Assessment and Facilitation of Diagnostic Competences with Simulations in Medical Education

Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.)

am Munich Center of the Learning Sciences

der Ludwig-Maximilians-Universität

München

vorgelegt von

Maximilian Christian Fink

München, 18. Mai 2021

“Learning and acting are interestingly indistinct, learning being a continuous, life-long process resulting from acting in situations.”
(Brown, Collins, & Duguid, 1989, p. 33)

Erstgutachter: Prof. Dr. Martin Fischer, MME (Bern)
Zweitgutachter: Prof. Dr. Frank Fischer
Weitere Mitglieder der Prüfungskommission:
Prof. Dr. Matthias Siebeck, MME (Universität Heidelberg)
PD Dr. Jan Kiesewetter, Dipl.-Psych.
Datum der mündlichen Prüfung: 27. September 2021

Acknowledgments

First of all, I would like to express my gratitude to Prof. Dr. Martin Fischer, Prof. Dr. Frank Fischer, and Prof. Dr. Matthias Siebeck. Their advice, continuous scaffolding, and trust in me helped me to develop and find my way as a researcher. Moreover, COSIMA, the research program they led, supported me in my research. Without the funding and input from this research program, it would not have been possible to conduct the sophisticated studies and analyses presented in this dissertation.

Next, I would like to thank Dr. Anika Radkowitsch**,** Amadeus Pickal, and Sonja Heuser. We shared an office, became good friends, and developed our skills as early career researchers through our interactions. Moreover, I thank Prof. Dr. Ralf Schmidmaier, Dr. Markus Berndt, Dr. Nicole Heitzmann, Dr. Jan Kiesewetter, Dr. Michael Sailer, Dr. Matthias Stadler, Dr. Jan Zottmann, Elisabeth Bauer, Elias Codreanu, Maria Kramer, Stephanie Kron, Angelika Simonsohn, Angelika Wildgans, and all other colleagues from the COSIMA research program and the institute of Medical Education. I am grateful for the invaluable feedback I received from them and the good times we had together at work. I would also like to thank Dr. Victoria Reitmeier dearly for her exemplary commitment and clever creation of instruments. In addition, I thank Johannes Kissel, Ana Maria Semm, and Renke Biallas, who substantially contributed to conducting the studies and preparing the research materials. My full appreciation for the thorough proofreading and editing goes to Keri Hartman.

Besides my colleagues, I would like to wholeheartedly thank Larissa Kaltefleiter for her advice and support. Having her by my side makes me truly happy and enables me to reach my full potential. My grateful thanks are also extended to the family Nystrom in Virginia, who hosted me as an exchange student and helped me to improve my English language skills. Moreover, I thank my friends Johannes Krause, Dominik Wasner, Junus Ergin, Felix Steiner, Bastian Siebenwirth, Philipp Schenke, Christian Höldrich, Theresa Hoffmann, Adrien Gaube, and Sebastian Höpfl for their encouragement through my studies and during the COVID-19 pandemic. Special thanks also go to my parents, Andrea Fink and Christian Fink, my sisters Eva-Maria Gaube and Katharina Fink, and my grandmothers Maria Fink and Theresia Baumgärtl. I thank them for the continuous support that paved my way and the wonderful family life with them.

Executive Summary

Diagnostic competences have always been a focus of medical education, as inaccurate or incorrect diagnoses in medicine can result in serious negative consequences. More recently, the importance of diagnostic competences has also been recognized in other fields of education. Consequently, the assessment and facilitation of diagnostic competences using various methods need to be further explored. The literature indicates that simulation-based methods could be effective for both of these purposes. For assessing diagnostic competences, simulations are an important performance- and competency-based approach. For facilitating diagnostic competences, large effects of simulation-based learning have already been reported. Against this background, this dissertation contributes to the following three central research questions: 1) What differences and similarities emerge in the assessment of diagnostic competences with the two simulation modalities of standardized patients and virtual patients? 2) How are process variables related to diagnostic quality in simulation-based assessment? 3) To what extent can the simulation-based learning of diagnostic competences be facilitated with scaffolding? Two empirical studies were conducted with medical students for this dissertation. The results of the studies are reported in three articles.

Article 1 compared assessment with standardized patients and virtual patients. In a repeated-measures study, participants worked with both of these simulation modalities. It was investigated whether diagnostic accuracy, perceived authenticity, and cognitive load differ or are equivalent in the two simulation modalities. Consistent with research on modality differences, diagnostic accuracy was higher in standardized patients than in virtual patients. Therefore, the use of standardized patients rather than virtual patients in assessment could positively affect examinee scores. Further analysis revealed that standardized patients elicited higher perceived authenticity than virtual patients, in line with prior research. However, perceived authenticity and diagnostic accuracy were minimally associated, also consistent with the literature. Due to this minimal association, it must be critically questioned whether extensive resources should be devoted to increasing perceived authenticity in simulations above a certain necessary level. Cognitive load was equivalent in both simulation modalities, showing that this variable can reach similar levels in digital and non-digital simulations. Moreover, extraneous and intrinsic cognitive load were negatively related to diagnostic accuracy. These relationships underscore that cognitive load

should be controlled and monitored when designing and conducting formative and summative assessments with simulations.

Article 2 examined the contribution of knowledge and the diagnostic process to diagnostic quality. In the underlying study, all participants completed conceptual and strategic knowledge tests and worked with the same virtual patients without additional scaffolding. The diagnostic process with the virtual patients was assessed in terms of the diagnostic activities of hypothesis generation, evidence generation, and evidence evaluation. Diagnostic quality for the virtual patients was measured with a separate diagnostic accuracy score and a comprehensive diagnostic score which included instruments for diagnostic accuracy, treatment selected, diagnostic measures for clarification, and expected findings in a physical examination. Knowledge and diagnostic activities each uniquely explained medium amounts of variance in the comprehensive diagnostic score. These results highlight that the diagnostic process, as operationalized by diagnostic activities, is more than merely an embodiment of knowledge and makes a unique contribution to diagnostic quality. Moreover, these results substantiate the assumption of several frameworks for diagnostic competences that these competences encompass knowledge, the diagnostic process, and diagnostic quality.

Article 3 investigated the effectiveness of reflection phases on learning to diagnose accurately through virtual patients. In an experiment with a pre- and a post-test, two intervention groups learned from virtual patients and additionally completed slightly different types of reflection phases. A control group learned only from the virtual patients. In all virtual patients, diagnostic accuracy was measured as outcome; participants' hypotheses were tracked to investigate the diagnostic process. The analyses showed that reflection phases had no added benefit on learning to diagnose accurately. This finding contradicts the positive effects of reflection phases reported for learning with text-based cases but aligns with prior results for simulation-based learning. Case format and information-processing differences between the two contexts could potentially explain this finding. If this finding is replicated, other types of scaffolding could be more effective for supporting simulation-based learning among medical students with low to medium levels of expertise. In addition, associations between prior knowledge and learning to diagnose accurately from reflection phases were examined. Prior knowledge was not associated with improving one's diagnostic accuracy from the pre- to the post-test. This result indicates that the effectiveness of reflection phases could perhaps depend more on expertise differences than on

prior knowledge differences. Additional analyses showed that participants improved their diagnostic process more during simulation-based learning than during reflection phases.

Overall, this dissertation highlights the potential of simulations for assessing and facilitating diagnostic competences. With respect to modality differences, this dissertation indicates that there are differences between standardized patients and virtual patients in diagnostic accuracy that may affect grading. Therefore, standardized patients should not be directly substituted with virtual patients without sensible contextual adaptations. With respect to the diagnostic process, this dissertation revealed that diagnostic activities have a unique, medium-sized contribution to diagnostic quality. Consequently, diagnostic activities should become an important part of assessing diagnostic competences in simulations. With respect to the effect of scaffolding, this dissertation demonstrated that reflection phases had no added benefit on the acquisition of diagnostic competences in virtual patients. If this finding is replicated, other types of scaffolding could be more effective for learners with low to medium expertise in this particular context. In addition to these points, this dissertation also provides limitations and directions for future research. The limitations discussed include the case specificity of simulations, the measures used and conceptualization of the diagnostic process, and the behavior- and prompt-based methodology employed. Promising directions for future research include case characteristics, adaptivity of instruction and assessment, and the possible transfer of results from assessment to facilitation settings. The dissertation concludes with a summary of the main findings and a personal remark.

Deutsche Zusammenfassung

Diagnosekompetenzen stehen seit jeher im Mittelpunkt der medizinischen Ausbildung, da ungenaue oder falsche Diagnosen in der Medizin zu schwerwiegenden negativen Folgen führen können. In jüngerer Zeit wurde die Bedeutung von Diagnosekompetenzen auch in anderen Bildungsbereichen erkannt. Folglich muss die Beurteilung und Förderung von Diagnosekompetenzen mit verschiedenen Methoden weiter erforscht werden. Die Literatur deutet darauf hin, dass simulationsbasierte Methoden für beide Zwecke effektiv sein könnten. Für die Beurteilung von Diagnosekompetenzen sind Simulationen ein wichtiger leistungs- und kompetenzorientierter Ansatz. Hinsichtlich der Förderung von Diagnosekompetenzen sind bereits große Effekte des simulationsbasierten Lernens bekannt. Vor diesem Hintergrund leistet diese Dissertation einen Beitrag zu den folgenden drei zentralen Forschungsfragen: 1) Welche Unterschiede und Gemeinsamkeiten ergeben sich bei der Beurteilung von Diagnosekompetenzen mit den beiden Simulationsmodalitäten standardisierte Patienten und virtuelle Patienten? 2) Wie hängen Prozessvariablen mit der Diagnosequalität in der simulationsbasierten Beurteilung zusammen? 3) Inwieweit kann simulationsbasiertes Lernen von Diagnosekompetenzen durch Scaffolding gefördert werden? Für diese Dissertation wurden zwei empirische Studien mit Medizinstudierenden durchgeführt. Die Ergebnisse der Studien werden in drei Artikeln berichtet.

Artikel 1 verglich die Beurteilung mit standardisierten Patienten und virtuellen Patienten. In einer Studie mit Messwiederholung absolvierten die Teilnehmer beide Simulationsmodalitäten. Es wurde untersucht, ob sich Diagnoseakkuratheit, wahrgenommene Authentizität und kognitive Belastung in den beiden Simulationsmodalitäten unterscheiden oder gleichwertig sind. In Übereinstimmung mit der Forschung zu Modalitätsunterschieden war die Diagnoseakkuratheit bei standardisierten Patienten höher als bei virtuellen Patienten. Daher könnte sich eine Verwendung von standardisierten Patienten anstelle von virtuellen Patienten bei Beurteilungen positiv auf die Bewertungen der Prüflinge auswirken. Eine weitere Analyse ergab, dass standardisierte Patienten, in Übereinstimmung mit früheren Untersuchungen, eine höhere wahrgenommene Authentizität hervorriefen als virtuelle Patienten. Die wahrgenommene Authentizität und die Diagnoseakkuratheit waren jedoch, ebenfalls in Übereinstimmung mit der Literatur, nur minimal assoziiert. Aufgrund dieses minimalen Zusammenhangs muss kritisch hinterfragt werden, ob umfangreiche Mittel dafür aufgewendet werden sollten, die wahrgenommene Authentizität in Simulationen über ein bestimmtes, notwendiges Maß hinaus zu erhöhen. Die kognitive Belastung

war in beiden Simulationsmodalitäten gleich hoch, was zeigt, dass diese Variable in digitalen und nicht-digitalen Simulationen ähnliche Werte erreichen kann. Darüber hinaus hingen die extrinsische und intrinsische kognitive Belastung negativ mit der Diagnoseakkuratheit zusammen. Diese Zusammenhänge unterstreichen, dass die kognitive Belastung bei der Entwicklung und Durchführung von formativen und summativen Beurteilungen mit Simulationen kontrolliert und überwacht werden sollte.

Artikel 2 untersuchte den Beitrag von Wissen und dem Diagnoseprozess zur Diagnosequalität. In der zugrundeliegenden Studie absolvierten alle Teilnehmer konzeptuelle und strategische Wissenstests und bearbeiteten die gleichen virtuellen Patienten ohne zusätzliches Scaffolding. Der Diagnoseprozess wurde in den virtuellen Patienten durch die diagnostischen Aktivitäten der Hypothesengenerierung, der Evidenzgenerierung und der Evidenzevaluation gemessen. Die Diagnosequalität wurde in den virtuellen Patienten durch eine separate Diagnoseakkuratheits-Skala und eine umfassende Diagnose-Skala erfasst, die Instrumente für die Diagnoseakkuratheit, ausgewählte Therapieverfahren, diagnostische Maßnahmen zur Abklärung, und erwartete Befunde in einer körperlichen Untersuchung beinhaltete. Wissen und diagnostische Aktivitäten erklärten jeweils einen mittleren Anteil der Varianz in der umfassenden Diagnose-Skala. Diese Ergebnisse unterstreichen, dass der Diagnoseprozess, wie er durch diagnostische Aktivitäten operationalisiert wird, mehr als eine bloße Verkörperung von Wissen ist und einen einzigartigen Beitrag zur Diagnosequalität leistet. Darüber hinaus untermauern diese Ergebnisse die Annahme mehrerer Rahmenmodelle zu Diagnosekompetenzen, dass diese Kompetenzen Wissen, den Diagnoseprozess und die Diagnosequalität umfassen.

Artikel 3 untersuchte die Wirksamkeit von Reflexionsphasen auf das Erlernen des akkuraten Diagnostizierens anhand von virtuellen Patienten. In einem Experiment mit Prä- und Posttest lernten zwei Interventionsgruppen von virtuellen Patienten und bearbeiteten zusätzlich leicht unterschiedliche Arten von Reflexionsphasen. Eine Kontrollgruppe lernte nur anhand von virtuellen Patienten. In allen virtuellen Patienten wurde die Diagnoseakkuratheit als Outcome gemessen; zur Untersuchung des Diagnoseprozesses wurden Hypothesen der Teilnehmer aufgezeichnet. Die Analysen zeigten, dass Reflexionsphasen keinen zusätzlichen Nutzen für das Erlernen des akkuraten Diagnostizierens hatten. Dieses Ergebnis steht im Widerspruch zu den positiven Effekten von Reflexionsphasen, die für das Lernen anhand von textbasierten Fällen berichtet wurden, stimmt aber mit früheren Ergebnissen für simulationsbasiertes Lernen überein.

Möglicherweise können Unterschiede im Fallformat und in der Informationsverarbeitung zwischen beiden Kontexten diesen Befund erklären. Wenn dieser Befund repliziert wird, könnten andere Arten von Scaffolding effektiver sein, um simulationsbasiertes Lernen von Medizinstudierenden mit geringem bis mittlerem Expertiseniveau zu fördern. Außerdem wurden Zusammenhänge von Vorwissen mit dem Erlernen des akkuraten Diagnostizierens anhand von Reflexionsphasen untersucht. Hierbei war Vorwissen nicht mit dem individuellen Erwerb von Diagnoseakkuratheit vom Prä- zum Posttest assoziiert. Dieses Ergebnis deutet darauf hin, dass die Wirksamkeit von Reflexionsphasen möglicherweise eher von Expertiseunterschieden als von Vorwissensunterschieden abhängen könnte. Zusätzliche Analysen ergaben, dass die Teilnehmer ihren Diagnoseprozess während des simulationsbasierten Lernens stärker verbesserten als in den Reflexionsphasen.

