14044
[6] Collins LC, Gablasova D, Pill J. 'Doing questioning' in the emergency department (ED). Health Commun. 2023;38(12):2721–2729. doi: 10.1080/10410236.2022.2111630
[7] Keifenheim KE, Teufel M, Ip J, et al. Teaching history taking to medical students: a systematic review. BMC Med Educ. 2015;15(1):159. doi: 10.1186/s12909-015-0443-x
[8] Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE guide no 42.

Med Teach. 2009;31(6):477–486. doi: 10.1080/01421590903002821
[9] Kaplonyi J, Bowles KA, Nestel D, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ. 2017;51(12):1209–1219. doi: 10.1111/medu.13387
[10] Maicher KR, Zimmerman L, Wilcox B, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach. 2019;41(9):1053–1059. doi: 10.1080/0142159X.2019.1616683
[11] Alyami H, Alawami M, Lyndon M, et al. Impact of using a 3D visual metaphor serious game to Teach history-taking content to medical students: longitudinal mixed methods pilot study. JMIR Serious Games. 2019;7(3):e13748. doi: 10.2196/13748
[12] Adamopoulou E, Moussiades L. An overview of chatbot technology. Artif Intel Appl Innovations. 2020:373–383. doi: 10.1007/978-3-030-49186-4_31
[13] Raupach T, de Temple I, Middeke A, et al. Effectiveness of a serious game addressing guideline adherence: cohort study with 1.5-year follow-up. BMC Med Educ. 2021;21(1):189. doi: 10.1186/s12909-021-02591-1
[14] Co M, John Yuen TH, Cheung HH. Using clinical history taking chatbot mobile app for clinical bedside teachings - a prospective case control study. Heliyon. 2022;8(6):e09751. doi: 10.1016/j.heliyon.2022.e09751
[15] Holderried F, Stegemann-Philipps C, Herrmann-Werner A, et al. A language model-powered simulated patient with automated feedback for history taking: prospective study. JMIR Med Educ. 2024;10:e59213. doi: 10.2196/59213
[16] Holderried F, Stegemann-Philipps C, Herschbach L, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ. 2024;10:e53961. doi: 10.2196/53961
[17] Raafat NH, Harbourne AD, Radia K, et al. Virtual patients improve history-taking competence and confidence in medical students. Med Teach. 2024;46(5):682–688. doi: 10.1080/0142159X.2023.2273782
[18] Aster A, Laupichler MC, Zimmer S, et al. Game design elements of serious games in the education of medical and healthcare professions: a mixed-methods systematic review of underlying theories and teaching effectiveness. Adv Health Sci Educ Theory Pract. 2024;29(5):1825–1848. doi: 10.1007/s10459-024-10327-1
[19] Deci EL, Ryan RM. The general causality orientations scale: self-determination in personality. J Res Personality. 1985;19(2):109–134. doi: 10.1016/0092-6566(85)90023-6
[20] Csikszentmihalyi M. Beyond boredom and anxiety. San Francisco (CA): Jossey-Bass; 1975.
[21] Krath J, Schürmann L, von Korflesch HFO. Revealing the theoretical basis of gamification: a systematic review and analysis of theory in research on gamification, serious games and game-based learning. Comput Hum Behav. 2021;125. doi: 10.1016/j.chb.2021.106963
[22] Sharifzadeh N, Kharrazi H, Nazari E, et al. Health education serious games targeting health care providers, patients, and public health users: scoping review. JMIR Serious Games. 2020;8(1):e13459. doi: 10.2196/13459
[23] Tan AJQ, Lee CCS, Lin PY, et al. Designing and evaluating the effectiveness of a serious game for safe administration of blood transfusion: a randomized controlled trial. Nurse Educ Today. 2017;55:38–44. doi: 10.1016/j.nedt.2017.04.027
[24] Knowles MS. Self-directed learning: a guide for learners and teachers. New York (NY): Association Press; 1975.
[25] Loyens SMM, Magda J, Rikers RMJP. Self-directed learning in problem-based learning and its relationships with self-regulated learning. Educ Psychol Rev. 2008;20(4):411–427. doi: 10.1007/s10648-008-9082-7
[26] Ryan RM, Deci EL. Self-regulation and the problem of human autonomy: does psychology need choice, self-determination, and will? J Pers. 2006;74(6):1557–1585. doi: 10.1111/j.1467-6494.2006.00420.x
[27] Dankbaar ME, Roozeboom MB, Oprins EA, et al. Preparing residents effectively in emergency skills training with a serious game. Simul Healthc. 2017;12(1):9–16. doi: 10.1097/SIH.0000000000000194
[28] Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol. 2000;25(1):54–67. doi: 10.1006/ceps.1999.1020
[29] Buijs-Spanjers KR, Harmsen A, Hegge HH, et al. The influence of a serious game's narrative on students' attitudes and learning experiences regarding delirium: an interview study. BMC Med Educ. 2020;20(1). doi: 10.1186/s12909-020-02210-5
[30] Mohan D, Angus DC, Ricketts D, et al. Assessing the validity of using serious game technology to analyze physician decision making. PLOS ONE. 2014;9(8):e105445. doi: 10.1371/journal.pone.0105445
[31] Berglund L, von Knorring J, McGrath A. When theory meets reality- a mismatch in communication: a qualitative study of clinical transition from communication skills training to the surgical ward. BMC Med Educ. 2023;23(1):728. doi: 10.1186/s12909-023-04633-2
[32] van den Eertwegh V, van Dulmen S, van Dalen J, et al. Learning in context: identifying gaps in research on the transfer of medical communication skills to the clinical workplace. Patient Educ Couns. 2013;90(2):184–192. doi: 10.1016/j.pec.2012.06.008
[33] Higham PA, Guzel MA. Cued recall. In: Seel NM, editor. Encyclopedia of the sciences of learning. Boston (MA): Springer; 2012. p. 868–871. doi: 10.1007/978-1-4419-1428-6
[34] Young M, Thomas A, Lubarsky S, et al. Drawing boundaries: the difficulty in defining clinical reasoning. Acad Med. 2018;93(7):990–995. doi: 10.1097/ACM.0000000000002142
[35] Aster A, Lotz A, Raupach T. Theoretical background of the game design element "chatbot" in serious games for medical education. Adv Simul. 2025;10(1):Article 10. doi: 10.1186/s41077-025-00341-7
[36] Raupach T, Munscher C, Beissbarth T, et al. Towards outcome-based programme evaluation: using student comparative self-assessments to determine teaching effectiveness. Med Teach. 2011;33(8):e446–453. doi: 10.3109/0142159X.2011.586751
[37] Aster A, Hütt C, Morton C, et al. Development and evaluation of an emergency department serious game for undergraduate medical students. BMC Med Educ. 2024;24(1):1061. doi: 10.1186/s12909-024-06056-z
[38] Middeke A, Anders S, Raupach T, et al. Transfer of clinical reasoning trained with a serious game to comparable clinical problems: a prospective randomized