Insgesamt zeigt diese Dissertation das Potenzial von Simulationen zur Beurteilung und Förderung von Diagnosekompetenzen auf. Hinsichtlich der Modalitätsunterschiede verdeutlicht diese Dissertation, dass es zwischen standardisierten Patienten und virtuellen Patienten Unterschiede in der Diagnoseakkuratheit gibt, die sich auf die Benotung auswirken können. Daher sollten standardisierte Patienten nicht direkt durch virtuelle Patienten ersetzt werden, ohne sinnvolle kontextspezifische Anpassungen vorzunehmen. Hinsichtlich des Diagnoseprozesses verdeutlicht diese Dissertation, dass diagnostische Aktivitäten einen einzigartigen, mittleren Beitrag zur Diagnosequalität leisten. Folglich sollten diagnostische Aktivitäten ein wichtiger Bestandteil der Beurteilung von Diagnosekompetenzen in Simulationen werden. Hinsichtlich des Effekts von Scaffoldings konnte in dieser Arbeit gezeigt werden, dass Reflexionsphasen keinen zusätzlichen Nutzen für den Erwerb von Diagnosekompetenzen anhand von virtuellen Patienten haben. Sollte dieser Befund repliziert werden, könnten andere Arten von Scaffolding für Lernende mit geringer bis mittlerer Expertise in diesem speziellen Kontext effektiver sein. Zusätzlich zu diesen Punkten werden in dieser Dissertation auch Limitationen und vielversprechend Themen für zukünftige Forschung aufgezeigt. Zu den diskutierten Limitationen gehören die Fallspezifität der Simulationen, die verwendeten Maße und Konzeptualisierung des Diagnoseprozesses, sowie die eingesetzte verhaltens- und promptbasierte Methodik. Vielversprechende Themen für zukünftige Forschung beinhalten Fallcharakteristika, die Adaptivität von Instruktion und Beurteilung und die mögliche Übertragung der Ergebnisse von Beurteilungs- auf Lernkontexte. Die Dissertation endet mit einer Zusammenfassung der wichtigsten Ergebnisse und einer persönlichen Bemerkung.

Table of Contents

1. General Introduction

1.1 Relevance, Goals, and Outline of the Dissertation

I begin this section with a discussion of the relevance of this dissertation for medical education and education in general. I then outline the three main goals that the dissertation pursues and refer to important literature within which this work is situated. At the end of the section, I present an outline of the dissertation that succinctly summarizes the enclosed empirical articles and the chapters' contents.

1.1.1 Relevance of the dissertation

The rate of diagnostic errors in medicine lies between ten and twenty percent, depending on the specialty (Berner & Graber, 2008). Diagnostic errors range from small misses with harmless consequences to cases with a severe negative impact. Indeed, it has been estimated that each year about 40,000 patients die due to diagnostic errors in American intensive care units (Winters et al., 2012). Other severe negative impacts of diagnostic errors include (permanent) injuries and disabilities, which result in healthcare and economic costs (Saber Tehrani et al., 2013). Due to the consequences of incorrect diagnoses, diagnostic competences have always been at the heart of medical education (Norman, 2005). Recently, however, the importance of diagnostic competences has also been acknowledged in other fields, including teacher education and social work education (Ghanem, Kollar, Fischer, Lawson, & Pankofer, 2016; Loibl, Leuders, & Dörfler, 2020).

Simulation-based learning is becoming increasingly popular in higher education, and there are at least two reasons for this development. First, some fields, including medical education, have widely adopted an approach known as outcome-based or competency-based education (Morcke, Dornan, & Eika, 2013). In such an approach, many of the outcomes and competences graduates should obtain can be conveyed well by simulations (Scalese, Obeso, & Issenberg, 2008). Second, simulation-based learning is becoming increasingly accessible (Chernikova, Heitzmann, Stadler et al., 2020; Gegenfurtner, Quesada-Pallarès, & Knogler, 2014), likely due to reduced costs, improved infrastructure, and enhanced training for facilitators. At the same time, the use of simulation-based assessments is growing. This type of assessment can measure the outcomes described in competency frameworks[1] with high face validity due to its focus on assessees'

[1] Please see two examples of competency frameworks in the domain of medical education by Frank and Danoff (2007) and MFT (Medizinischer Fakultätentag der Bundesrepublik Deutschland e.V.) (2015).

observable performance (Miller, 1990). Moreover, this type of assessment is conceptually closely aligned to simulation-based learning and can be used for both formative and summative purposes (Boulet, 2008; van der Vleuten & Schuwirth, 2019).

In light of these developments, how diagnostic competences can be assessed and facilitated through new simulation-based methods is an important issue.

1.1.2 Goals of the dissertation

I pursue three goals with this dissertation. First, I want to improve knowledge on the assessment of diagnostic competences with different simulation modalities. I will focus this inquiry mainly on a comparison of standardized patients and virtual patients. In medical education, these two types of simulation modalities have only been contrasted directly in a few studies (Edelstein, Reid, Usatine, & Wilkes, 2000; Guagnano, Merlitti, Manigrasso, Pace-Palitti, & Sensi, 2002; Hawkins et al., 2004). Second, I want to add to research on the relationships among process variables relevant in diagnosing. Although various theoretical frameworks for diagnostic competences describe important process variables, these process variables' relationships with and exact contribution to diagnostic quality remain largely unquantified (Heitzmann, Fischer, & Fischer, 2017). Third, I want to contribute to research on the effect of scaffolding in simulation-based learning. While it is well-established that simulation-based learning has a large effect on acquiring diagnostic competences (Chernikova, Heitzmann, Stadler et al., 2020), the effect of various types of additional scaffolding in simulation-based learning remains unclear. This is particularly the case for reflection phases, which have mainly been investigated in the context of learning from text-based cases (Mamede & Schmidt, 2017), but rarely in simulation-based learning. Next, I provide an outline of the dissertation.

1.1.3 Outline of the dissertation

This dissertation is subdivided into five chapters. Chapter 1 provides the theoretical background. It addresses the topics of diagnostic competences and the diagnostic process, simulations and simulation-based education, assessment of diagnostic competences in the simulation modalities of standardized patients and virtual patients, and facilitating diagnostic competences with scaffolding. Chapters 2 to 4 contain the three empirical articles. Chapter 2 investigates the assessment of diagnostic competences with standardized patients and virtual patients. Chapter 3 analyzes the interplay of the diagnostic process, knowledge and diagnostic performance in diagnosing virtual patients. Chapter 4 examines the effect of scaffolding in the form of reflection phases on learning to diagnose accurately through virtual patients. Chapter 5 summarizes the three empirical articles' findings and discusses this dissertation's contributions to answering the central research questions (see Section 1.6). Moreover, it discusses limitations and directions for future research on a more global level than the individual articles.

1.2 Diagnostic Competences

I start this section by providing definitions of terms related to diagnostic competences. I then describe essential frameworks for diagnostic competences and discuss relationships between diagnostic process variables and diagnostic quality. Finally, I conclude the section with a summary of conceptualizations of diagnostic quality.

1.2.1 Definitions and terms

According to Blömeke, Gustafsson, and Shavelson (2015), competences should be defined on a continuum from latent cognitive and affective dispositions to observable performance in real situations. Situation-specific skills lie at the middle of this continuum and play a mediating role between the two poles. This definition has two significant implications. First, competences are, at least to some extent, trainable and should be systematically fostered in higher education. Second, competences can be relatively objectively measured with various suitable methods, such as simulations, and should play a role in selecting persons to undergo training and awarding licenses for professional practice (Blömeke et al., 2015). I will now turn to diagnosing. A first domain-general notion of diagnosing comes from the literature on scientific problem-solving. Klahr and Simon (2001) define problem-solving as a process in which an individual tries to reach a goal by carrying out operations. Only certain operations with specific characteristics are permissible at certain stages of problem-solving and must be determined by the individual in a search process (Klahr & Simon, 2001). This definition of problem-solving applies to many different scientific problems and domains. While it provides a first glimpse into major characteristics of diagnosing, it is too general to capture the full complexity of this process. Therefore, a definition of diagnosing that is more suitable for the field of medical education but also other educational contexts is required. Heitzmann et al. (2019) define diagnosing as gathering and evaluating information for professional decision-making. This definition highlights several points. First, diagnosing can be regarded as a decision-making or problem-solving process. Second, the diagnostician must have received training in this competence. Third, the diagnostician strives to decrease uncertainty in their hypothesis via professional means until a conclusion can be reached. In this dissertation, I operationalize competences following Blömeke et al. (2015), consider diagnosing a problem-solving process in accordance with Klahr and Simon (2001), and follow the definition of diagnosing as a professional decision-making process by Heitzmann et al. (2019).

1.2.2 Frameworks for diagnostic competences and the diagnostic process

There are a plethora of frameworks for diagnostic competences in medical education that cannot be described in their entirety here (Elstein, 2009; Ericsson, 2007; Norman, 2005). Nevertheless, three major types of frameworks focusing on the diagnostic process can be distinguished: the knowledge-centered illness script theory (Charlin, Boshuizen, Custers, & Feltovich, 2007), the cognitive-oriented dual-process theory (Kahneman, 2011), and theories that operationalize diagnosing as problem-solving process (Heitzmann et al., 2019).

Illness script theory (Schmidt, Norman, & Boshuizen, 1990) assumes that a variety of knowledge structures, such as biomedical and clinical knowledge, develop during medical training. Biomedical knowledge refers to knowledge from basic science fields relevant to medicine, such as physiology (Boshuizen & Schmidt, 1992). Clinical knowledge, on the other hand, consists of knowledge about clinical symptoms and procedures (Patel, Evans, & Groen, 1989). In the later stages of medical training, medical students and physicians build large knowledge networks called *illness scripts* that are enriched with experiences from patients (Schmidt & Rikers, 2007). These illness scripts focus on the underlying conditions and symptoms of diagnoses and only contain small amounts of encapsulated knowledge. Illness script theory postulates and has empirically demonstrated that diagnostic performance frequently relies on a rapid, automatic process of pattern recognition that draws primarily on illness scripts rather than other types of knowledge (Charlin et al., 2007).

According to dual-process theory (Kahneman, 2011), there are two different systems used in diagnosing. *System I* comprises fast, heuristic, and unconscious processes. *System II* encompasses slow, reflective, and conscious processes. In medical education, numerous dual-processing models with heterogeneous assumptions have emerged (Evans, 2008). Most of these models agree that both systems can be used together or at least affect each other during diagnosing (Croskerry, 2009; Eva, 2004; Evans, 2008). Moreover, empirical research has shown that both systems can be used simultaneously and lead to similar misdiagnosis rates (Eva, Hatala, LeBlanc, & Brooks, 2007; Monteiro & Norman, 2013). While dual-process theories assign unique characteristics to two different cognitive systems, more specific diagnostic processes and their associations with diagnostic quality are not at the heart of this theory.

Based on the aforementioned notion of problem-solving, Heitzmann et al. (2019) presented a conceptual framework for the simulation-based learning of diagnostic competences. This framework assumes that success in diagnosing is a result of individual prerequisites, such as prior knowledge. Knowledge is operationalized in this theory as *conceptual* and *strategic* knowledge (Stark, Kopp, & Fischer, 2011). Conceptual knowledge is declarative knowledge about concepts and their interconnections. Strategic knowledge refers to knowledge on making decisions and how to proceed in specific clinical situations (Stark et al., 2011). Empirical studies have shown that these types of knowledge are interconnected and associated with performance in diagnosing (Schmidmaier et al., 2013; Stark et al., 2011). Moreover, this framework posits that the diagnostic process comprises eight diagnostic activities that are potentially related to learning and diagnostic quality. Diagnostic activities are relatively general epistemic practices (Kelly, 2008) that are taught and evaluated by a community of practice and can be assessed across different domains. Table 1 provides definitions of all eight diagnostic activities. This dissertation, however, will primarily focus on the diagnostic activities *hypothesis generation*, *evidence generation,* and *evidence evaluation*. There is already some support for relationships between these three diagnostic activities and diagnostic quality. Several correlational studies support bivariate relationships between diagnostic quality and hypothesis generation (Coderre, Wright, & McLaughlin, 2010; LeBlanc, Brooks, & Norman, 2002; LeBlanc, Norman, & Brooks, 2001), as well as evidence generation (Stillman et al., 1986; Woolliscroft et al., 1989). Moreover, an association between evidence evaluation and diagnostic quality is implied in theorizing on the script concordance test (Charlin, Roy, Brailovsky, Goulet, & van der Vleuten, 2000). In addition, two studies from teacher education and medical education investigated the contribution of multiple diagnostic processes to diagnostic quality (Groves, O'Rourke, & Alexander, 2003; Kramer, Förtsch, Seidel, & Neuhaus, 2021). These two studies demonstrated that the diagnostic process predicted a substantial amount of variance in diagnostic quality. Despite these findings, the exact contribution of knowledge and the diagnostic process to diagnostic quality still requires further research (Heitzmann et al., 2017).

Table 1

Overview of Diagnostic Activities

Diagnostic activity	Definition
Identifying problems	The diagnostician recognizes a conspicuous (problematic) phenomenon.
Questioning	The diagnostician asks themselves questions that determine the direction of the subsequent process.
Hypothesis generation	The diagnostician names or indicates initial or preliminary diagnoses.
Constructing artefacts	The diagnostician constructs (physically existing) products that can be used repeatedly to gather diagnostic information.
Evidence generation	The diagnostician gathers additional information.
Evidence evaluation	The diagnostician determines the reliability and interprets the meaning of one or multiple pieces of information.
Drawing conclusions	The diagnostician decides on their final diagnosis.
Communicating the process/results	The diagnostician passes on information from the diagnostic process to a recipient in written or oral form.

The described framework by Heitzmann et al. (2019) and two frameworks for diagnostic competences used in teacher education (Herppich et al., 2018; Loibl et al., 2020) operationalize diagnostic competences in accordance with Blömeke et al. (2015). Therefore, these frameworks conceptualize diagnostic competences as consisting of knowledge, the diagnostic process, and diagnostic quality. Empirical results that support this broad conceptualization of diagnostic competences are so far lacking, and more research on this topic is needed.

1.2.3 Diagnostic quality

In the framework by Heitzmann et al. (2019), diagnostic performance in case-based learning and assessment is defined as *diagnostic quality* and consists of *diagnostic accuracy* and *diagnostic efficiency*. Diagnostic accuracy is the correspondence of a diagnosis with a correct expert solution. Diagnostic efficiency relates the achieved diagnostic accuracy to the time, cost, or damage created (Braun et al., 2017; Radkowitsch, Fischer, Schmidmaier, & Fischer, 2020; Towle, 1998). However, a literature review showed that there are many additional operationalizations of diagnostic quality in medical education (Daniel et al., 2019). Following these operationalizations, diagnostic quality can be determined based on justifications, case summaries, alternative diagnoses, treatment decisions, and requested follow-up tests (Graber, Tompkins, & Holland, 2009; Nendaz & Bordage, 2002; Radkowitsch et al., 2020; Stojan, Daniel, Morgan, Whitman, & Gruppen, 2017; Williams et al., 2014). Consequently, it can be concluded that both more specific and broader conceptualizations of diagnostic quality are possible and applied in the literature. This dissertation draws in Article 1 and Article 3 on a more specific conceptualization of diagnostic quality as diagnostic accuracy. Article 2, however, additionally uses a broader conceptualization of diagnostic quality, which is called *diagnostic success* in this text, by measuring it with a *comprehensive diagnostic score*. Next, I discuss the topic of simulations and simulation-based education.

1.3 Simulations and Simulation-Based Education

At the beginning of this section, I provide definitions of terms related to simulations. Afterward, I sketch out the differences between simulations by providing an overview of different simulation modalities, simulation properties, and design features. At the end of the section, I synthesize theories on simulation-based education, including the concepts of scaffolding, authenticity, and cognitive load. Moreover, this last part further elaborates on the conceptual framework for simulation-based learning of diagnostic competences by Heitzmann et al. (2019) that was partly introduced in Section 1.2.

1.3.1 Definitions and terms

Simulations have been used for decades in various domains, including medicine, management, the military, the aviation industry, and education (Hallinger & Wang, 2020; Salas, Rosen, Held, & Weissmuller, 2009). Different definitions of simulations concur that simulations are (partial) representations of a situation, task, or system that can be manipulated by a participant (Heitzmann et al., 2019; Jones, Passos-Neto, & Braghiroli, 2015; Kaufman & Ireland, 2016). The term *simulation* is sometimes differentiated from the term *simulator*. In computer science, hardware and software are distinguished from one another. Following such a distinction, simulators are devices and environments that are used for carrying out and running simulations, which are particular scenarios and represented situations (Khan, Tolhurst-Cleaver, White, &

Simpson, 2011). I will not pursue this distinction further in this dissertation. Thus, when I use the term simulations in this text, both or only one of the two components (simulation and simulator) may be meant depending on the context.

1.3.2 Differences between simulations

Differences between simulations used in education can be described with several terms. On the one hand, simulations can be distinguished by their modality. Simulation modalities frequently used in education include live simulations, role-play simulations, and digital simulations (Fink, Radkowitsch et al., 2021). In live simulations, (professional) actors and trained confederates who are briefed about the scenario interact with participants (Barrows & Abrahamson, 1964). Role-play simulations are a modality in which participants act out a case after briefly preparing for the session (Simpson, 1985). In digital simulations, participants learn through a computer simulation that includes digital agents or environments (de Jong, 1991). In addition to these three modalities, there are extended-virtuality simulations in which participants immerse themselves in virtual or augmented environments (Kaplan et al., 2020). On the other hand, simulations can be distinguished by their fidelity (Hamstra, Brydges, Hatala, Zendejas, & Cook, 2014; Maran & Glavin, 2003), where fidelity is defined as the degree of realism a simulation possesses. The physical resemblance between the simulation and the real-life task is called physical fidelity. The alignment between the task in the simulation and the real-life task is called functional fidelity (Hamstra et al., 2014; Maran & Glavin, 2003). In addition, simulations can be differentiated based on the degree of technology use (Jones et al., 2015). On a continuum of technology use, there are simulations with low technology, such as simple mannequins, simulations with medium technology, like screen-based computer simulations, and simulations using high technology, such as virtual reality simulations (Jones et al., 2015).

Differences between simulations can also be characterized by the design features they include. First of all, it should be noted that numerous lists of design features exist (Chernikova, Heitzmann, Stadler et al., 2020; Crawford, 1966; Davidsson & Verhagen, 2013; Gaba, 2004; Gegenfurtner et al., 2014; Huwendiek et al., 2009; Issenberg, McGaghie, Petrusa, Gordon, & Scalese, 2005; Kim et al., 2006; Meller, 1997). The most typical design features reported in the literature are the following: duration, case difficulty, type of user interface, degree of interactivity, paths through the simulation, underlying model of the simulation, extent to which reality is represented, number of participants, outcome measured, type of assessment, and type of scaffolding. One other design feature that should be mentioned is the case format (Huwendiek et al., 2009; Kiesewetter et al., 2020). In simulations, the serial cue format or branching format are typically used. In the serial cue format, participants acquire different pieces of information step-by-step, in a largely linear way. In the branching format, participants acquire information by browsing freely through a large number of paths and end up at different end points within the simulation. Both of these case formats can be contrasted to the whole case format used in regular instruction with text-based cases, in which all information about a case is provided to the learner all at once (Huwendiek et al., 2009; Kiesewetter et al., 2020). These design features highlight that even simulations with the same simulation modality, fidelity, or degree of technology use can be highly heterogeneous. Moreover, simulations contain many design features that could be potential moderators of performance in learning and assessment.

1.3.3 Simulation-based education

Two theories can be considered more or less as direct precursors of theories on simulation-based education: problem-based learning and case-based reasoning. Problem-based learning refers to instruction in which small groups of students learn via authentic exercises or tasks (Barrows, 1996; Wood, 2003). Moreover, it typically includes a phase of independent study and group discussion as well as support by a facilitator. Case-based reasoning (Kolodner, 1992) theorizes that throughout their lives, persons acquire a set of situations that are used in problem-solving and to acquire knowledge and skills. Both theories highlight the role of the context in education. Problem-based learning provides principles and findings on effective instruction that can be transferred to simulation-based learning. Case-based reasoning outlines in more detail the cognitive mechanisms that affect problem-solving and learning in simulations that consist of cases.

Another theory that is relevant for simulation-based education is the cognitive theory of multimedia learning (Mayer & Moreno, 1998). According to this theory, separate verbal and visual channels process sounds and images in sensory memory. In working memory, the cognitive processes of selecting, organizing, and integration are employed to foster learning. Attention processes and the retrieval of content from long-term memory play a crucial part in these cognitive processes (Mayer & Moreno, 1998). This theory seems particularly suitable for learning contexts in which audiovisual media and texts are used. Moreover, the theory highlights that simulations that contain only text might be processed differently from simulations that also contain audiovisual materials.

Moreover, several theories propose specific concepts that might play a role in simulation-based learning and assessment. I now focus on the three particularly important concepts of scaffolding, cognitive load, and authenticity.

Wood, Bruner, and Ross (1976) define scaffolding as supporting learners in a goal-oriented way either by reducing the complexity of a task or by helping to regulate their learning process. Scaffolding can include human support and/or support from technological systems and can either be relatively static or relatively adaptive (Belland, 2014). With respect to the tasks and materials used, scaffolding theory posits, in line with the idea of the *zone of proximal development* (Vygotsky, 1978), that learners acquire knowledge and skills most effectively when solving tasks that are challenging but can be performed well with external support. In terms of the scope and duration of external support, it should be noted that external support may be permanent or be removed in the course of the learning process through fading (Tabak & Kyza, 2018). The scaffolding concept is vital for simulation-based learning because it emphasizes that learners should encounter challenging materials and receive adequate instructional support to promote learning. Moreover, there is empirical evidence that the inclusion of scaffolding may have an additional benefit in simulation-based learning (see Section 1.5).

According to cognitive load theory, working memory capacity is strictly limited, and high mental effort can lead to performance issues in problem-solving (Sweller, van Merriënboer, & Paas, 1998). Cognitive load encompasses three different facets: intrinsic load, extraneous load, and germane load. Intrinsic load is mainly determined by the difficulty and complexity of the material or task at hand. Extraneous load is the cognitive load created by the learning environment. Germane cognitive load is the cognitive load that results from mental processes such as schema abstraction that may bolster learning (Sweller et al., 1998). Numerous empirical studies have supported the assumption that primarily intrinsic and extraneous cognitive load are negatively correlated with performance in problem-solving and diagnosing (Sweller, van Merriënboer, & Paas, 2019; Young, van Merriënboer, Durning, & ten Cate, 2014). Due to the possible effects of cognitive load, simulation-based learning and assessment environments must be designed optimally in the following way. Unnecessary demands resulting from the context that increase extraneous cognitive load should be reduced, and demands productive for learning that raise germane cognitive load should be promoted (Young et al., 2014).

Theories on authenticity are also crucial for simulation-based learning and assessment. There are various notions of authenticity, including the concepts of perceived authenticity (Schubert, Friedman, & Regenbrecht, 2001), thick authenticity (Shaffer et al., 2009), and fidelity (Hamstra et al., 2014; Maran & Glavin, 2003). These concepts differ in their definitions and measurement of authenticity. In this dissertation, I mainly focus on the concept of perceived authenticity, which encompasses the three facets realness, spatial presence, and involvement (Schubert et al., 2001). Realness refers to the extent to which a situation or task resembles the actual, simulated situation or task. Spatial presence is the physical immersion experienced in the simulated situation (Schubert et al., 2001). Involvement refers to a feeling of cognitive immersion and a sense of relevancy (Hofer, 2016). Findings on the association between authenticity and performance are mixed. On the one hand, there is meta-analytic evidence that higher authenticity in simulation-based learning environments is associated with increased acquisition of complex skills (Chernikova, Heitzmann, Stadler et al., 2020). On the other hand, a literature review from medical education that investigated the relationship between fidelity and learning discovered only minimal performance differences in learning between simulations with high and low fidelity (Norman, Dore, & Grierson, 2012). We can conclude from this brief look at the literature that different operationalizations of authenticity may potentially be associated with different findings on its relationship with performance.

I now describe a comprehensive conceptual framework for the simulation-based learning of diagnostic competences in teacher education and medical education (Heitzmann et al., 2019). Both the dissertation overall and the included empirical articles draw on this framework. The framework posits that simulations act as “approximations of practice” (Grossman et al., 2009, p. 2058) and provide authentic and relevant problems. Participants then acquire knowledge and competences through problem-solving in simulations that include scaffolding. In general, the framework assumes that individual learning prerequisites, instructional support, the simulation context, and processes in simulation-based learning environments affect the acquisition of diagnostic competences. Figure 1 provides a graphical depiction of this framework.

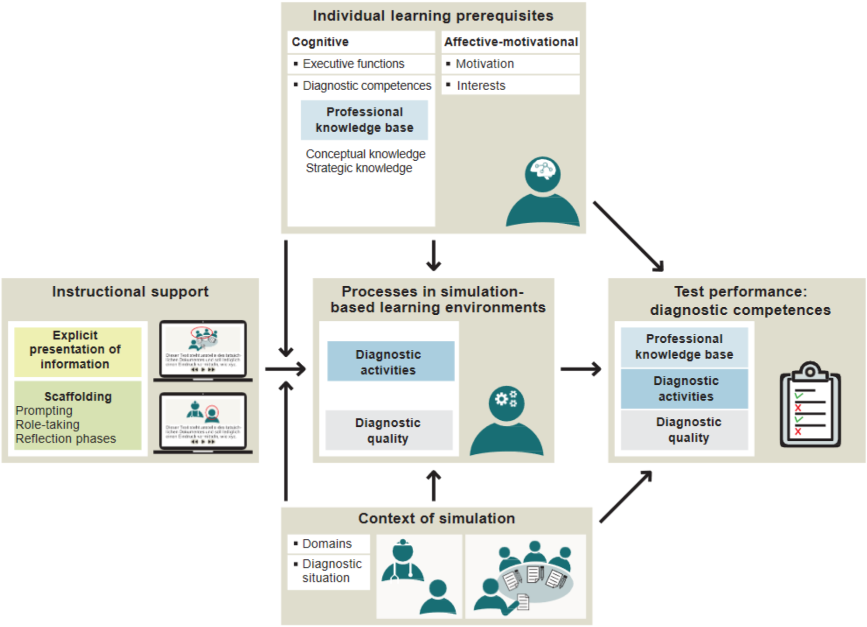


Figure 1. The COSIMA framework model (adapted from COSIMA research unit, 2021). © COSIMA research unit. This is an open-access figure distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The complete bibliographic information, a link to the original publication as well as this copyright and license information must be included.

Next, I summarize the framework's assumptions regarding instructional support, relationships with individual prerequisites, the conceptualization of the diagnostic process, and the operationalization of diagnostic competences. Concerning instructional support, the framework expects the explicit presentation of information and scaffolding to improve the acquisition of diagnostic competences. The explicit presentation of information encompasses conveying knowledge before, during, or after the participant takes part in the simulation. Moreover, the framework distinguishes between three different types of scaffolding: prompting, role-taking, and reflection phases. Prompts are notifications supplied to the participant during problem-solving. Role-taking refers to the three possible roles a participant can take in diagnosing simulations: an active diagnostician, a person whose features are diagnosed, or a passive observer. In reflection phases, participants answer a set of pre-defined questions to improve engagement, recapitulate their diagnoses, and plan and monitor the diagnostic process (see more in Section 1.5). Regarding individual prerequisites, the framework stresses the importance of professional knowledge. It differentiates between strategic and conceptual knowledge, which have already been described (see Section 1.2). With respect to the diagnostic process, the framework assumes that the eight diagnostic activities defined in Table 1 (see Section 1.2) are related to learning and diagnostic quality. The framework stresses that the quality and sequence of diagnostic activities, not their sheer quantity, could be a suitable performance indicator for the diagnostic process. In the framework, diagnostic competences encompass professional knowledge, diagnostic activities, and diagnostic quality. Diagnostic quality is in turn operationalized in terms of diagnostic accuracy and diagnostic efficiency (Heitzmann et al., 2019). For more details and a comparison with other conceptualizations of diagnostic competences, see Section 1.2. In the next section, I discuss the assessment of diagnostic competences with simulations.

1.4 Assessing Diagnostic Competences with Simulations

I begin this section with a discussion of the relevance of performance-based assessment in medical education. I then describe the two simulation types, standardized patients and virtual patients, frequently used for performance-based assessment in medical education. Finally, I synthesize the literature on differences between standardized patients and virtual patients regarding the variables perceived authenticity, cognitive load, and diagnostic accuracy.

1.4.1 Performance-based assessment in medical education

Performance-based assessment places an “emphasis on testing complex, ‘higher order’ knowledge and skills in the real-world context” (Swanson, Norman, & Linn, 1995, p. 5). Due to this focus on outcomes and competences, performance-based assessment is becoming increasingly common in medical education (Boulet & Durning, 2019; Swanson & Roberts, 2016). Performance-based assessment can be used for summative and formative purposes. In summative assessment, one or more assessors judge a person’s knowledge or competences at a specific time point for evaluation purposes (Taras, 2005). In formative assessment, one or more assessors judge a person’s knowledge or competences repeatedly and also inform the person of their evaluations in order to promote learning (Taras, 2005). There are many benefits of conducting performance-based assessment. One of the main benefits is encapsulated in the saying “assessment drives learning” (Wormald, Schoeman, Somasunderam, & Penn, 2009, p. 199). This saying alludes to the positive learning effects that assessment may provide, amongst other things, through practice testing (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013) and obtaining feedback (Hattie & Timperley, 2007). Another benefit is that assessment allows for the monitoring and goal-oriented improvement of the educational system. This benefit has been impressively demonstrated in K-12 education by the Programme for International Student Assessment (PISA), which led to substantial education reforms in many countries (Schleicher & Zoido, 2016). A third benefit is that performance-based assessment can protect the community from professionals who do not fulfill the minimum requirements for safe practice. This effect can be achieved through performance-based assessment by selecting suitable candidates, conducting strict licensure examinations, and requiring continuous recertification during practice (Boulet & Durning, 2019). Various assessment methods are used in performance-based assessment. These assessment methods include multiple-choice tests that provide contextual information, virtual patients, standardized patients, the Objective Structured Clinical Examination, oral examinations, and workplace-based observation (Boulet & Durning, 2019; Swanson et al., 1995). In the following section, I elaborate on the two assessment methods relevant for this dissertation, standardized patients and virtual patients, which can be considered different simulation modalities but nevertheless have certain commonalities.

1.4.2 Standardized patients and virtual patients

Standardized patients and virtual patients are two popular assessment methods in medical education. I now define these two assessment methods, briefly describe their inception, and discuss their use in formative and summative assessment.

For standardized patients, laypeople, actors, or medical students are taught by acting coaches and physicians to convincingly and consistently display the signs and symptoms of a patient (van der Vleuten & Swanson, 1990). Standardized patients were first conceived by Barrows and Abrahamson (1964) to practice history-taking and physical examinations. In the 1980s, standardized patients were integrated into multiple-station evaluations called Objective Structured Clinical Examination (Harden, Stevenson, Wilson Downie, & Wilson, 1975). In the Objective Structured Clinical Examination, participants move from station to station and are typically judged by multiple raters at each station. The judgment process for standardized patients and the Objective Structured Clinical Examination has been standardized through the implementation of checklists and rating-scale forms that aim to ensure reliability and objectivity (Cohen, Colliver, Marcy, Fried, & Swartz, 1996; van Thiel, Kraan, & van der Vleuten, 1991). The use of standardized patients in summative and formative assessments varies across the globe. Today, standardized patients are a crucial part of licensure examinations in several countries, including the US, UK, and Canada (Swanson & Roberts, 2016). In Germany, the introduction of the Objective Structured Clinical Examination containing standardized patients to the second part of the national licensure examination has only recently been agreed upon and is currently in the initial stages of implementation (Jünger, 2018).