study. Simul Healthc. 2020;15(2):75–81. doi: 10.1097/SIH.0000000000000407

[39] Middeke A, Anders S, Schuelper M, et al. Training of clinical reasoning with a serious game versus small-group problem-based learning: a prospective study. PLOS ONE. 2018;13(9):e0203851. doi: 10.1371/journal.pone.0203851

[40] Glass GV, Peckham PD, Sanders JR. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Rev Educ Res. 1972;42(3):237–288. doi: 10.3102/00346543042003237

[41] Rasch D, Guiard V. The robustness of parametric statistical methods. Psychol Sci. 2004;46:175–208. doi: 10.1007/978-94-009-6528-7_20

[42] Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284. doi: 10.1037/1040-3590.6.4.284

[43] Tulving E, Pearlstone Z. Availability versus accessibility of information in memory for words. J Verbal Learn Verbal Behav. 1966;5(4):381–391. doi: 10.1016/S0022-5371(66)80048-8

[44] Aster A, Ragaller SV, Raupach T, et al. ChatGPT as a virtual patient: written empathic expressions during medical history taking. Med Sci Educ. 2025. doi: 10.1007/s40670-025-02342-7

[45] Aster A, Scheithauer S, Middeke AC, et al. Use of a serious game to teach infectious disease management in medical school: effectiveness and transfer to a clinical examination. Front Med. 2022;9:863764. doi: 10.3389/fmed.2022.863764

[46] Maicher K, Danforth D, Price A, et al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. Simul Healthc. 2017;12(2):124–131. doi: 10.1097/SIH.0000000000000195

Medical Science Educator (2025) 35:1513–1522
https://doi.org/10.1007/s40670-025-02342-7

ORIGINAL RESEARCH

ChatGPT as a Virtual Patient: Written Empathic Expressions During Medical History Taking

Alexandra Aster[1] · **Sophia Viktoria Ragaller**[1] · **Tobias Raupach**[1] · **Ambra Marx**[2]

Accepted: 17 February 2025 / Published online: 27 February 2025

Abstract

Objective Virtual patients are already utilized in the teaching of medical history taking. Since its emergence, ChatGPT has been integrated into several areas of medical education. This study aimed to examine whether ChatGPT can be used to train empathic history taking while fostering students' subjective autonomy.

Methods Third-year medical students took histories with ChatGPT 3.5 after entering a predefined prompt covering cardiological diseases. Afterwards, students answered a questionnaire regarding their experienced autonomy. All chats were analyzed using the Empathic Communication Coding System measuring ChatGPT's given empathic opportunities as well as students' responses.

Results Out of 659 interactions, 93 were identified as empathic. ChatGPT provided opportunities mostly through reporting emotional statements or challenges. Students sometimes missed reacting adequately to ChatGPT's opportunities but more often responded by implicit recognition of patient perspective and reported a high level of experienced autonomy.

Conclusions The study yielded preliminary results that ChatGPT might be suitable as a tool mimicking a virtual patient while enabling an empathic history taking. To date, ChatGPT seems valid as a supplement to training with simulated patients. Medical faculty could consider integrating ChatGPT into teaching, such as through a flipped classroom approach, to guide students in its use as ChatGPT continues to gain attention.

Keywords Empathy · ChatGPT · Medical education · History taking · Virtual patient

Introduction

Empathy has not only been shown to improve the physician–patient-interaction, patients' therapy adherence, and patient treatment outcomes but has also been shown to be a protective factor for physicians' well-being [1–4]. Conversely, a physicians' burnout negatively influences patients' experience during a consultation [5]. Communication trainings, especially empathy interventions, significantly improve medical students' empathy and have an impact on physicians showing more empathy in patient encounters afterwards [6, 7]. In this sense, the relevance of strengthening empathy during medical education and especially teaching empathic history taking becomes apparent. Generally, Cuff, Brown [8] contextualize empathy as follows:

> "Empathy is an emotional response (affective), dependent upon the interaction between trait capacities and state influences. […] The resulting emotion is similar to one's perception (directly experienced or imagined) and understanding (cognitive empathy) of the stimulus emotion, with recognition that the source of the emotion is not one's own." (p. 150)

During communication, empathy can be expressed threefold: through verbal, non-verbal, and paraverbal components. Non-verbal aspects are expressed through perceptible behaviors such as general posture and body movements [4, 9]. Paraverbal components are implicit aspects of communication that pertain to the expression of speech, such as vocal quality and prosodic characteristics [9]. Verbal aspects cover explicit as well as implicit reactions and responses to a message, for instance confirmation or support [10]. Another related factor in empathic communication is the technique of