Virtual patients are computer simulations of important clinical situations and tasks, such as history-taking (Cook, Erwin, & Triola, 2010). The emergence of virtual patients goes back to the development of text-based simulations. Among the most prominent text-based simulations were patient management problems (McCarthy & Gonnella, 1967; McGuire & Babbott, 1967) and modified essay questions (Knox, 1975). Text-based simulations contextualize the clinical problems of a case and allow participants to either branch through a scenario or gather data linearly. With the advent of personal computing, virtual patients were created on the basis of these text-based simulations. Virtual patients provide greater possibilities for interaction and integrate audiovisual media that text-based simulations cannot (Clyman, Melnick, & Clauser, 1999; Ellaway, Candler, Greene, & Smothers, 2006). Currently, virtual patients are used widely in summative and formative assessment. Regarding licensing, virtual patients have been used in the US since 1999 (Dillon, Boulet, Hawkins, & Swanson, 2004). In Germany, however, virtual patients are not used in licensing in medicine, and concrete steps towards their inclusion had not been planned until recently (Jünger, 2018).

Categorized by simulation modality, standardized patients in the domain of medical education largely fall under live simulations and virtual patients under digital simulations (see Section 1.3).

1.4.3 Differences between standardized patients and virtual patients

Only a few prior studies in medical education have directly compared standardized patients and virtual patients (Edelstein et al., 2000; Guagnano et al., 2002; Hawkins et al., 2004). Such direct comparisons may help to highlight the differences between assessment methods and may quantify their agreement in measuring performance. Next, I outline three different aspects along which standardized patients and virtual patients should be compared with each other and may differ.

One aspect that seems worthwhile to compare for the two assessment methods is perceived authenticity (see Section 1.3). Several studies have demonstrated the exceptionally high perceived authenticity of standardized patients (Luctkar-Flude, Wilson-Keates, & Larocque, 2012; Rethans, Sturmans, Drop, & van der Vleuten, 1991). It has also been shown that virtual patients can obtain high perceived authenticity scores (Friedman, France, & Drossman, 1991). However, even modern virtual patients still lack some of the features that make standardized patients particularly authentic (e.g., being in a physical environment during the assessment). Since no empirical studies have so far directly compared the authenticity of the two assessment methods, it is still an open question whether standardized patients are indeed substantially more authentic than virtual patients.

Another aspect that should be examined in assessment with standardized patients and virtual patients is cognitive load (see Section 1.3). In medical education, cognitive load has so far only been contrasted in learning but not in assessment environments. Haji et al. (2016) compared a group of medical students learning with a complex simulation to a group learning with a simple simulation. In both groups, the simulations used were part task trainers enabling repeated practice of a physical skill. Cognitive load was lower in the simple simulation than in the complex simulation (Haji et al., 2016). Dankbaar et al. (2016) contrasted learning with a text-based simulation to learning with a more authentic simulation game. Cognitive load was higher in the simulation game than in the text-based simulation (Dankbaar et al., 2016). These findings suggest that cognitive load in standardized patients and virtual patients could depend on task complexity as well as authenticity. In addition to these findings, it can also be argued, based on theory, that cognitive load mainly depends on the sources of the three facets intrinsic, germane, and extraneous cognitive load described in Section 1.3. Following this argument, cognitive load should be similar in standardized patients and virtual patients if the same cases are used and the instructional design poses comparable demands. However, it is an open question whether the design features of standardized patients and virtual patients, such as the type of user input, themselves create different levels of cognitive load even when a relative level of standardization is attempted.

A final important aspect is how standardized patients and virtual patients measure diagnostic competences. If both assessment methods measure the same construct, performance in the two assessment methods should be interrelated. In line with this idea, the existing studies comparing standardized patients and virtual patients in medical education have reported correlations between diagnostic accuracy in both assessment methods (Edelstein et al., 2000; Guagnano et al., 2002; Hawkins et al., 2004). Contrary to the results described above, it could also be argued that standardized patients and virtual patients possess different design features and characteristics (see Section 1.3) which could, in turn, result in lower performance in one of the two simulation modalities. Moreover, it should be highlighted that the literature has so far not investigated whether standardized patients and virtual patients lead to the same educational decisions and grades. This research question is crucial because assessment with virtual patients may be used as a substitute for assessment with standardized patients. This is especially the case because training and implementing assessment with standardized patients involves a great deal of person-hours and ongoing costs (Ziv, 2009). In contrast, assessment with virtual patients only requires the one-time expenditure of high costs and development efforts (Huang, Reynolds, & Candler, 2007). In the following section, I shift focus from assessing diagnostic competences to facilitating diagnostic competences.

1.5 Facilitating Diagnostic Competences with Simulations

In this section, I complement the theories of simulation-based learning and scaffolding described in Section 1.3 with a summary of empirical results on the effects of simulation-based learning, as well as scaffolding and reflection in simulation-based learning.

1.5.1 The effects of simulation-based learning

Three meta-analyses that synthesized studies from different domains, including the health sciences, provide a detailed picture of the effects of simulation-based learning (Chernikova, Heitzmann, Stadler et al., 2020; Cook et al., 2010; Cook et al., 2012). In comparison to no intervention, large positive effects of simulation-based learning have been shown for knowledge, skills, clinical reasoning, and complex cognitive skills (Chernikova, Heitzmann, Stadler et al., 2020; Cook et al., 2010). Compared to other types of instruction, the effects of simulation-based learning seem to depend primarily on the outcome examined. Simulation-based learning is particularly effective for physical procedures and competences like diagnosing (Chernikova, Heitzmann, Stadler et al., 2020; Cook et al., 2012). However, when it comes to conveying knowledge, the advantage over conventional types of instruction, such as lectures, seems relatively small (Cook et al., 2010; Cook et al., 2012). Additional support for the effectiveness of simulation-based learning comes from studies investigating different types of problem-based learning. This is the case because problem-based learning is, like simulation-based learning, a highly contextualized form of learning that frequently takes place using cases (see Section 1.3). Meta-analyses on the problem-based learning of skills and diagnostic competences have reported medium effects in comparison to conventional types of instruction (Chernikova, Heitzmann, Fink et al., 2020; Dochy, Segers, Van den Bossche, & Gijbels, 2003).

1.5.2 The effects of scaffolding in simulation-based learning

A meta-analysis by Cook et al. (2013) more closely explored the effects of technology-enhanced simulation-based learning with instructional support in the health sciences. This meta-analysis was not based on a clear theory of scaffolding and conflated diverse characteristics of simulations, such as feedback, mastery learning, and distributed practice as well as so-called key instructional design features. The inclusion of key instructional design features had a small to moderate additional benefit on various outcomes, including knowledge and skills (Cook et al., 2013). The meta-analysis by Chernikova, Heitzmann, Stadler et al. (2020) goes beyond this prior meta-analysis because it draws upon a clear framework of simulation-based learning with scaffolding by exploring the effects of examples, prompts, and reflection phases on the acquisition of complex skills like diagnosing. Moreover, it investigated different types of simulations with and without technology use. The results only partly corroborated the expected additional effects of including scaffolding in simulation-based learning. In fact, only certain combinations of scaffolding and types of scaffolding that fit the learners' prerequisites were highly beneficial for acquiring complex skills. On average, simulations that included scaffolding and other instructional supports were only slightly more beneficial for acquiring complex skills than simulations that included no scaffolding or instructional supports at all (Chernikova, Heitzmann, Stadler et al., 2020). However, research on different types of problem-based learning offers some additional, indirect support for a medium-sized added benefit of including scaffolding for fostering cognitive outcomes and competences in diagnosing (Belland, Walker, Kim, & Lefler, 2017; Chernikova, Heitzmann, Fink et al., 2020).

1.5.3 The effects of reflection in simulation-based learning

Reflection itself can be defined as a process in which a learner engages with his or her thoughts, actions, and their bases with the intention of changing them (Nguyen, Fernandez, Karsenti, & Charlin, 2014). In medical education, the form of instructional support known as reflection phases is used to trigger reflective processes beneficial for learning. In this context, reflection phases consist of predefined questions and prompts that ask the learner to reconsider the solution of their problem-solving process (Mamede et al., 2012; Mamede et al., 2014; Mamede, Schmidt, & Penaforte, 2008). The reasoning instructions used in reflection phases can either be rather general or quite specific (Mamede & Schmidt, 2017). For instance, general reasoning instructions might ask learners to interpret all of the provided data when diagnosing. Specific reasoning instructions, on the other hand, might ask learners to provide reasons for and against their diagnoses. In terms of timing, this instructional support can take place before, during, or after simulation-based learning (Beauchamp, 2015; Mamede & Schmidt, 2017). Moreover, there are at least four different mechanisms that can explain the effectiveness of reflection for learning and are also applicable to learning from simulations. First, following the dual-process theory (see Section 1.2), reflection can induce a slow System II process that can prevent biases and correct mistakes stemming from intuitive pattern recognition (Mamede & Schmidt, 2017). Second, reflection may add a pause to the learning process that allows for better application of available but hard to access knowledge in problem-solving (Renkl, Mandl, & Gruber, 1996). Third, reflection may improve the creation of self-generated feedback that could be beneficial for problem-solving (Butler & Winne, 1995). Fourth, reflection could improve metacognitive processes. For example, reflection, especially when it occurs during a task, could have a positive impact on the problem-solving process by improving the metacognitive processes of planning and monitoring.

The literature reports mixed findings on the inclusion of reflection phases as scaffolding to promote diagnostic competences. In a literature review in medical education, Mamede and Schmidt (2017) argue that reflection phases were beneficial in studies on diagnosing under two conditions. First, reflection phases were effective when specific rather than general reasoning instructions were utilized. Second, reflection phases were beneficial when they were used to scrutinize the generated hypotheses and not for other purposes like generating initial hypotheses (Mamede & Schmidt, 2017). In addition to this literature review, results from two meta-analyses are available. The meta-analysis by Chernikova, Heitzmann, Fink et al. (2020) on different types of problem-based learning of diagnostic competences in teacher education and medical education reported medium positive effects of including reflection phases. However, these findings are contradicted by the meta-analysis by Chernikova, Heitzmann, Stadler et al. (2020) on the simulation-based learning of complex skills like diagnosing in various domains. In this meta-analysis, including reflection phases in simulations had no additional benefit (Chernikova, Heitzmann, Stadler et al., 2020). Taken together, the mixed findings from the literature indicate that further research and explanations for the differential effects of reflection phases are warranted. In addition, two other topics could be promising to investigate. As a first additional research topic, it should be examined to what extent the effect of reflection phases depends on individuals' prior knowledge. On the one hand, the previously described meta-analyses (Chernikova, Heitzmann, Fink et al., 2020; Chernikova, Heitzmann, Stadler et al., 2020) showed that learners with high prior knowledge, as operationalized by content familiarity and level of education, experienced greater improvements in diagnostic competences from reflection phases than learners with low prior knowledge. On the other hand, an experiment by Mamede et al. (2010) suggested that a high level of expertise is a necessary prerequisite for learning through reflective thought. In this experiment, only physicians in specialist training, but not medical students, profited from reflective thought when solving complex problems. In sum, it seems likely that prior knowledge is associated with acquiring diagnostic competences from reflection phases when medical students solve regular rather than complex cases. As a second additional research topic, empirical evidence on the diagnostic process during reflection phases should be gathered. To my knowledge, only one study by Mamede et al. (2020) provided insights into learners' hypotheses during reflection phases. In this study, participants completed different types of reflection phases in four experimental groups containing either no instructions or instructions to argue for, against, or both for and against their hypotheses. The participants were prompted before and after each type of reflection phase to name their current hypothesis. Analyses showed that the accuracy of hypotheses increased from the first to the second measurement point regardless of the type of reflection phase used (Mamede et al., 2020).

The next section states the central research questions underlying this dissertation, elaborates on the simulation-based environment used, and provides an overview of the conducted studies and enclosed articles.

1.6 Research Questions, Simulation-Based Environment, and Overview of Research

As we have seen, many important research questions on the simulation-based assessment and facilitation of diagnostic competences remain unresolved. This dissertation aims to enhance knowledge on the following three central research questions:

1. What differences and similarities emerge in the assessment of diagnostic competences with the two assessment methods of standardized patients and virtual patients?
2. How are process variables related to diagnostic quality in simulation-based assessment?
3. To what extent can the simulation-based learning of diagnostic competences be facilitated with scaffolding?

I now briefly describe the simulation-based environment developed before giving an overview of the conducted studies and enclosed articles.

1.6.1 Simulation-based environment

The design and development process for the studies making up this dissertation can only briefly be summarized at this point. A more detailed account can be found in Fink, Reitmeier, Siebeck, Fischer, and Fischer (2022). I developed materials for use with the two assessment methods of standardized patients and virtual patients together with an interdisciplinary team. The interdisciplinary team consisted of a board-certified physician, two medical education professors, an educational psychology professor, and a computer scientist. History-taking was selected as the professional situation to simulate for two reasons. First, history-taking generates a large share of the information used in diagnosing (Keifenheim et al., 2015). Second, diagnoses can frequently be made correctly after history-taking and without any further diagnostic steps (Peterson, Holbrook, Von Hales, Smith, & Staker, 1992). The topic of dyspnea was chosen because of its high relevancy for emergency medicine and general medicine (Berliner, Schneider, Welte, & Bauersachs, 2016). Next, a framework for the simulations was created. In the simulations, the participant first obtains prior information (e.g., ECG results) before watching or observing a patient's chief complaint. Afterward, a phase of independent history-taking takes place, followed by the completion of a case summary. Based on this framework, nine case vignettes and instruments for assessing diagnostic competences were developed. An expert workshop validated these materials, and minor revisions were made. Subsequently, the following steps were taken. For the standardized patients, professional actors were trained by a physician and an acting coach. For the virtual patients, professional actors were filmed displaying the signs and symptoms involved in the cases. Then, Study 1 was conducted using these materials. For Study 2, the virtual patients created for Study 1 were utilized, with a few additional minor changes. Next, I describe the conducted empirical studies.

1.6.2 Overview of the studies

Two empirical studies were conducted for this dissertation. Study 1 contrasted the assessment of standardized patients and virtual patients in a sample of *N* = 86 medical students and is reported in Article 1. Study 2 focused on simulation-based education in virtual patients. Two different articles included in this dissertation were created from Study 2. However, the samples for the two articles differ due to their different research questions and analyses. Article 2 investigated the relationships between diagnostic activities, professional knowledge, and diagnostic quality in diagnosing virtual patients in a sample of *N* = 106 medical students. Article

3 examined the effect of reflection phases on learning to diagnose accurately through virtual patients in a sample of $N = 121$ medical students. It should also be noted what data was used in each article. Article 2 used only data from the test cases of Study 2 and thus contained no scaffolding. Article 3, on the other hand, examined data from all phases of Study 2 and therefore included scaffolding within the learning phase in the intervention groups. Below, I outline how the articles are connected to the central research questions, summarize their design and assessed variables, and provide an overview of their research questions.

1.6.3 Overview of Article 1

The first article compared the assessment of diagnostic competences with standardized patients and virtual patients, mainly contributing to the first central research question. The underlying study employed a repeated measures design in which participants worked with both standardized patients and virtual patients. Diagnostic accuracy, perceived authenticity, and cognitive load were measured after each simulation. Moreover, the quality of evidence generation was assessed. The following research questions were investigated:

1. To what extent does perceived authenticity differ across the two assessment methods, and how is it associated with diagnostic accuracy?
2. Is cognitive load equivalent for standardized patients and virtual patients, and how is it related to diagnostic accuracy?
3. To what extent are the diagnostic competences components diagnostic accuracy, quantity of evidence generation, and quality of evidence generation equivalent or differ for standardized patients and virtual patients, and how are they related to each other?