✉ Alexandra Aster
alexandra.aster@ukbonn.de

[1] Institute of Medical Education, University Hospital Bonn, Bonn, Germany

[2] Department of Psychosomatic Medicine and Psychotherapy, University Hospital Bonn, Bonn, Germany

active listening as proposed by Rogers [11]. Active listening consists of three components: nonverbal engagement, verbal paraphrasing, and further questioning [12]. It has been shown to foster feelings of understanding and greater satisfaction [12] and is significantly associated with empathy in healthcare professions [13]. Beyond general empathy, the construct of clinical empathy that is defined as "a kind of emotional reasoning that allows physicians to incorporate emotional experiences as part of clinical decision-making" [14, p. 97] counts for the medical context. Clinical empathy has two sides of the same coin: it is emotional labor that healthcare professionals must manage effectively, but when routinized, it enhances medical encounters [15]. Researchers recommend incorporating practical communication training into medical education, emphasizing the hands-on aspect, as empathy must be developed through practice rather than theoretical instruction [4, 16]. Conducting communication trainings have been shown to help cardiologists improve their empathy [17]. To date, communication trainings in medical education are mostly conducted by using simulated patients (SP) [18]. According to Barrows [19], a simulated patient is a non-ill individual who, after thorough training, acts as a patient with a specified disease. Additionally, actual patients trained to present their own diseases in a standardized manner can also be classified under the umbrella term "standardized patients" along with simulated patients [19, 20]. However, virtual patients (VPs) have been increasingly and effectively used in recent years [21, 22]. Although SPs are widely recognized as a valuable component of communication training, implementing and maintaining an SP program is time-, labor-, and resource-intensive [23]. In contrast, while VPs do require resources to be set up, they offer learners the opportunity to repeat their training indefinitely [21] and shape their learning in a self-paced manner. Additionally, it is conceivable that using VPs may appeal to students' feeling of autonomy, a basic psychological need according to the self-determination theory [24]. In a previous study, students favored the use of VPs compared to lectures when it comes to promoting self-directed learning [25]. Self-directed learning is linked with feelings of autonomy [26, 27], suggesting that students' subjectively experienced autonomy is addressed by using VPs. In communication training, it is essential to develop not only hard skills, such as information gathering, but also soft skills, like empathy. Studies have shown that VPs can be effectively used to train empathic history taking [28, 29]. The advent of large language models, such as ChatGPT in November 2022, has created an opportunity to implement VPs. Literature suggests that for VPs to be effective in terms of learning outcomes, their textual setup must be based on valid theoretical foundations [21], which may also apply to the setup of ChatGPT as a VP. A handful of studies have utilized custom generative pre-trained transformers (GPTs), primarily based on ChatGPT, to train medical history taking in various fields, including dental [30] and medical education [31]. While SPs and VPs are already known to be feasible for practicing empathic history taking, it remains unknown whether ChatGPT can provide a feasible environment conducive to practicing empathic history taking that also supports students' feelings of autonomy.

Therefore, within this study, the following research questions were assessed:

RQ 1: Can ChatGPT be effectively used to conduct empathic history taking?
RQ 2: How do students rate their experienced autonomy when taking a history with ChatGPT?

Material and Methods

Study Procedure

The study took place at a German medical school in summer term 2024 after being approved by the local ethics committee (application number: 2024–96-BO). Data collection directly followed each session of a course teaching the foundations of empathic history taking for third-year medical students. Each student participated in one session of the course. All attending students were invited to voluntarily participate in the study and were quasi-randomly assigned to one of four conditions. In the third year of their degree program, students begin their clinical training. Training in empathic communication is also provided through seminars and lectures in the semesters during the clinical training. Students provided implied informed consent by completing the study. Initially, they were instructed to take a written medical history with ChatGPT as a VP. However, they were not explicitly instructed to practice empathic communication, as this could have biased the results, for example due to social desirability. At this stage of their studies, students had not yet attended any courses covering clinical content. However, this is not a limitation, as the study aimed to assess empathic history taking, which was the primary learning objective of the course preceding the data collection. From a patient's perspective, empathy is ranked among the top three most important factors in an emergency department [32]. Therefore, the storylines for the presented cases were based on cardiology and emergency medicine (Table 1). More specifically, endocarditis and heart failure were chosen as the two cardiological diseases for the study. Additionally, it was of interest whether a change in the temperature of ChatGPT's response has an impact on students' expressed empathy. Specifying the temperature adjusts ChatGPT's responses in terms of creativity by influencing how likely words are strung together.

Table 1 Structure of applied ChatGPT prompt

Structure	Prompt elements
Basic structure part 1	Hello. Assume the role of a standardized patient, adjust your responses according to the patient's condition, and independently decide the level of detail in your answers. A standardized patient is an actor who portrays the role of a real patient as authentically as possible. Therefore, the standardized patient does not understand or interpret the clinical findings provided below and cannot answer any regarding questions. Do not use medical jargon in your responses. Response temperature: 0.2 / 0.8
Medical information 1 (endocarditis)	An 81-year-old man presents to the emergency department, reporting fatigue. He arrived at the emergency department on his own and has a reddish face color. Additionally, he reports suffering from muscle pain. Osler nodes and positive blood cultures are observed
Medical information 2 (heart failure)	A 75-year-old man presents to the emergency department reporting shortness of breath. He was brought in by emergency services and appears pale. Additionally, he reports frequent evening ankle swelling. Initial physical examination reveals the following clinical findings: the first heart sound is relatively quiet, a holosystolic murmur at the 5th intercostal space, left midclavicular line, and a slightly elevated jugular venous pressure
Basic structure part 2	You will now be asked medical history questions, which you should answer accurately in your role as a standardized patient, according to the case. The illness itself must not be mentioned at any point during the medical history. The history-taking process ends only when the correct diagnosis is made. Introduce yourself with a full name at the beginning of the history-taking. Do not ask any questions at the start

A single response temperature was chosen for each condition

In this sense, a lower temperature (e.g., 0.2) yields more focused and concrete answers, as subsequent words are chosen based on higher probabilities [33, 34]. Conversely, a higher temperature (e.g., 0.8) results in more random and creative responses by selecting words based on lower probabilities, making it more suitable for training communicative skills [33, 34]. Both diseases were presented with a response temperature of either 0.2 or 0.8, resulting in the four conditions: endocarditis 0.2, endocarditis 0.8, heart failure 0.2, and heart failure 0.8. The respective prompt was given to students via an online storage platform. Students pasted the prompt in their own ChatGPT 3.5 accounts and had 60 min to perform the history taking. Throughout and after the session, no feedback was provided to the students regarding either the content of the medical histories or their expressed empathy. Students anonymously shared their chat transcripts for data analysis by providing the export link for the respective chat and completed the short questionnaire described below.