1.6.4 Overview of Article 2

The second article investigated the relationships between diagnostic activities, professional knowledge, and diagnostic quality in order to gain insights regarding the second central research question. The procedure was as follows. Participants first filled out conceptual and strategic professional knowledge tests before completing an assessment with virtual patients without scaffolding. Diagnostic quality was measured with a diagnostic accuracy score and a comprehensive diagnostic score that included 1) diagnostic accuracy, 2) treatment selected, 3) diagnostic measures for medical clarification, and 4) expected findings in a physical examination. To assess diagnostic activities, the variables hypothesis generation, evidence generation, and evidence evaluation were tracked while participants worked with the virtual patients. This article sought to answer the research question: To what extent do diagnostic activities and professional knowledge uniquely explain variance in diagnostic quality?

1.6.5 Overview of Article 3

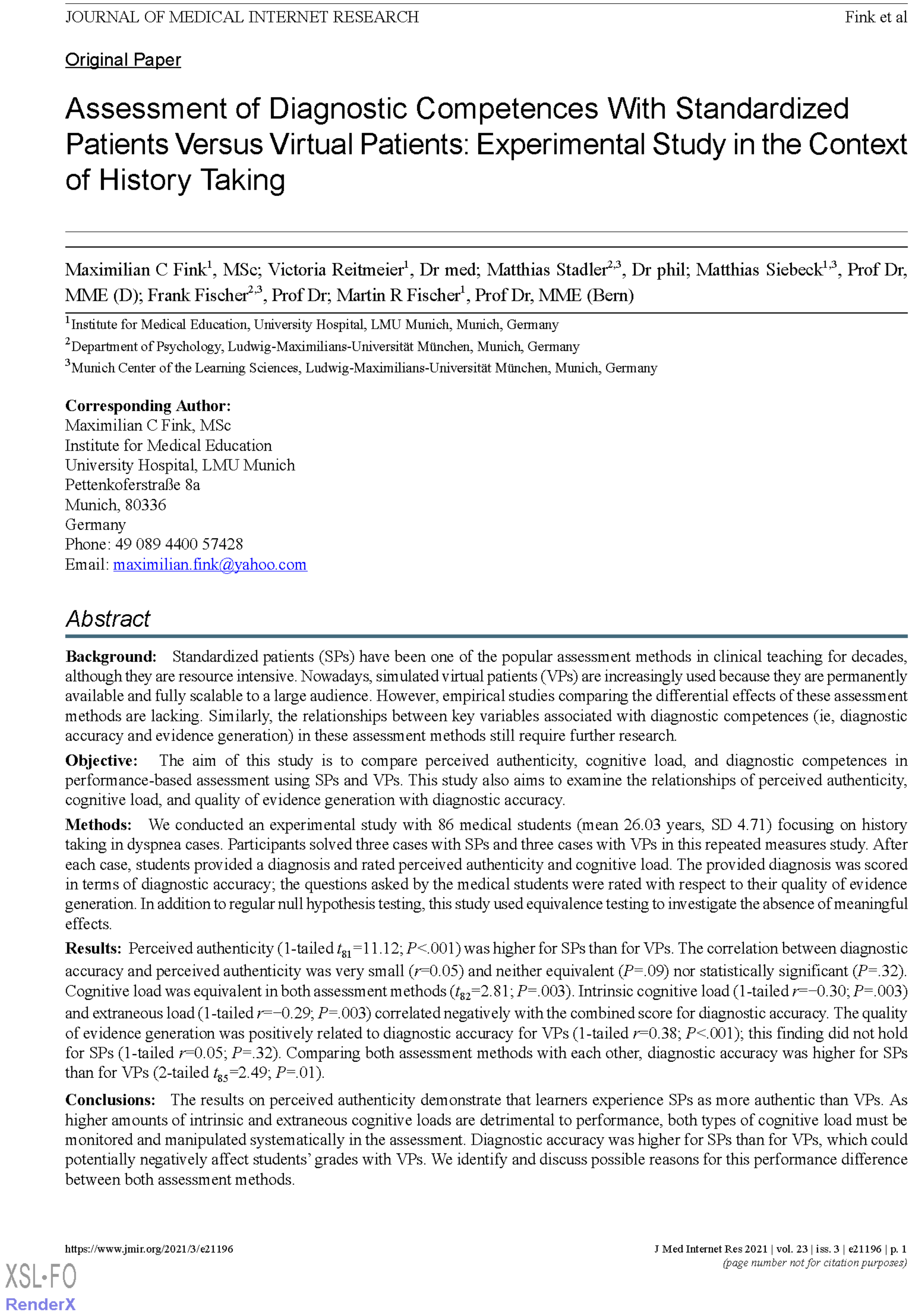
The third article examined the effects of reflection phases on learning to diagnose accurately through virtual patients, thus contributing to the third central research question. An experiment with a pre- and post-test and a control group was conducted. In the pretest, prior conceptual and strategic knowledge were assessed, and participants diagnosed virtual patients. In the intervention, two groups learned from virtual patients and completed two slightly different types of reflection phases. The control group learned from regular virtual patients without reflection phases. In the post-test, the participants once again diagnosed virtual patients. Diagnostic accuracy was measured for all virtual patients as the primary outcome. Moreover, hypotheses were tracked while working with the virtual patients and in reflection phases as a measure of the diagnostic process. This article investigated the following research questions:

1. To what extent do reflection phases affect learning to diagnose accurately in virtual patients?
2. To what extent is prior knowledge associated with learning to diagnose accurately through reflection phases?
3. To what extent does the diagnostic process improve during simulation-based learning with virtual patients and during reflection phases, in the sense of enhancements in current hypotheses and diagnostic accuracy over the course of cases?

2. Article 1: Assessment of Diagnostic Competences with Standardized Patients versus Virtual Patients: Experimental Study in the Context of History Taking

Reference

Fink, M. C., Reitmeier, V., Stadler, M., Siebeck, M., Fischer, F., & Fischer, M. R. (2021). Assessment of diagnostic competences with standardized patients versus virtual patients: Experimental study in the context of history taking. *Journal of Medical Internet Research*, *23*(3), e21196. https://doi.org/10.2196/21196



Original Paper

Assessment of Diagnostic Competences With Standardized Patients Versus Virtual Patients: Experimental Study in the Context of History Taking

Maximilian C Fink[1], MSc; Victoria Reitmeier[1], Dr med; Matthias Stadler[2,3], Dr phil; Matthias Siebeck[1,3], Prof Dr, MME (D); Frank Fischer[2,3], Prof Dr; Martin R Fischer[1], Prof Dr, MME (Bern)

[1]Institute for Medical Education, University Hospital, LMU Munich, Munich, Germany

[2]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

[3]Munich Center of the Learning Sciences, Ludwig-Maximilians-Universität München, Munich, Germany

Corresponding Author:
Maximilian C Fink, MSc
Institute for Medical Education
University Hospital, LMU Munich
Pettenkoferstraße 8a
Munich, 80336
Germany
Phone: 49 089 4400 57428
Email: maximilian.fink@yahoo.com

Abstract

Background: Standardized patients (SPs) have been one of the popular assessment methods in clinical teaching for decades, although they are resource intensive. Nowadays, simulated virtual patients (VPs) are increasingly used because they are permanently available and fully scalable to a large audience. However, empirical studies comparing the differential effects of these assessment methods are lacking. Similarly, the relationships between key variables associated with diagnostic competences (ie, diagnostic accuracy and evidence generation) in these assessment methods still require further research.

Objective: The aim of this study is to compare perceived authenticity, cognitive load, and diagnostic competences in performance-based assessment using SPs and VPs. This study also aims to examine the relationships of perceived authenticity, cognitive load, and quality of evidence generation with diagnostic accuracy.

Methods: We conducted an experimental study with 86 medical students (mean 26.03 years, SD 4.71) focusing on history taking in dyspnea cases. Participants solved three cases with SPs and three cases with VPs in this repeated measures study. After each case, students provided a diagnosis and rated perceived authenticity and cognitive load. The provided diagnosis was scored in terms of diagnostic accuracy; the questions asked by the medical students were rated with respect to their quality of evidence generation. In addition to regular null hypothesis testing, this study used equivalence testing to investigate the absence of meaningful effects.

Results: Perceived authenticity (1-tailed t_{81}=11.12; *P*<.001) was higher for SPs than for VPs. The correlation between diagnostic accuracy and perceived authenticity was very small (*r*=0.05) and neither equivalent (*P*=.09) nor statistically significant (*P*=.32). Cognitive load was equivalent in both assessment methods (t_{82}=2.81; *P*=.003). Intrinsic cognitive load (1-tailed *r*=−0.30; *P*=.003) and extraneous load (1-tailed *r*=−0.29; *P*=.003) correlated negatively with the combined score for diagnostic accuracy. The quality of evidence generation was positively related to diagnostic accuracy for VPs (1-tailed *r*=0.38; *P*<.001); this finding did not hold for SPs (1-tailed *r*=0.05; *P*=.32). Comparing both assessment methods with each other, diagnostic accuracy was higher for SPs than for VPs (2-tailed t_{85}=2.49; *P*=.01).

Conclusions: The results on perceived authenticity demonstrate that learners experience SPs as more authentic than VPs. As higher amounts of intrinsic and extraneous cognitive loads are detrimental to performance, both types of cognitive load must be monitored and manipulated systematically in the assessment. Diagnostic accuracy was higher for SPs than for VPs, which could potentially negatively affect students' grades with VPs. We identify and discuss possible reasons for this performance difference between both assessment methods.

(J Med Internet Res 2021;23(3):e21196) doi: 10.2196/21196

KEYWORDS

clinical reasoning; medical education; performance-based assessment; simulation; standardized patient; virtual patient

Introduction

Performance-Based Assessment With Standardized Patients and Virtual Patients

Since the turn of the millennium, performance-based assessment has become a mandatory part of medical licensure examinations in various countries [1], complementing traditional assessment formats, such as text vignettes, with methods including standardized patients (SPs) and simulated virtual patients (VPs). SPs have been used for performance-based assessment in health care since the 1960s [2]. However, VPs have only recently become more widely employed in this domain [3].

The term SPs refers to (trained) actors or real former patients who act as if they display symptoms of a disease [4]. Usually, students encounter several SPs in assessment settings to reliably measure clinical variety [5]. Performance is then scored by a trained faculty member or the SPs themselves using a rating scheme. Although we will elaborate on the specific features used for this assessment method later, it should be noted here that organizing an assessment with SPs is relatively resource intensive [6].

VPs are a type of computer simulation and typically include an authentic model of a real-world situation that can be manipulated by the participant [7]. VPs can use avatars or realistic videos with SPs as stimuli and offer varying degrees of interaction [8]. Moreover, assessment through VPs can take place automatically, and a recent study showed that such an automatic assessment corresponds well to ratings from clinician-educators [9]. The production of authentic VPs can frequently produce considerable costs above $10,000 [10]. Although the initial production of VPs is often more resource intensive than organizing SPs, this assessment method is then permanently available and fully scalable to a large audience.

Next, we summarize a conceptual framework. This framework provides, on the one hand, a precise operationalization of diagnostic competences. On the other hand, the framework includes a research agenda that summarizes essential moderators of performance that should be examined systematically in research on simulation-based assessment.

A Framework for the Assessment of Diagnostic Competences With Simulations

The framework developed by Heitzmann et al [10] to facilitate diagnostic competences with simulations operationalizes diagnostic competences in assessment settings as a disposition. This disposition encompasses the components of diagnostic knowledge, diagnostic quality, and diagnostic activities. Diagnostic knowledge includes conceptual and strategic knowledge [11]. Conceptual knowledge encompasses concepts and their relationships. Strategic knowledge comprises possible avenues and heuristics in diagnosing. Diagnostic quality consists of components' diagnostic accuracy and efficiency that can serve as major outcome measures in empirical studies. Diagnostic activities entail the actions of persons assessed during the diagnostic process, such as evidence generation by asking questions in history taking. The framework proposes that context is an important moderator in assessment. Therefore, more research on the effects of the assessment methods SPs and VPs seems to be warranted. A meta-analysis on simulation-based learning of complex skills [12] added to this framework that authenticity should also be explored as an important moderator in assessment and learning. Similarly, a meta-analysis on instructional design features in simulation-based learning indicated that certain types of cognitive load could be detrimental to performance [13]. Therefore, it could be fruitful to explore the relationship between cognitive load and diagnostic competences within SP and VP assessments.

Perceived Authenticity and Diagnostic Competences With SPs and VPs

There is a multitude of conceptualizations of authenticity. In our study, we focus on *perceived authenticity* [14] because this concept can be assessed entirely internally by learners' judgment. Other related concepts such as *thick authenticity* [15] and *fidelity* [16] can, at least to some extent, also be determined externally.

According to a factor analysis by Schubert et al [14], perceived authenticity—sometimes also called presence—comprises the facets of realness, involvement, and spatial presence. Realness describes the degree to which a person believes that a situation and its characteristics resemble a real-life context [14]. Involvement is defined as a feeling of cognitive immersion and judgment that a situation has personal relevancy [17]. Spatial presence denotes the feeling of physical immersion in a situation [14]. SPs are considered highly authentic because they are carefully trained to realistically portray symptoms and allow for natural interactions [18]. Empirical studies support this claim, reporting high values of perceived authenticity for SPs [19,20]. VPs also received rather high perceived authenticity scores in empirical studies [21] but lacked some of the features that may make SPs particularly authentic, such as high interactivity in oral conversations. Thus, VPs could potentially evoke lower perceived authenticity than SPs. Findings on the effect of authenticity on diagnostic competences are mixed. On the one hand, it has been argued that higher authenticity is associated with higher engagement and better performance [22]. On the other hand, literature reviews [23,24] that compared the relationship between perceived authenticity and clinical performance in simulation-based learning only reported minimal effects of authenticity. In addition, an empirical study [25] showed that above a certain threshold, further increases in perceived authenticity do not improve diagnostic accuracy.

Cognitive Load and Diagnostic Competences With SPs and VPs

Cognitive load theory posits that performance can be inhibited through high situational demands that stress working memory and attention [26]. The cognitive load consists of the following 3 different facets [27]: *Intrinsic* load results from the interplay between certain topics and materials and the assessed person's expertise. *Extraneous load* is created exclusively by characteristics of the assessment environment that strain memory and attention without being necessary for performance. *Germane load* refers to the cognitive load created through the assessed person's cognitive processes, including schema construction and abstraction. Intrinsic and extraneous cognitive loads are considered additive and can inhibit performance in complex tasks [27]. Germane load, however, is theorized to bolster performance [27]. A few primary studies from medical education have already contrasted the cognitive load of different assessment methods and reported their relationship with diagnostic competences. Dankbaar et al [28] demonstrated that intrinsic and germane cognitive loads were higher for a group learning emergency skills with a simulation game than for a group learning with a text-based simulation. Extraneous load did not differ between these groups, and none of the groups differed in performance. Haji et al [29] compared surgical skills training with less complex and more complex simulation tasks. The total cognitive load was higher in the more complex simulation than in the less complex simulation, and cognitive load was negatively associated with performance. As a result of these findings, we can conclude that SPs and VPs generally do not differ in different facets of cognitive load if the assessment methods are of equal complexity, and the main characteristics related to the facets are similar. The literature summarized earlier also shows that intrinsic and extraneous cognitive loads are negatively associated with diagnostic competences.

Assessment Method and Diagnostic Competences

Before we discuss diagnostic accuracy and evidence generation—2 important aspects of diagnostic competences—it should be noted that diagnostic competences are only a part of the broader concept of clinical reasoning. Clinical reasoning emphasizes the process of diagnosing and encompasses the full process of making clinical decisions, including the selection, planning, and reevaluation of a selected intervention [30]. In line with the conceptual framework by Heitzmann et al [10] for facilitating diagnostic competences, *diagnostic accuracy* denotes the correspondence between the learner's diagnoses and the solutions determined by experts for the same cases. According to this framework, *evidence generation* (ie, actions related to the gathering of data in a goal-oriented way) is also an important quality criterion for the diagnostic process and a crucial aspect of diagnostic competences.

Diagnostic Accuracy

Currently, there are only a few studies in the health care domain that contrast assessments using VPs and SPs directly in one experiment. Edelstein et al [1] investigated assessments with SPs and computer-based case simulations in advanced medical students using a repeated measures design. A moderate positive correlation was found between diagnostic accuracy in the two assessment formats that used different cases. Guagnano et al [31] examined SPs and computer-based case simulations in a medical licensing exam. Participants first completed the computer-based case simulations and then completed the SPs. The two assessment methods correlated positively with each other. Hawkins et al [32] compared the assessment of patient management skills and clinical skills with SPs and computer-based case simulations in a randomized controlled trial. Participating physicians completed both assessment methods, and a positive correlation of diagnostic accuracy with both assessment methods was reported. Outside the health care domain, a meta-analysis of studies from different domains reported a robust modality effect for students in problem-solving tasks. Students who solved problems presented in the form of illustrations accompanied by text were more successful than students who solved problems presented merely in text form [33]. Similarly, it seems reasonable to assume that one assessment method could lead to higher diagnostic accuracy than the other assessment method because of its different characteristics. The described findings from the health care domain tentatively indicate that SPs and VPs could result in relatively equivalent diagnostic accuracy. Such a finding would contradict the modality effect reported in other domains.

Evidence Generation

Comparable empirical studies on evidence generation for SPs and VPs are lacking. Nevertheless, we can assume that the quantity of evidence generation should be higher for SPs than for VPs. The main reason for this is that students can ask questions of SPs more quickly orally than by selecting questions from a menu of options with VPs. Apart from this difference in evidence generation between the 2 assessment methods, the relationships between evidence generation and diagnostic accuracy are interesting. The relationship between the quantity of evidence generation and diagnostic accuracy is relatively complex. The ideal amount of evidence generation may depend strongly on the case difficulty, the diagnostic cues contained in the evidence, and learner characteristics. For these reasons, the framework by Heitzmann et al [10] for facilitating diagnostic competences argues that the sheer quantity of evidence generation is not a dependable quality criterion for the diagnostic process. However, the quality of evidence generation is hypothesized by Heitzmann et al [10] to be a rather dependable quality criterion for the diagnostic process. This agrees with the literature, as we know from studies on SPs using observational checklists that the quality of evidence generation is positively associated with diagnostic accuracy [34]. Moreover, one study with specialists in internal medicine and real patients demonstrated that asking specific questions in history taking correlated positively with clinical problem solving [35].

Study Aim, Research Questions, and Hypotheses

We aim to compare the perceived authenticity, cognitive load, and diagnostic competences in SPs and VPs. We also aim to examine the relationships of perceived authenticity, cognitive load, and quality of evidence generation with diagnostic accuracy. Thus, we address the following 3 research questions: To what extent does perceived authenticity differ across the 2 assessment methods, and how is it associated with diagnostic accuracy (RQ1)? We hypothesize that SPs induce higher perceived authenticity than VPs (H1.1). Moreover, we expect to be able to demonstrate with equivalence tests for correlations (given in the *Statistical Analyses* section) that perceived authenticity is not associated meaningfully with diagnostic accuracy (H1.2). Next, is cognitive load equivalent for SPs and VPs, and how is it related to diagnostic accuracy (RQ2)? We assume to find equivalent cognitive load for SPs and VPs (H2.1). Moreover, we expect that intrinsic and extraneous loads are negatively related to diagnostic accuracy (H2.2-H2.3). To what extent are the diagnostic competences components diagnostic accuracy, quantity of evidence generation, and quality of evidence generation equivalent or differ for SPs and VPs, and how are they related to each other (RQ3)? We hypothesize that SPs and VPs evoke equivalent diagnostic accuracy (H3.1). In addition, we assume that the quantity of evidence generation is higher for SPs than for VPs (H3.2). We also expect that the quality of evidence generation is positively related to diagnostic accuracy (H3.3).

Methods

Participant Characteristics and Sampling Procedures

A sample of 86 German medical students (with a mean age of 26.03 years, SD 4.71) made up the final data set. This sample consisted of 63% (54/86) females and 37% (32/86) males. Medical students in years 3-6 of a 6-year program with a good command of German were eligible. Medical students in years 3-5 (44/86, 51%) were considered novices, as they were still completing the clinical part of the medical school. Medical students in year 6 (42/86, 49%) were regarded as intermediates as they had passed their second national examination and worked full time as interns in a medical clinic or practice. We provide a detailed overview of participant characteristics across all conditions and a CONSORT (Consolidated Standards of Reporting Trials)–style diagram of participant flow in Multimedia Appendix 1.

We collected data from October 20, 2018, to February 20, 2019, in the medical simulation center of the University Hospital, LMU Munich. We recruited participants via on-campus and web-based advertising. Participants were randomly assigned to conditions by the first author by drawing a pin code to log in to an electronic learning environment without knowing the condition assigned to the pin. In the final data collection sessions, the conditions were filled by the first author with random participants from specific expertise groups (novices vs intermediates). This procedure was applied to achieve a comparable level of expertise in all conditions. As expected, the proportion of participants from different expertise groups did not differ across conditions (χ^2_3=0.2; *P*=.99).

Research Design

The study used a repeated measures design with assessment method (SPs vs VPs) as the key factor. In addition, we varied the between-subjects factor case group (CG) order and assessment method order. In total, students encountered 6 different cases. We provide an overview of the experiment in Table 1. Details of the succession through cases and medical content in the experimental conditions are provided in Table 2. We attempted to ensure similar topics and difficulty for both CGs by conducting an expert workshop and adapting cases based on the experts' feedback as part of creating the experimental materials.

Table 1. General overview of the experiment.

Part of the experiment	Activity or test	Duration (min)
Pretest	Briefing	10
	Conceptual knowledge test	40
	Strategic knowledge test	40
Break	—[a]	10
Assessment phase I (cases 1-3)	VPs[b] or SPs[c]	70
Break and change of modality	—	5
Assessment phase II (cases 4-6)	VPs or SPs	70
Posttest and debriefing	Working memory test	15
	End-debriefing	5

[a] No activity or test takes place.

[b] VP: virtual patient.

[c] SP: standardized patient.

Table 2. Succession through cases and medical content in the experimental conditions[a,b].

Cases	Condition 1A	Condition 1B	Condition 2A	Condition 2B
1-3	CG[c] A (SPs[d])	CG B (VPs[e])	CG B (SPs)	CG A (VPs)
4-6	CG B (VPs)	CG A (SPs)	CG A (VPs)	CG B (SPs)

[a] Case group A: (1) pulmonary embolism with lymphoma, (2) congestive heart failure with atrial fibrillation, and (3) hyperventilation tetany caused by a panic attack.

[b] Case group B: (1) pulmonary embolism with coagulation disorder, (2) community-acquired pneumonia, and (3) hypertrophic obstructive cardiomyopathy.

[c] CG: case group.

[d] SP: standardized patient.

[e] VP: virtual patient.

Procedure and Materials

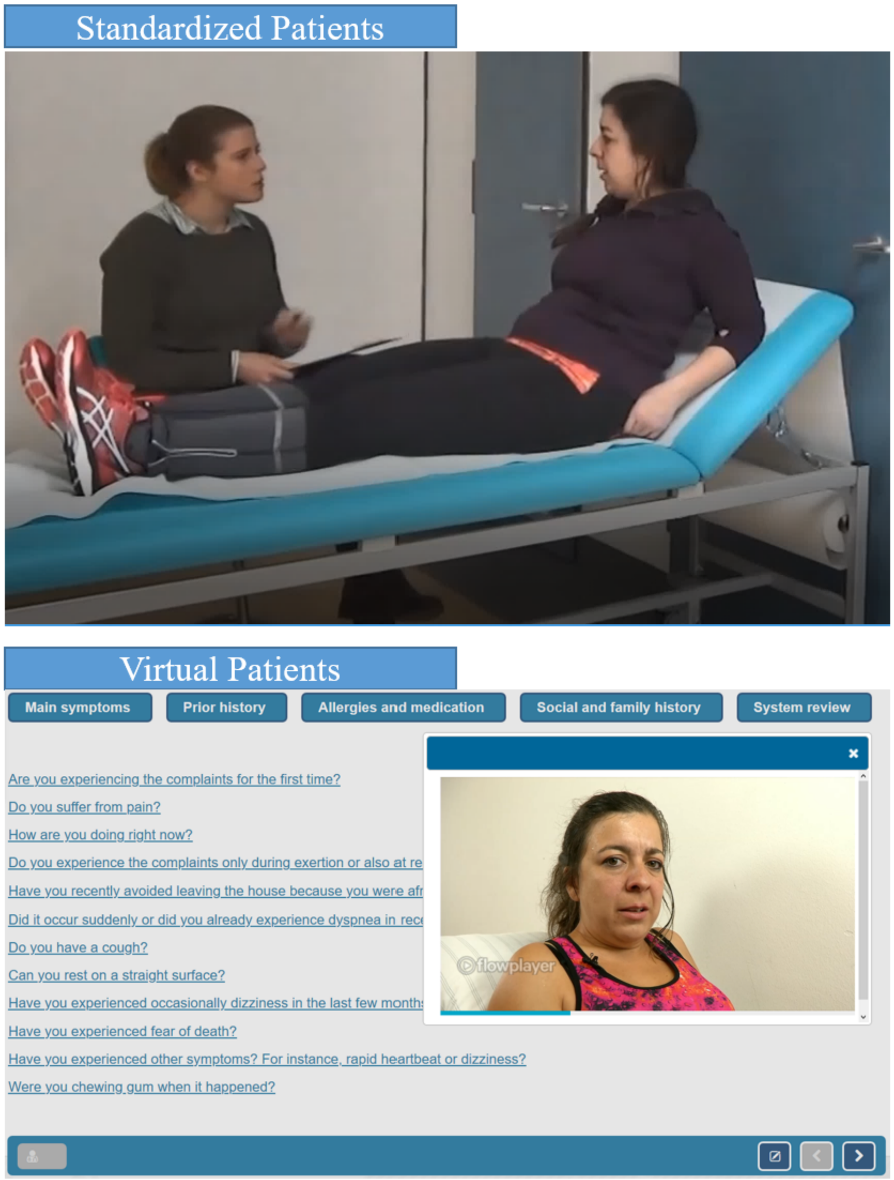
Participants completed a pretest of conceptual knowledge and strategic knowledge at the beginning of the experiment. Afterward, participants took part in the assessment phase, solving the first 3 cases with SPs and the next 3 cases with VPs or vice versa. All cases were drafted by a specialist in general practice and evaluated positively by an expert panel. The cases were not adapted from real clinical cases but based on cases from textbooks and symptoms reported in guidelines. A short familiarization phase preceded each assessment phase and included a motivational scale. For all cases in both assessment methods, assessment time was held constant at 8 minutes and 30 seconds for history taking and 5 minutes for writing up a diagnosis for the case in an electronic patient file. At the end of the experiment, participants were debriefed. A more detailed overview of the procedure can be found in Multimedia Appendix 2.

Assessment with SPs was conducted in a simulated emergency room. All SPs were (semi-) professional actors who were financially compensated; most had previous experience working in an SP program. All SPs were extensively trained by an acting coach and a physician, memorized their symptoms and scripts, and were not aware of their patient’s diagnosis. Participants first received prior information (eg, electrocardiogram and lab results) and presentation of the chief complaint for each case. Next, participants formulated and asked questions independently, and the SPs responded. The interaction was recorded on a video. After each case, the participants completed a patient file, including measures of diagnostic accuracy and other scales. A screenshot of this assessment method is provided in Figure 1.

The assessment with the VPs was carried out in a simulated assessment environment in a computer room. First, participants received prior information and a video with a chief complaint for each case. The participants then selected questions independently from a menu with up to 69 history-taking questions. The VP’s answer was streamed as a video, including a recorded response by an actor. After each case, the participants completed a patient file, including a measure of diagnostic accuracy and other scales. A screenshot of this assessment method is provided in Figure 1.

The VPs, patient file, and other measures were implemented in the electronic assessment environment CASUS [36]. The questions provided for the VPs were based on a structural and topical analysis of history-taking forms by Bornemann [37] and are displayed in Multimedia Appendix 3. According to this analysis, physician questions in history taking can fall under the 5 categories of main symptoms, prior history, allergies and medication, social and family history, and system review. Participants with SPs received empty history-taking forms for all cases and time to formulate possible history-taking questions during the familiarization phase, at which point participants in the VPs only read all questions from the menu. Without this additional structuring support in the SP condition, the participants in the VP condition would have received additional support in the form of a list of questions in the menu.

Figure 1. History-taking with standardized patients and virtual patients.



Are you experiencing the complaints for the first time?

Do you suffer from pain?

How are you doing right now?

Do you experience the complaints only during exertion or also at re

Have you recently avoided leaving the house because you were afr

Did it occur suddenly or did you already experience dyspnea in rece

Do you have a cough?

Can you rest on a straight surface?

Have you experienced occasionally dizziness in the last few months

Have you experienced fear of death?

Have you experienced other symptoms? For instance, rapid heartbeat or dizziness?

Were you chewing gum when it happened?

Measures and Covariates

Perceived Authenticity

Perceived authenticity was operationalized as a construct with the 3 dimensions of realness, involvement, and spatial presence [14]. All 3 authenticity scales used a 5-point scale ranging from (1) *disagree* to (5) *agree* and were taken from multiple validated questionnaires [14,38-40]. The items were slightly adapted to simulation-based assessment and are included in Multimedia Appendix 4. A combined score for all 3 dimensions was built by calculating the mean. This scale achieved a reliability of Cronbach α=.88.

Cognitive Load

The cognitive load scale by Opfermann [41] used in this study assessed the extraneous cognitive load with 3 items and germane and intrinsic cognitive loads with 1 item each. A 5-point scale from (1) *very easy*, (2) *rather easy*, (3) *neutral*, (4) *rather hard*, to (5) *very hard* was used. The scale is included in Multimedia Appendix 4. A combined score for all 3 facets was built by calculating the mean. This scale achieved a reliability of Cronbach α=.88.

Motivation, Diagnostic Knowledge, and Other Control Variables

We assessed motivation as a control variable because it could differ between assessment methods and potentially affect performance. The expectancy component of motivation was assessed with a 4-item, 7-point scale adapted from Rheinberg et al [42]. The motivation expectancy scale ranged from (1) strongly disagree to (7) strongly agree. The value component of motivation was measured with a 4-item, 5-point scale based on a questionnaire by Wigfield [43]. The motivation value scale ranged from (1) strongly disagree to (5) strongly agree. The full scales are provided in Multimedia Appendix 4. Diagnostic knowledge was also measured in this study but later not taken into account in the analyses because it was similar in VPs and SPs because of the repeated measures design. We measured diagnostic knowledge using a conceptual and strategic knowledge test. Both types of knowledge have been identified as predictors of clinical reasoning [44]. The maximum testing time was set to 40 minutes per test. More details on both diagnostic knowledge tests are reported in Multimedia Appendix 4. Apart from this, demographic data were collected, including participants' sex, age, and expertise (year of medical school).