Materials

Development of the ChatGPT Prompt

To enable ChatGPT to act as a VP, a basic structure instructing ChatGPT how to take on its role was created for each prompt, which was then tailored with the specific story of the respective disease. Table 1 shows the prompt structure, which — for the purpose of this publication — was translated to English since the study was conducted in German. For each condition, a disease storyline and a response temperature were selected. To ensure external validity, the disease was not explicitly named in the prompt, as most patients do not present with a final diagnosis in real life. The basic structure part 2 included instructions that history taking should continue until the correct diagnosis was made. However, entering a diagnosis was not mandatory, and students could conclude history taking whenever they felt they had gathered sufficient information. However, it has proven to be effective during the preceding prompt testing to add this instruction as a security measure for ChatGPT to stay in its role and to not accidentally name the disease.

Questionnaire

A short questionnaire was developed to measure two variables. The first variable assessed students' self-reported feelings of autonomy during the history taking with ChatGPT, using two autonomy scales—i.e., freedom of choice, and task relevance—postulated by Sailer [35]. Both scales were adapted to fit the specific context of history taking (e.g., "I was able to decide for myself, which history taking questions I wanted to ask."). All six questions were measured using a 7-point Likert scale. The second variable assessed students' prior use of ChatGPT, ranging from no experience to extensive experience on a 5-point scale.

Choice of Tool for Data Analysis

To assess empathic interactions, the Empathic Communication Coding System (ECCS) [10, 36] was applied, as it allowed for separately coding ChatGPT's provided empathic opportunities and medical students' corresponding empathic responses. ChatGPT's statements were analyzed as opportunities according to the ECCS. According to the manuals

provided by the International Association for Communication in Healthcare [EACH, 37], opportunities can be provided on three different levels while responses can be provided on seven levels. For an overview and a description of the levels, see Table 2.

Coding ChatGPT's provided opportunities enabled verification of whether it had offered empathic opportunities at all. For ChatGPT to be effectively used as a VP for training empathic communication, it must independently generate empathic opportunities, as these cannot be as easily preconfigured in the publicly available version as in predefined VPs. Thus, it is essential not only to analyze students' written empathic responses but also to evaluate ChatGPT's suitability as a VP in imitating human conversation and interaction. This approach enables the examination of empathic interplay and the analysis of interactions between medical students and ChatGPT as a VP.

Data Analysis

Two psychologists independently rated each student's history taking chat protocol using the ECCS. Beforehand, both raters became acquainted with the coding scheme by studying the manuals provided by EACH [37]. During the rating, the raters adhered to the manuals and referred to it in case of ambiguous ratings. Additionally, three test ratings with chat protocols, which were not included in the final analysis, were conducted to ensure both raters reached a common understanding. In contrast to a face-to-face conversation, interaction with ChatGPT does not constitute a traditional dialogue, as ChatGPT provides a response to which the user reacts in isolation. Consequently, both raters evaluated each pair comprising ChatGPT's statement and the subsequent medical students' reaction. If ChatGPT did not present an empathic opportunity, it was coded as "not applicable". Students' responses were also coded as "not applicable" due to the absence of an opportunity for the student to react empathetically. Interrater reliability was assessed using the intraclass correlation coefficient, resulting in a good agreement of 0.770, according to Cicchetti [38]. Both raters resolved discrepant ratings, resulting in agreed-upon values for each rating used for the analysis.

Results

Thirty-five third-year medical students participated in the study, with a total of 28 valid chat protocols included. The discrepancy between the number of participants and valid chat protocols was due to technical issues with exporting the chat data. No additional exclusion criteria for chat protocols were applied. On average, students made 24.96 ($SD = 9.41$) entries in ChatGPT during their history takings.

Table 2 Overview and descriptions of the levels used for coding ChatGPT's opportunities and students responses

ChatGPT's opportunities		Student responses	
Level	Description	Level	Description
1	*Emotion statements* ChatGPT describing current feelings of emotions	0	*Denial of ChatGPT's perspective* Ignoring or disconfirming ChatGPT's perspective
2	*Progress statements* ChatGPT describing positive developments or an improved quality of life	1	*Perfunctory recognition of ChatGPT's perspective* Superficial recognitions of ChatGPT's perspective without explicitly acknowledging it (e.g. "hmm"). Hence, this level cannot be adequately represented in exclusively written communication with ChatGPT
3	*Challenge statements* ChatGPT describing the negative impact that problems have on the quality of life or reports of radical life events	2	*Implicit recognition of ChatGPT's perspective* Implicit recognition without focusing on the central issue but pointing out a peripheral aspect
		3	*Acknowledgement without pursuit* Acknowledging ChatGPT's statements without further pursuit
		4	*Acknowledgement with pursuit* Acknowledging ChatGPT's statements with further pursuit
		5	*Confirmation* Confirmation of ChatGPT's perspective that can be expressed through various ways
		6	*Shared Feeling or Experience* Sharing their own feelings or experiences with ChatGPT

Based on the works by Bylund and Makoul [10] and Bylund and Makoul [36], and the manuals provided by EACH [37]. All headings are extracted exactly from the manuals, with the exception of the word patient being changed to ChatGPT

Empathy Ratings

Out of 659 general interactions (i.e., all students' entries and ChatGPT statements taken together irrespective of empathic content), 93 (14%) empathic interactions between ChatGPT and medical students could be identified across both diseases and response temperatures. The highest number of empathic interactions was observed for endocarditis, with 27 interactions at both response temperatures. All empathic opportunities and responses broken down to each disease and the respective response temperature can be found in Table 3. A detailed analysis of both interaction components, ChatGPT's empathic opportunities and students' responses, is presented in the following paragraphs.