Diagnostic Competences

Diagnostic Accuracy

Diagnostic accuracy was assessed based on the answer to the prompt "Please choose your final diagnosis after history taking" from a long menu containing 239 alternative diagnoses. Two physicians created a coding scheme for scoring diagnostic accuracy in all cases (Multimedia Appendix 4). To do that, the physicians rated all 239 alternative diagnoses for all cases and resolved the disagreements until they reached full agreement. One of the physicians was a specialist in general practice who also drafted the cases. The other physician was a board-certified doctor familiar with medical assessment through her dissertation. The latter physician, who is also the second author of this paper, then scored diagnostic accuracy based on the coding scheme: 1 point was allocated for the designated correct answer, 0.5 point for a partially correct answer, and 0 point for an incorrect answer. Due to having only 1 rater to score the diagnostic accuracy with the comprehensive coding scheme, a reliability estimate cannot be reported. However, this is also not necessary because the exact diagnostic accuracy score for all selectable diagnoses included in the electronic assessment environment was determined upfront in the coding scheme.

Evidence Generation

The second author classified the quality of evidence generation by determining the essential questions relevant for the correct diagnosis for each VP case (the coding scheme is given in Multimedia Appendix 4). This process took part before looking at the experimental data. All solutions were discussed with a specialist in general practice, and all disagreements were resolved. Student assistants transcribed all utterances recorded in the videos of the SP encounters, and the electronic assessment environment stored all selected questions during the VP encounters. The *R* scripts automatically classified the log data from the VPs using the coding scheme. Student assistants had no medical background and were trained by the second author to code the transcripts from the SP encounters. This task mainly implied recognizing the intent of history-taking questions and linking them, if possible, to the most similar question in the coding scheme. After training the raters, 20% of this complex and extensive SP data were coded by 2 raters to check interrater agreement. This data set encompassed SP data from 18 of the 86 participants of our study with all three SP cases in which the participants took part. Fleiss κ=0.74 demonstrated that agreement was substantial, and the rest of the data were coded by the same raters individually. The score for quantity of evidence generation corresponded to the total number of questions posed for each case. To calculate the score for quality of evidence generation for each case, we counted the number of relevant questions posed and divided this score by the number of relevant questions that could potentially be posed.

Scale Construction

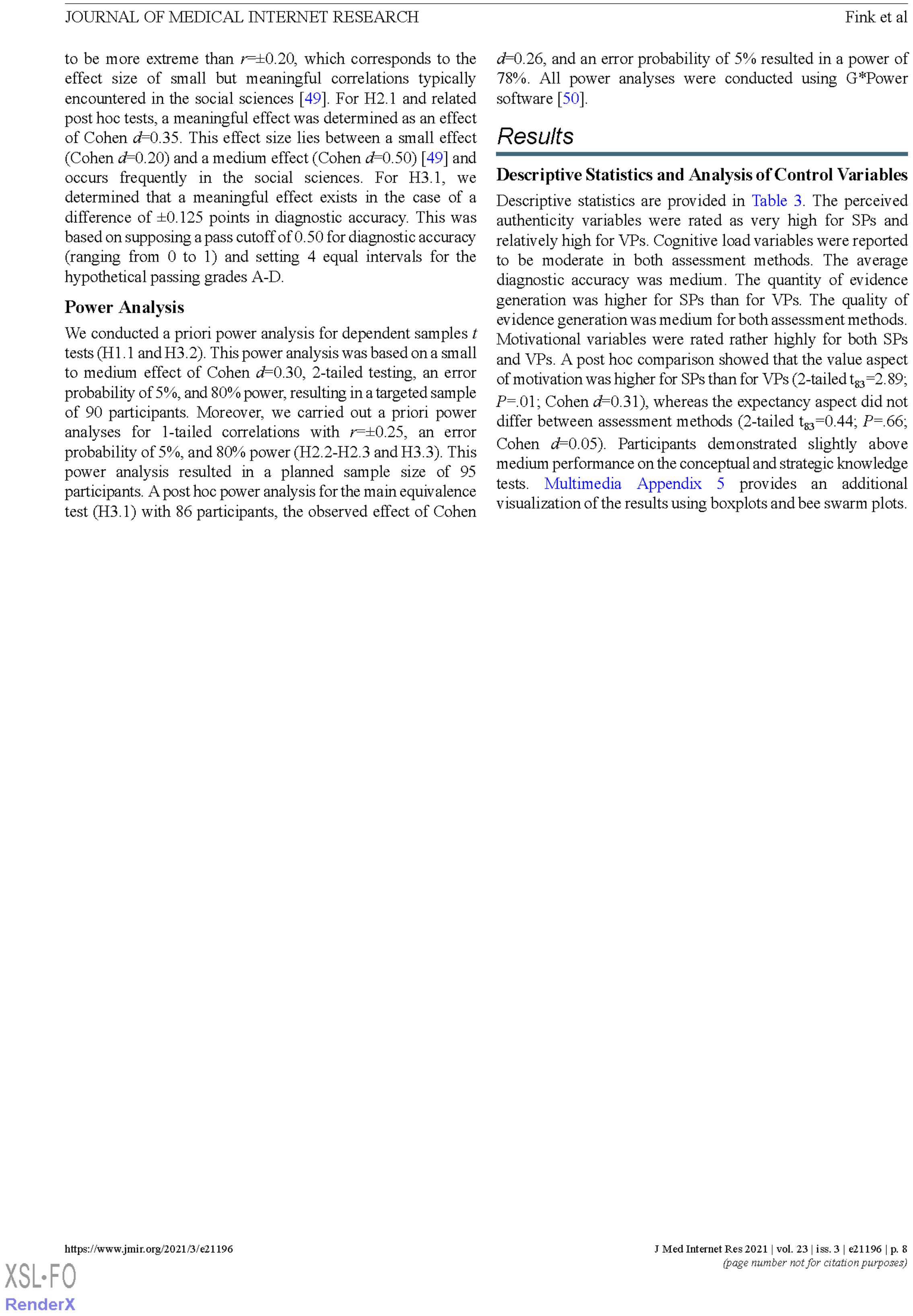
Diagnostic accuracy and evidence generation scales for each assessment method and combining the 2 methods were built by calculating the mean of the included cases. Case 1 in CS A was excluded from all analyses because of high difficulty (mean diagnostic accuracy 0.05, SD 0.18).

Statistical Analyses

This study answers the proposed research questions using traditional null hypothesis significance testing (NHST) and equivalence testing. In contrast to NHST, equivalence testing can be used to investigate "whether an observed effect is surprisingly small, assuming that a meaningful effect exists in the population" [45]. For this type of test, first, the smallest effect size of interest, that is, the threshold for a meaningful effect, is specified based on the literature. The null hypothesis that the effect is more extreme than the smallest effect size of interest is then investigated. To do this, 2 separate 1-sided tests (TOST; eg, *t* tests) are conducted [46]. These tests examine whether the observed effect is more extreme than the specified smallest effect size of interest. If both 1-sided tests are significant, the null hypothesis that there is a meaningful effect that is more extreme than the smallest effect size of interest is rejected. Thus, equivalence is supported. For more convenient reporting, only the *t* test with a higher P value is reported. In cases in which equivalence cannot be supported, NHST is performed for follow-up analyses.

All statistical analyses were performed using *R* version 3.6.1 [47]. The TOST procedure and the corresponding package TOSTER [45] were used to conduct the equivalence tests. In all statistical analyses, the alpha level was set to 5%; 1-tailed tests were used where applicable. The Bonferroni-Holm method [48] was used to correct P values for multiple comparisons in post hoc and explorative tests.

For all equivalence tests, the smallest effect size of interest was determined based on the discussed literature. For H1.2 and related post hoc tests, the smallest effect size of interest was set



to be more extreme than r=±0.20, which corresponds to the effect size of small but meaningful correlations typically encountered in the social sciences [49]. For H2.1 and related post hoc tests, a meaningful effect was determined as an effect of Cohen d=0.35. This effect size lies between a small effect (Cohen d=0.20) and a medium effect (Cohen d=0.50) [49] and occurs frequently in the social sciences. For H3.1, we determined that a meaningful effect exists in the case of a difference of ±0.125 points in diagnostic accuracy. This was based on supposing a pass cutoff of 0.50 for diagnostic accuracy (ranging from 0 to 1) and setting 4 equal intervals for the hypothetical passing grades A-D.

Power Analysis

We conducted a priori power analysis for dependent samples t tests (H1.1 and H3.2). This power analysis was based on a small to medium effect of Cohen d=0.30, 2-tailed testing, an error probability of 5%, and 80% power, resulting in a targeted sample of 90 participants. Moreover, we carried out a priori power analyses for 1-tailed correlations with r=±0.25, an error probability of 5%, and 80% power (H2.2-H2.3 and H3.3). This power analysis resulted in a planned sample size of 95 participants. A post hoc power analysis for the main equivalence test (H3.1) with 86 participants, the observed effect of Cohen d=0.26, and an error probability of 5% resulted in a power of 78%. All power analyses were conducted using G*Power software [50].

Results

Descriptive Statistics and Analysis of Control Variables

Descriptive statistics are provided in Table 3. The perceived authenticity variables were rated as very high for SPs and relatively high for VPs. Cognitive load variables were reported to be moderate in both assessment methods. The average diagnostic accuracy was medium. The quantity of evidence generation was higher for SPs than for VPs. The quality of evidence generation was medium for both assessment methods. Motivational variables were rated rather highly for both SPs and VPs. A post hoc comparison showed that the value aspect of motivation was higher for SPs than for VPs (2-tailed t_{83}=2.89; P=.01; Cohen d=0.31), whereas the expectancy aspect did not differ between assessment methods (2-tailed t_{83}=0.44; P=.66; Cohen d=0.05). Participants demonstrated slightly above medium performance on the conceptual and strategic knowledge tests. Multimedia Appendix 5 provides an additional visualization of the results using boxplots and bee swarm plots.

Table 3. Descriptive statistics.

Variable	Both methods, mean (SD)	SPs[a], mean (SD)	VPs[b], mean (SD)
Perceived authenticity[c]	3.62 (0.67)	4.02 (0.67)	3.23 (0.84)
Realness[c]	3.71 (0.79)	4.13 (0.74)	3.28 (1.07)
Involvement[c]	3.82 (0.66)	4.03 (0.73)	3.61 (0.83)
Spatial presence[c]	3.35 (0.80)	3.89 (0.83)	2.80 (1.05)
Cognitive load[c]	2.88 (0.61)	2.88 (0.74)	2.90 (0.69)
Intrinsic load[c]	3.18 (0.68)	3.20 (0.78)	3.14 (0.80)
Extraneous load[c]	2.84 (0.65)	2.82 (0.79)	2.87 (0.76)
Germane load[c]	2.74 (0.76)	2.73 (0.88)	2.76 (0.84)
Diagnostic competences			
Diagnostic accuracy[d]	0.46 (0.18)	0.51 (0.28)	0.41 (0.24)
Quantity of evidence generation	22.26 (4.88)	29.01 (8.03)	17.34 (4.21)
Quality of evidence generation[d]	0.40 (0.11)	0.37 (0.18)	0.43 (0.13)
Control variables			
Motivation expectancy aspect[e]	5.07 (0.91)	5.10 (0.88)	5.05 (1.08)
Motivation value aspect[c]	4.44 (0.51)	4.54 (0.54)	4.34 (0.67)
Conceptual knowledge[d]	0.65 (0.14)	—[f]	—
Strategic knowledge[d]	0.66 (0.15)	—	—

[a] SP: standardized patient.

[b] VP: virtual patient.

[c] Scale range: 1-5.

[d] Scale range: 0-1.

[e] Scale range: 1-7.

[f] Knowledge was assessed before taking part in SPs and VPs.

Perceived Authenticity and Diagnostic Accuracy (RQ1)

A paired sample *t* test demonstrated that in line with hypothesis H1.1, perceived authenticity was considered higher for SPs than VPs in terms of the combined score (1-tailed t_{81}=11.12; *P*<.001; Cohen *d*=1.23). Post hoc tests showed that this was also the case for realness (t_{80}=8.83; *P*<.001; Cohen *d*=0.98), involvement (t_{81}=4.60; *P*<.001; Cohen *d*=0.51), and spatial presence (t_{79}=10.65; *P*<.001; Cohen *d*=1.19). Our expectation in H1.2 was that perceived authenticity would not be meaningfully associated with diagnostic accuracy. The TOST procedure for correlations showed that the relationship between diagnostic accuracy and the combined perceived authenticity score (*r*=0.05; *P*=.09) was outside the equivalence bounds of a meaningful effect of *r*=±0.20. Post hoc equivalence tests demonstrated that this also holds for the relationship of diagnostic accuracy with realness (*r*=0.03; *P*=.06), involvement (*r*=0.07; *P*=.11), and spatial presence (*r*=0.05; *P*=.08). Reanalyzing these correlations with regular 1-tailed NHST tests also yielded nonsignificant results for the combined score (*P*=.32), realness (*P*=.39), involvement (*P*=.28), and spatial presence (*P*=.33). These results mean that there is neither evidence for the absence of meaningful correlations nor evidence for significant correlations. These inconclusive findings may stem from the lack of statistical power because of the relatively small sample size [45].

Cognitive Load and Diagnostic Accuracy (RQ2)

We hypothesized in H2.1 that we would find equivalent cognitive load scores for SPs and VPs. Equivalence testing with the TOST procedure for paired samples indicated that for both assessment methods, the scores for combined cognitive load (t_{82}=2.81; *P*=.003) were significantly within the equivalence bounds of an effect of Cohen *d*=0.35. Adjusted post hoc equivalence tests showed that this is also the case for intrinsic load (t_{82}=−2.47; *P*=.008), extraneous load (t_{82}=2.55; *P*=.01), and germane load (t_{82}=2.64; *P*=.01). We expected in H2.2-H2.3 to uncover negative correlations between diagnostic accuracy and intrinsic cognitive load and extraneous load. As assumed, intrinsic cognitive load (1-tailed *r*=−0.30; *P*=.003) and extraneous load (1-tailed *r*=−0.29; *P*=.003) correlated negatively with the combined score for diagnostic accuracy. Adjusted explorative follow-up analyses showed that germane load (*r*=−0.25; *P*=.010) and the total score for cognitive load

(r=−0.31; P=.004) also correlated negatively with the combined score for diagnostic accuracy.

Assessment Method and Diagnostic Competences (RQ3)

Diagnostic Accuracy

In H3.1, we hypothesized finding equivalent diagnostic accuracy scores for SPs and VPs. H3.1 was first examined by applying a paired samples TOST procedure. According to our data, we cannot reject hypothesis H3.1 that a difference in diagnostic accuracy of at least ±0.125 points (1 grade) exists between the 2 assessment methods (t_{85}=–0.60; P=.28). A follow-up 3-way mixed design analysis of variance demonstrated that neither the CG order nor the assessment method order ($F_{3,82}$=2.49; P=.12; η^2=0.03, respectively, $F_{3,82}$=0.02; P=.88; η^2=0.01) had a significant effect on diagnostic accuracy. The assessment method itself, however, had a significant main effect ($F_{3,82}$=6.30; P=.01; η^2=0.07), indicating that diagnostic accuracy was higher for SPs than for VPs. The finding that diagnostic accuracy was higher for SPs than for VPs also corresponds to the result of a paired sample t test (2-tailed t_{85}=2.49; P=.01; Cohen d=0.27).

Evidence Generation

H3.2 that students display an increased quantity of evidence generation with SPs than with VPs was supported (1-tailed t_{69}=12.26; P<.001; Cohen d=1.47). However, in an explorative follow-up analysis, we found no evidence that the *quantity* of evidence generation was related to diagnostic accuracy (1-tailed r=0.11; P=.15). This finding holds equally for SPs (r=−0.09; P=.76) and VPs (r=−0.10; P=.82). Moreover, H3.3 that the *quality* of evidence generation is positively related to diagnostic accuracy in both assessment methods was not supported (1-tailed r=0.18; P=.05). Corrected post hoc analyses showed, however, that the quality of evidence generation was positively related to diagnostic accuracy for VPs (r=0.38; P<.001); this finding did not hold for SPs (r=0.05; P=.32). Additional post hoc exploratory analyses revealed that the quality of evidence generation was higher for VPs than for SPs (2-tailed t_{74}=–2.47; P=.02; Cohen d=0.29).