ChatGPT primarily provided empathic opportunities at level 1 ($n = 50$) across all conditions, directly followed by level 3 ($n = 41$) and level 2 ($n = 2$). The number of provided opportunities did not differ significantly between response temperatures ($U = 1045.00$, $p = 0.852$). ChatGPT predominantly used statements of emotion or challenge, while progress statements were rarely expressed. It maintained its role stringently and coherently. Descriptively, ChatGPT provided slightly more emotion-related statements at the lower response temperature, where only two progress statements were detected. A closer look at the questions posed by students revealed that ChatGPT allowed students to ask a substantial number of relevant questions and provided satisfactory answers. From a qualitative perspective, it is striking that ChatGPT commonly draws on the same storylines, expanding beyond the information provided in the prompt, particularly within the same session. Moreover, it is evident that ChatGPT adjusts its responses based on the students' questioning behavior.

Students' mostly showed empathic responses at level 2 ($n = 41$) across all diseases and response temperatures. The number of empathic responses did not differ significantly between the response temperatures ($U = 912.00$, $p = 0.200$). Although only a descriptive trend, slightly more responses were observed at levels 5 and 6 for the lower response temperature. Across all chats, students most commonly provided either one or three empathic responses per chat (Fig. 1). However, in 34 instances, students did not respond to ChatGPT's empathic opportunities at all.

Specifically, the most frequent empathic interactions occurred with the level combinations 1–0, 1–2, and 3–2 with 20 occurrences each, irrespective of disease and response temperature. Figure 2 illustrates the distribution as well as the different levels on which opportunity-response pairs occurred.

Students' Feelings of Autonomy

Twenty-five students completed the questionnaire regarding perceived autonomy. When considering both scales together, students reached an average of $M = 38.2$ ($SD = 3.44$) out of a maximum of 42. Broken down to the single scales, students rated their autonomy on the scale "freedom of choice" with an average of $M = 6.8$ ($SD = 0.54$) out of a maximum of 7. On the scale "task relevance", students rated their experienced autonomy with an average of $M = 5.93$ ($SD = 1.09$) out of a maximum of 7. The study in which the scale was developed [35] can be referenced to obtain mean values and

Table 3 Descriptive distribution of levels in opportunities and responses

Levels of the respective opportunities and responses			
All diseases and response temperatures ($N = 28$)			
ChatGPT's opportunities		*Student responses*	
Level 1 = 50		Level 0 = 34	
Level 2 = 2		Level 1 = 0	
Level 3 = 41		Level 2 = 41	
		Level 3 = 3	
		Level 4 = 7	
		Level 5 = 8	
		Level 6 = 0	
Response temperature 0.2 across both diseases ($n = 16$)		**Response temperature 0.8 across both diseases** ($n = 12$)	
Opportunity	*Response*	*Opportunity*	*Response*
Level 1 = 28	Level 0 = 18	Level 1 = 22	Level 0 = 16
Level 2 = 2	Level 1 = 0	Level 2 = 0	Level 1 = 0
Level 3 = 22	Level 2 = 20	Level 3 = 19	Level 2 = 21
	Level 3 = 2		Level 3 = 1
	Level 4 = 6		Level 4 = 1
	Level 5 = 6		Level 5 = 2
	Level 6 = 0		Level 6 = 0

The levels arranged parallel do not match correspondingly to each other but independently represent the respective numbers

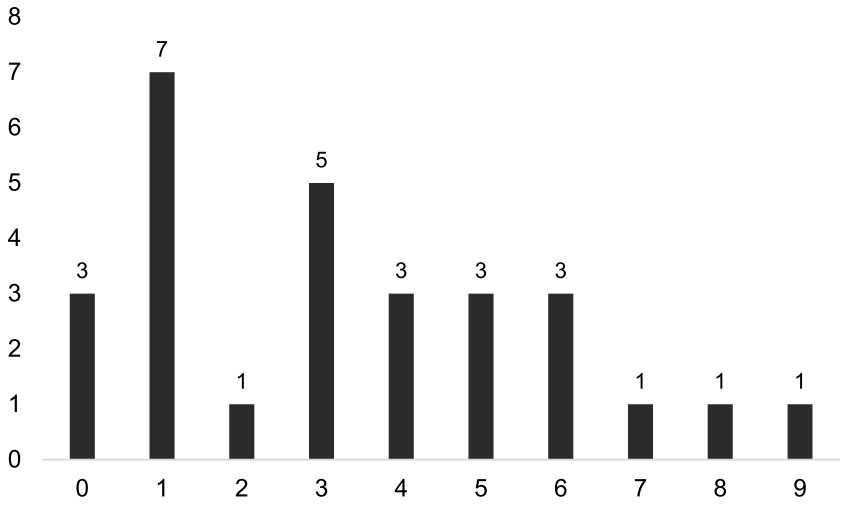


Fig. 1 Number of empathic responses per chat and distribution across all chats. X-axis, number of empathic responses per chat; Y-axis, number of occurrence of the respective number across all chats, which is also depicted above each bar

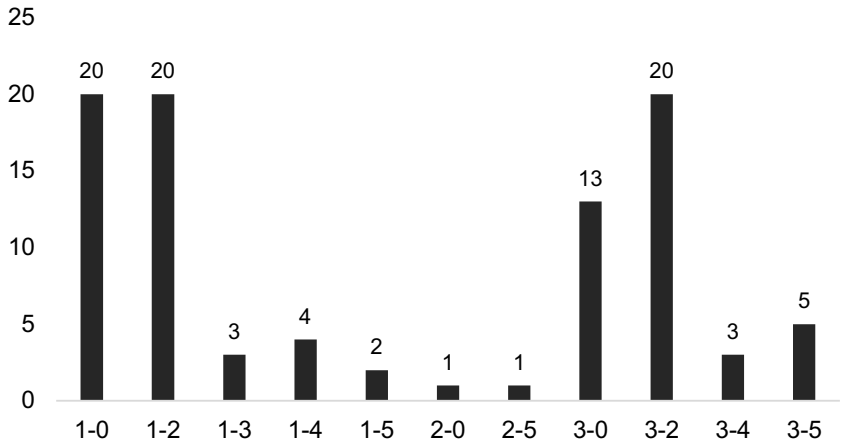


Fig. 2 Distribution and absolute number of opportunity-response pairs across all diseases and response temperatures. X-axis, opportunity-response pairs with the left number representing the level of the opportunity and the right number representing the level of the response; Y-axis, absolute number of occurrences, which is also depicted above each bar

standard deviations for comparing with autonomy scores in the current study. In the referenced study, the gamification group had an average of $M = 4.03$ ($SD = 1.49$) on the "freedom of choice" subscale and $M = 5.46$ ($SD = 1.06$) on the "task relevance" subscale [35]. In comparison, subjective autonomy when using ChatGPT was higher in the current study.

In a separate question, students rated their previous experience with ChatGPT. It showed that students' experience was nearly normally distributed with the majority of students reporting a medium level of previous experience ($M = 2.48$, $SD = 0.98$).

Discussion and Conclusion

To our knowledge, this is the first study to examine whether medical students can practice empathic history taking with ChatGPT as a virtual patient. Given ChatGPT's widespread recognition and accessibility, assessing its effectiveness for medical training is essential. ChatGPT's empathy has been examined in the context of responding to patient questions [39], and patients' perceptions of ChatGPT's responses compared to specialists' responses [40], but not yet within the context of practicing empathic history taking. The study results indicated that while ChatGPT was able to provide empathic opportunities, the majority of its statements were factual rather than empathic. Building on this, students were able to show empathic responses, although the responses were still predominantly factual. Nevertheless, a relatively small number of empathic interactions between students and ChatGPT occurred. Taken together, ChatGPT in its free version presents certain challenges but may also serve as a feasible tool for practicing empathic communication. Adjusting the response temperature in the prompt did not result in any significant differences in the number of provided opportunities or in the responses given. Moreover, students reported high feelings of autonomy during history taking. Collectively, the results suggest that ChatGPT might be a potentially valuable and feasible option for practicing history taking, even though the results can only be interpreted as preliminary at this stage and further research is needed.

Given that students demonstrated adequate questions on a medical level, it is worth questioning whether ChatGPT might be more suitable for learning the content of history taking rather than training empathic communication. Supporting this assumption, Deladisma and Cohen [41] demonstrated that medical students interacted empathically with VPs, but interactions were insufficient on a quantitative and qualitative level to replace real-life interactions such as encounters with SPs. At the end of the session, some students verbally reported feeling hindered in expressing empathy due to the nature of ChatGPT requiring questions to be entered in writing into a technical device. Others reported difficulty in demonstrating empathy towards an AI as a communication partner. However, since ChatGPT provided humanoid conversation, the technical aspect might be a more valid reason than the quality of communication. These assumptions should be validated in future studies by systematically assessing students' perception of their learning experience when using ChatGPT. In the future, there will be further developments of ChatGPT (e.g., communicating via vocal or visual input) that might facilitate showing empathy through imitating an even more humanoid conversation possibly including non-verbal communication aspects. In addition to the forthcoming updates to ChatGPT, previous research has attempted to integrate ChatGPT or other large language models into interfaces designed to facilitate empathic, dialog-based interactions [42]. One advantage is the inclusion of an integrated feedback system [42], which, in our case, had to be manually added into the prompt. Although only a few empathic interactions were observed, students conducted sufficient history takings, supporting the assumption that ChatGPT is not yet unconditionally applicable as a tool for practicing empathic history taking, but can be useful for general history taking. However, it is plausible that embedding the desired ECCS levels directly into the prompt could encourage ChatGPT to provide empathic opportunities aligned with these levels. Furthermore, it is conceivable that specifying these levels, along with a corresponding request, could prompt ChatGPT to provide targeted feedback on students' responses. Nevertheless, ChatGPT might be an adjunct to learning with SPs. Future studies should compare the effectiveness of training empathic history taking using ChatGPT with SP-based training. Communication generally consists of verbal and non-verbal aspects and it has already been shown that empathy is correlated with non-verbal communication in medical encounters but not with verbal communication [4]. An obstacle to consider when using ChatGPT for training history taking is that non-verbal aspects cannot yet be addressed by using ChatGPT. The same applies to minimal implicit verbal expressions (e.g., "hmm"), as these are inserted spontaneously during a conversation that cannot be effectively replicated in interactions with ChatGPT. The study results accentuate this assumption as no empathic responses at level 1 were found. On the other side, there were also no findings for responses at level 6. It should be discussed whether responding at level 6 with shared feelings or expressions constitutes too much self-disclosure, and whether the optimal empathic response might be found closer to the middle of the spectrum with balanced and personalized self-disclosure. A study conducted with patients suffering from chronic pain found that self-disclosure is conducive to patients' own self-disclosure and is perceived as empathic; however, patients also reported concerns that excessive self-disclosure by the physician might lead to an insufficient focus on the patient [43]. Nevertheless, self-disclosure is a connective element in a physician–patient relationship as physicians sharing appropriate self-disclosure are perceived as empathic and facilitate patients' self-disclosure [44]. Since expressing empathy is highly dependent on culture [45], the results may differ across different regions, and, therefore, may not be fully generalizable.

Strengths and Limitations

To our knowledge, this is the first study using ChatGPT as a virtual patient presenting with cardiological diseases in an emergency ward. Without a doubt, empathy plays an important role in the practice of medicine and is stated as crucial for patient satisfaction in the areas of cardiology and emergency medicine [32], which is why we combined both areas. During the analysis, both sides of the interaction between ChatGPT and medical students were rated, resulting in an assessment of ChatGPT's given empathic opportunities and students' empathic reactions. Two main conclusions can be drawn. The first concerns whether ChatGPT can be used for practicing empathic communication from the students' perspective. The second relates to ChatGPT's feasibility in creating an appropriate environment for such practice by providing empathic opportunities. In this context, it is noteworthy that students conducted an average of 24.96 interactions per chat, demonstrating ChatGPT's ability to maintain its persona over an extended conversation. Moreover, the use of their own ChatGPT accounts by students was advantageous, as it enhanced the external validity by providing insight into how students might use ChatGPT as a VP at home. Although most students reported a moderate experience with ChatGPT, its ever-increasing usage worldwide suggests that an increasing number of medical students will likely use ChatGPT at home.

The study has limitations, the most significant being the small sample size. As a result, the findings should be considered preliminary. Future research should include larger and more diverse samples. One shortcoming that simultaneously is an advantage is the use of students' own ChatGPT accounts. For this study, no separate interface was created;

and therefore, students had to use their own accounts. While this improved external validity, it reduced controllability. One obstacle was that students should not see the name of the disease in the applied prompt, which is why the disease had to be embedded in the storyline paraphrased by only stating relevant facts. This obstacle could be circumvented by creating an interface in which the disease could be deposited namely along with all specifications [e.g., 31]. The study results showed no significant differences in verbal empathy expressions between the two temperatures. Therefore, it should be noted that simply adding information about response temperature to the prompt is insufficient. A customized interface would also allow for clearly depositing the response temperature on which ChatGPT should provide its answers. Another limiting factor is that ChatGPT 3.5 is not as controllable as a customized VP, especially in terms of the storyline. During the data collection, ChatGPT occasionally provided trial versions of ChatGPT 4.0 for the first questions in some chats, which was not controllable. The course students attended beforehand focused on general history taking rather than disease-specific history taking. Although empathy is important in the field of cardiology and emergency medicine, applying the chosen diseases simultaneously is a limitation. Future studies should match the diseases with existing course contents. Nevertheless, students were able to transfer their learned skills to new contexts favoring the generalizability of the learned contents. Moreover, the emergency department as a chosen setting for the storyline in the prompt may have hindered students from taking an empathic medical history. Although the selected diseases were not life-threatening, the study should be conducted in general practice setting, where such conditions might be approached more empathically due to their less severe nature. The ECCS is a standardized and valid tool for analyzing empathic communication; however, until now, it has only been used in research involving human participants. Therefore, it should be further investigated whether the ECCS is also a valid tool for analyzing a conversation between a human and a large language model. For instance, empathic opportunities at level 1 could not be coded since this level refers to expressions that are shown within a human conversation but not in a conversation with a large language model.

Conclusion

The study showed initial results for ChatGPT being used as a tool for practicing empathic history taking. Although only a relatively small number of empathic interactions were identified among the many interactions between ChatGPT and medical students, students were still able to take comprehensive histories. Thus, ChatGPT might not yet be perfectly suited for practicing empathic history taking, but it might be effective for training comprehensive history taking. Future research should replicate the study with a larger sample size to draw more robust and reliable conclusions that extend beyond the preliminary findings. Conclusively, the free version of ChatGPT may serve as a useful and cost-effective supplement to SPs for factual learning and, initially, for empathic communication training.

Practice Implications

Using ChatGPT might be a feasible addition for practicing empathic history taking. ChatGPT is a freely accessible tool making it possible for students to practice their history taking skills self-paced and indefinitely in their everyday lives. Teachers need to moderate to ensure an effective learning environment, for example, by providing validated prompts. Taking this idea further, training history taking with ChatGPT might be implemented in a flipped-classroom model. In this sense, students prepare themselves for the course by taking a history with a predefined prompt asynchronously before the course itself. During the synchronous part of the course, the prepared chats are reviewed and teachers should teach the cognitive aspects of empathy during history taking. In the following asynchronous part, ChatGPT can again be used as a tool for practicing the learned aspects of empathic history taking. In a summative exam (e.g., OSCE), the effectiveness of ChatGPT as a training tool could be assessed based on performance improvement. Additionally, training with ChatGPT should be compared to learning with simulated patients.

Author Contribution Conceptualization: Alexandra Aster, Ambra Marx. Methodology: Alexandra Aster. Formal analysis and investigation: Alexandra Aster, Sophia Viktoria Ragaller. Writing — original draft preparation: Alexandra Aster. Writing — review and editing: Alexandra Aster, Sophia Viktoria Ragaller, Tobias Raupach, Ambra Marx. Supervision: Tobias Raupach, Ambra Marx.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics Approval The local ethics committee at University of Bonn approved this study in summer term 2024 (application number: 2024-96-BO).

Consent to Participate Students gave their implied informed consent by completing the study.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

1. Zolnierek KB, Dimatteo MR. Physician communication and patient adherence to treatment: a meta-analysis. Med Care. 2009;47(8):826–34.
2. Neumann M, Wirtz M, Bollschweiler E, Warm M, Wolf J, Pfaff H. Psychometrische evaluation der deutschen version des Messinstrumentes "Consultation and Relational Empathy" (CARE) am Beispiel von Krebspatienten. PPmP - Psychother Psychosom, Med Psychol. 2008;58(01):5–15.
3. Maguire P, Pitceathly C. Key communication skills and how to acquire them. BMJ. 2002;325(7366):697–700.
4. Vogel D, Meyer M, Harendza S. Verbal and non-verbal communication skills including empathy during history taking of undergraduate medical students. BMC Med Educ. 2018;18(1):157.
5. Chung S, Dillon EC, Meehan AE, Nordgren R, Frosch DL. The relationship between primary care physician burnout and patient-reported care experiences: a cross-sectional study. J Gen Intern Med. 2020;35(8):2357–64.
6. Bonvicini KA, Perlin MJ, Bylund CL, Carroll G, Rouse RA, Goldstein MG. Impact of communication training on physician expression of empathy in patient encounters. Patient Educ Couns. 2009;75(1):3–10.
7. Fragkos KC, Crampton PES. The effectiveness of teaching clinical empathy to medical students: a systematic review and meta-analysis of randomized controlled trials. Acad Med. 2020;95(6):947–57.
8. Cuff BMP, Brown SJ, Taylor L, Howat DJ. Empathy: a review of the concept. Emot Rev. 2014;8(2):144–53.
9. Rocco D, Pastore M, Gennaro A, Salvatore S, Cozzolino M, Scorza M. Beyond verbal behavior: an empirical analysis of speech rates in psychotherapy sessions. Front Psychol. 2018;9:978.
10. Bylund CL, Makoul G. Empathic communication and gender in the physician-patient encounter. Patient Educ Couns. 2002;48(3):207–16.
11. Rogers CR. Client-centered therapy. Boston: Houghton-Mifflin; 1951.
12. Weger H, Castle Bell G, Minei EM, Robinson MC. The relative effectiveness of active listening in initial interactions. Int J List. 2014;28(1):13–31.
13. Haley B, Heo S, Wright P, Barone C, Rao Rettiganti M, Anders M. Relationships among active listening, self-awareness, empathy, and patient-centered care in associate and baccalaureate degree nursing students. NursingPlus Open. 2017;3:11–6.
14. Underman K, Hirshfield LE. Detached concern?: emotional socialization in twenty-first century medical education. Soc Sci Med. 2016;160:94–101.
15. Vinson AH, Underman K. Clinical empathy as emotional labor in medical work. Soc Sci Med. 2020;251:112904.
16. Steinmair D, Zervos K, Wong G, Loffler-Stastka H. Importance of communication in medical practice and medical education: an emphasis on empathy and attitudes and their possible influences. World J Psychiatry. 2022;12(2):323–37.
17. Pollak KI, Olsen MK, Yang H, Prose N, Jackson LR 2nd, Pinheiro SO, et al. Effect of a coaching intervention to improve cardiologist communication: a randomized clinical trial. JAMA Intern Med. 2023;183(6):544–53.
18. Kaplonyi J, Bowles KA, Nestel D, Kiegaldie D, Maloney S, Haines T, et al. Understanding the impact of simulated patients on health care learners' communication skills: a systematic review. Med Educ. 2017;51(12):1209–19.
19. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC Acad Med. 1993;68(6):443–51.
20. Long-Bellil LM, Robey KL, Graham CL, Minihan PM, Smeltzer SC, Kahn P. Teaching medical students about disability: the use of standardized patients. Acad Med. 2011;86(9):1163–70.
21. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. Med Educ. 2020;54(9):786–95.
22. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: systematic review and meta-analysis by the digital health education collaboration. J Med Internet Res. 2019;21(7):e14676.
23. Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach. 2009;31(6):477–86.
24. Deci EL, Ryan RM. The general causality orientations scale: self-determination in personality. J Res Pers. 1985;19(2):109–34.
25. Benedict N, Schonder K, McGee J. Promotion of self-directed learning using virtual patient cases. Am J Pharm Educ. 2013;77(7):151.
26. Dankbaar ME, Roozeboom MB, Oprins EA, Rutten F, van Merrienboer JJ, van Saase JL, et al. Preparing residents effectively in emergency skills training with a serious game. Simul Healthc. 2017;12(1):9–16.
27. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol. 2000;25(1):54–67.
28. Fidler BD. Use of a virtual patient simulation program to enhance the physical assessment and medical history taking skills of doctor of pharmacy students. Curr Pharm Teach Learn. 2020;12(7):810–6.
29. Foster A, Chaudhary N, Kim T, Waller JL, Wong J, Borish M, et al. Using virtual patients to teach empathy. Simul Healthc. 2016;11(3):181–9.
30. Or AJ, Sukumar S, Ritchie HE, Sarrafpour B. Using artificial intelligence chatbots to improve patient history taking in dental education (pilot study). J Dent Educ. 2024;88:1988–90.
31. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, Griewatz J, et al. A generative pretrained transformer (GPT)-powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study. JMIR Med Educ. 2024;10:e53961.
32. Sonis JD, Aaronson EL, Lee RY, Philpotts LL, White BA. Emergency department patient experience: a systematic review of the literature. J Patient Exp. 2018;5(2):101–6.

33. Davis J, Van Bulck L, Durieux BN, Lindvall C. The temperature feature of ChatGPT: modifying creativity for clinical research. JMIR Hum Factors. 2024;11:e53559.
34. OpenAI. OpenAI Platform. 2024. Available from: https://platform.openai.com/docs/api-reference/chat. Accessed from 19 Apr 2024 to 19 Jul 2024.
35 Sailer M. Die Wirkung von Gamification auf Motivation und Leistung: Empirische Studien im Kontext manueller Arbeitsprozesse. Wiesbaden: Springer; 2016.
36. Bylund CL, Makoul G. Examining empathy in medical encounters: an observational study using the empathic communication coding system. Health Commun. 2005;18(2):123–40.
37. EACH. Empathic communication coding system. 2023. Available from: https://each.international/reachresources/empathic-communication-coding-system/. Accessed 24 May 2024.
38. Cicchetti D. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess. 1994;6(4):284–90.
39. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96.
40. Maida E, Moccia M, Palladino R, Borriello G, Affinito G, Clerico M, et al. ChatGPT vs. neurologists: a cross-sectional study investigating preference, satisfaction ratings and perceived empathy in responses among people living with multiple sclerosis. J Neurol. 2024;271(7):4057–66.
41. Deladisma AM, Cohen M, Stevens A, Wagner P, Lok B, Bernard T, et al. Do medical students respond empathetically to a virtual patient? Am J Surg. 2007;193(6):756–60.
42. Thesen T, Alilonu NA, Stone S. AI Patient Actor: An open-access generative-AI app for communication training in health professions. Med Sci Educ. 2024. https://doi.org/10.1007/s40670-024-02250-2.
43. Chang HA, Iuliano K, Tackett S, Treisman GJ, Erdek MA, Chisolm MS. Should physicians disclose their own health challenges? Perspectives of patients with chronic pain. J Patient Exp. 2022;9:23743735221128676.
44. Kadji K, Schmid MM. The effect of physician self-disclosure on patient self-disclosure and patient perceptions of the physician. Patient Educ Couns. 2021;104(9):2224–31.
45. Jami PY, Walker DI, Mansouri B. Interaction of empathy and culture: a review. Curr Psychol. 2023;43(4):2965–80.