Discussion

Principal Findings

With regard to perceived authenticity, our results showed that SPs and VPs achieved high scores on all 3 dimensions of realness, involvement, and spatial presence. Despite this high level of perceived authenticity in both assessment methods, perceived authenticity was higher for SPs than for VPs on all 3 dimensions. This finding is in line with the literature, which has long claimed that SPs achieve a very high level of perceived authenticity [18-20]. Other studies on perceived authenticity have so far focused on comparing formats such as SPs, video presentations, and text vignettes and different levels of authenticity within VPs [21]. Our study extends this literature by directly comparing SPs and VPs with respect to 3 frequently used perceived authenticity variables. This comparison seems particularly relevant, as both assessment formats are becoming increasingly popular. Our findings on the relationship between perceived authenticity and diagnostic accuracy are mixed. The equivalence test on correlations was not significant; therefore, we could not confirm the hypothesis that perceived authenticity is not meaningfully associated with diagnostic accuracy. However, a regular correlation between perceived authenticity and diagnostic accuracy that was calculated afterward was close to 0. Taken together, these findings of nonequivalence and nonsignificance indicate that we did not have sufficient power to draw a conclusion [45]. Nevertheless, we have found some indication that the correlation between perceived authenticity and diagnostic competences is rather small. This finding is in accordance with literature reviews [23,24], which reported small correlations between perceived authenticity and performance.

With regard to cognitive load, we found that the combined score is equivalent for SPs and VPs that use the same clinical cases. This finding substantiates the literature suggesting that cognitive load depends mainly on task complexity [29]. Moreover, the fact that the extraneous load was equivalent for SPs and VPs indicates that user interaction through a software menu does not substantially increase cognitive load. This finding is important because decreasing the cognitive load by allowing for user input using natural language processing [21] is still highly expensive. Our study also adds to the literature that the level of cognitive load is similar in SPs and VPs as assessment methods if the different types of cognitive load are systematically controlled for during the design process. In addition, we demonstrated that intrinsic and extraneous cognitive loads correlate negatively with diagnostic accuracy. The finding on intrinsic cognitive load corroborates that the interplay between materials and the assessed person's expertise is associated with performance. The finding on extraneous cognitive load shows that unnecessary characteristics of the assessment environment can strain memory and attention and be detrimental to performance in assessment settings. Together, these findings fit well with the literature, which has repeatedly reported negative effects of intrinsic and extraneous cognitive loads on complex problem solving in medical education [27] and other domains [51]. Our study unveils that a negative relationship between intrinsic and extraneous cognitive loads and performance in a simulation-based measure of diagnostic competences already shows when overall cognitive load is medium on average.

Our study found no evidence that diagnostic accuracy was equivalent for SPs and VPs. In contrast, higher diagnostic accuracy was achieved for SPs than for VPs. The small number of studies comparing both assessment methods so far [1,31,32] have reported medium correlations, not taking into account different case content or testing time. Using the TOST procedure as a novel methodological approach, our study contributes to the literature by finding that grading was not equivalent, as participants received a better hypothetical grade when the simulation-based assessment was administered with SPs than with VPs. On the one hand, we cannot rule out that this finding may be explained by additional support from the actors in the SP assessment. To avoid and mitigate such an effect, actors were trained by an acting coach and a physician, memorized their symptoms and scripts, and did not know the diagnosis of their case. Moreover, student assistants screened all SP assessments, and no additional systematic support by actors was discovered. On the other hand, this finding can be explained by the lower appraisal of motivational value and the lower quantity of evidence generation reported for VPs. Participants solving VP cases may thus have been less engaged and may have collected a smaller number of important diagnostic cues that supported their diagnostic process.

Contrary to our expectations, the quality of evidence generation was not positively correlated with the *combined* diagnostic accuracy score. Closer inspection of the data revealed that the quality of evidence generation was positively correlated with diagnostic accuracy in *VPs*. This confirmed relationship is in line with the theoretical assumptions of Heitzmann et al [10]. In *SPs*, however, the quality of evidence was not correlated with diagnostic accuracy. This finding contradicts the theoretical assumptions of Heitzmann et al [10] and empirical results from studies using observational checklists with SPs [34] and real patients [36]. There are 2 explanations for these conflicting findings. First, the quality of evidence generation was, as an exploratory follow-up *t* test indicated, higher in VPs than in SPs. This higher quality of evidence generation could have been caused by a slightly different process of history taking in both assessment methods. Participants working with VPs selected questions from a menu. In contrast, participants working with SPs formulated questions during history taking freely. Second, SPs could have offered additional support to assessed persons who displayed a low quality of evidence generation, whereas VPs reacted in a completely standardized way to all assessed persons.

Limitations

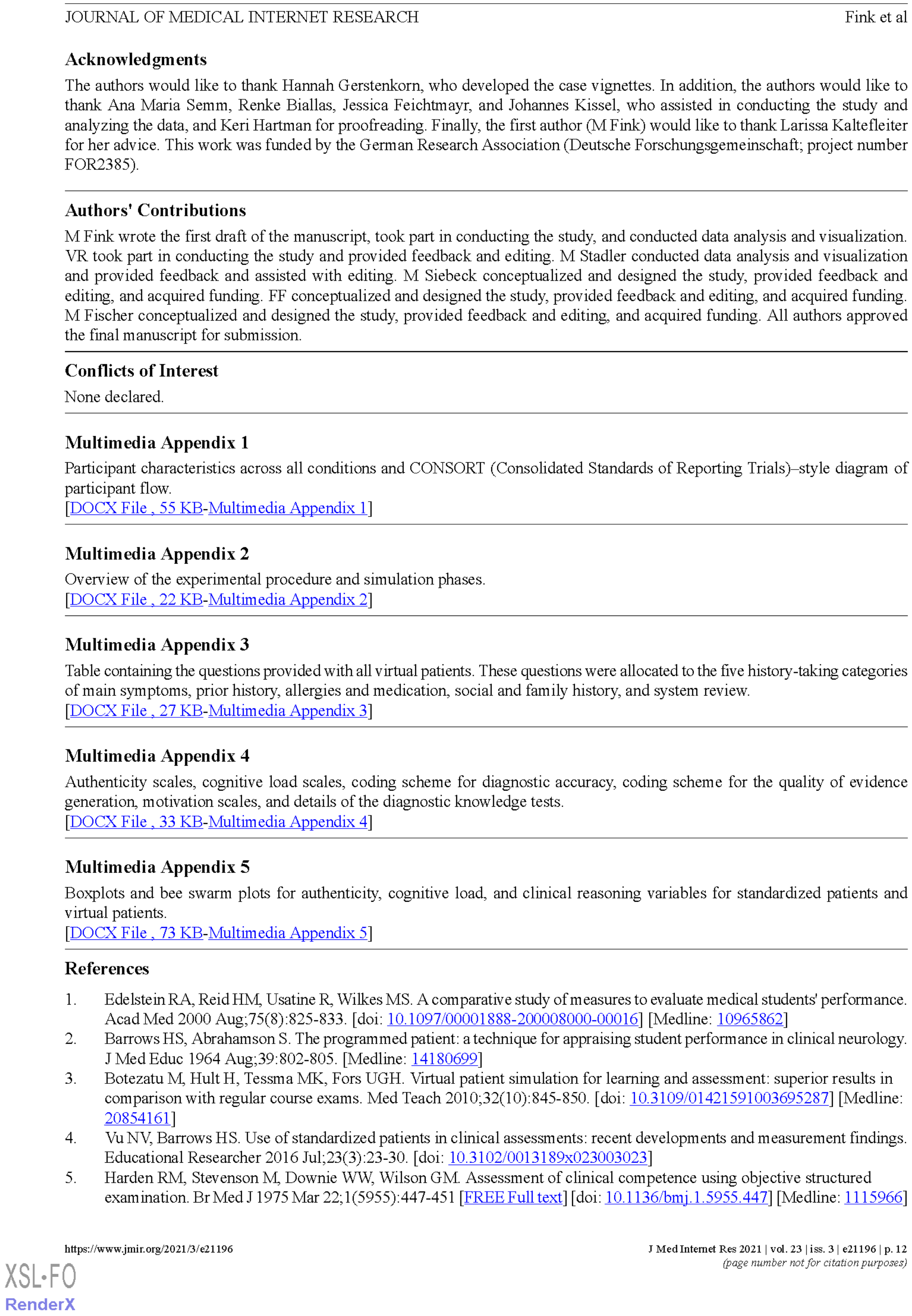
One methodological limitation of our study might be the low statistical power for the analysis of hypothesis H1.2 and related post hoc analyses that addressed the relationship between the perceived authenticity variables and diagnostic accuracy. This lack of statistical power can primarily be attributed to our investigation of whether a correlation of r=±0.20 or more extreme exists. As recommended by Lakens [46], the smallest effect size of interest was selected based on findings from the literature. Specifying the smallest effect size of interest to be larger would have increased power but not have contributed findings from a valuable equivalence test to the literature. This is the case because the literature already assumes a small effect size [23,24].

One theoretical limitation of the study is that the results on perceived authenticity may not generalize without restrictions to other related concepts of authenticity. Shaffer et al [15] argue that thick authenticity consists of four different aspects. An authentic task, situation, or material should (1) exist in real life, (2) be meaningful, (3) allow the learner to engage in professional activities of the discipline, and (4) be conducted rather similar in instruction and assessment. The authors assume that thick authenticity can only be achieved when all aspects of authenticity are adequate and that VPs could potentially achieve similar authenticity to SPs. Hamstra et al [16] proposed distinguishing fidelity using the terms physical resemblance and functional task alignment. The authors report weak evidence for the relationship between physical resemblance and performance, and strong evidence for the relationship between functional task alignment and performance. In our study, the concepts of thick authenticity and fidelity were not measured for two reasons. First, these concepts can, to some extent, only be judged externally by experts. Second, the repeated measures design of the study forced us to keep aspects such as thick authenticity, physical resemblance, and functional task alignment as similar as possible in SPs and VPs. Nevertheless, we believe that the relationship between different authenticity concepts and diagnostic competences still requires further research. Future studies should attempt to untangle the relationship between different authenticity concepts and diagnostic competences by measuring these systematically.

Conclusions

Our findings on the relationship between perceived authenticity and diagnostic accuracy contribute to the debate on the costs and benefits of perceived authenticity in performance-based assessments. These results relativize the importance of perceived authenticity in assessment. Increasing the perceived authenticity of assessment methods above a certain necessary threshold and thus raising their costs [23] does not seem to be of much benefit. Such spending could potentially squander a large share of the medical education budget [52] that could be put to more valuable use. Our results on cognitive load highlight its importance as a process variable in assessment settings. Performance-based assessment should thus attempt to reduce extraneous load and control for intrinsic load to measure performance in a standardized way that is still close to clinical practice [53].

Finally, the findings on diagnostic competences have some practical implications if VPs are used as an alternative to SPs in assessment. In particular, we found that VPs could lead to lower diagnostic accuracy scores than SPs, which could, in turn, negatively affect students' grades. There are 2 different mechanisms that could explain this finding: assessment with SPs could overestimate true performance or assessment with VPs could underestimate true performance. In accordance with SPs overestimating performance, we could not rule out additional support from the actors. In fact, the low, nonsignificant correlation between the quality of evidence generation and diagnostic accuracy in SPs, together with the higher diagnostic accuracy in SPs, could indicate that actors provided some additional support (eg, to participants who displayed low quality of evidence generation). Careful training [54] and screening thus seem to be of great importance to avoid additional support from actors during SP assessment to match the high level of standardization that VPs provide. The mechanism of possible underestimation of performance with VPs could be substantiated by the lower motivational value and quantity of evidence generation discovered for VPs. We suggest taking the following measures: students could be motivated additionally in VP assessment by more interactive environments (eg, using natural language processing) or providing automated elaborated feedback directly after the assessment. Moreover, the assessment time can be extended when menu-based VPs are used in practice. This way, the quantity of evidence generation could be raised to a level similar to that in the SP assessment.



Acknowledgments

The authors would like to thank Hannah Gerstenkorn, who developed the case vignettes. In addition, the authors would like to thank Ana Maria Semm, Renke Biallas, Jessica Feichtmayr, and Johannes Kissel, who assisted in conducting the study and analyzing the data, and Keri Hartman for proofreading. Finally, the first author (M Fink) would like to thank Larissa Kaltefleiter for her advice. This work was funded by the German Research Association (Deutsche Forschungsgemeinschaft; project number FOR2385).

Authors' Contributions

M Fink wrote the first draft of the manuscript, took part in conducting the study, and conducted data analysis and visualization. VR took part in conducting the study and provided feedback and editing. M Stadler conducted data analysis and visualization and provided feedback and assisted with editing. M Siebeck conceptualized and designed the study, provided feedback and editing, and acquired funding. FF conceptualized and designed the study, provided feedback and editing, and acquired funding. M Fischer conceptualized and designed the study, provided feedback and editing, and acquired funding. All authors approved the final manuscript for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Participant characteristics across all conditions and CONSORT (Consolidated Standards of Reporting Trials)–style diagram of participant flow.
[DOCX File , 55 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Overview of the experimental procedure and simulation phases.
[DOCX File , 22 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Table containing the questions provided with all virtual patients. These questions were allocated to the five history-taking categories of main symptoms, prior history, allergies and medication, social and family history, and system review.
[DOCX File , 27 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Authenticity scales, cognitive load scales, coding scheme for diagnostic accuracy, coding scheme for the quality of evidence generation, motivation scales, and details of the diagnostic knowledge tests.
[DOCX File , 33 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Boxplots and bee swarm plots for authenticity, cognitive load, and clinical reasoning variables for standardized patients and virtual patients.
[DOCX File , 73 KB-Multimedia Appendix 5]

References

1. Edelstein RA, Reid HM, Usatine R, Wilkes MS. A comparative study of measures to evaluate medical students' performance. Acad Med 2000 Aug;75(8):825-833. [doi: 10.1097/00001888-200008000-00016] [Medline: 10965862]
2. Barrows HS, Abrahamson S. The programmed patient: a technique for appraising student performance in clinical neurology. J Med Educ 1964 Aug;39:802-805. [Medline: 14180699]
3. Botezatu M, Hult H, Tessma MK, Fors UGH. Virtual patient simulation for learning and assessment: superior results in comparison with regular course exams. Med Teach 2010;32(10):845-850. [doi: 10.3109/01421591003695287] [Medline: 20854161]
4. Vu NV, Barrows HS. Use of standardized patients in clinical assessments: recent developments and measurement findings. Educational Researcher 2016 Jul;23(3):23-30. [doi: 10.3102/0013189x023003023]
5. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975 Mar 22;1(5955):447-451 [FREE Full text] [doi: 10.1136/bmj.1.5955.447] [Medline: 1115966]

6. Ziv A. Simulators and simulation-based medical education. In: A practical guide for medical teachers. Vol. 2. 3rd ed. Amsterdam: Elsevier; 2009.
7. de JT. Instruction based on computer simulations. In: Handbook of research on learning and instruction. New York: Routledge; 2011:446-466.
8. Villaume WA, Berger BA, Barker BN. Learning motivational interviewing: scripting a virtual patient. Am J Pharm Educ 2006 Apr 15;70(2):33 [FREE Full text] [doi: 10.5688/aj700233] [Medline: 17149413]
9. Setrakian J, Gauthier G, Bergeron L, Chamberland M, St-Onge C. Comparison of assessment by a virtual patient and by clinician-educators of medical students' history-taking skills: exploratory descriptive study. JMIR Med Educ 2020 Mar 12;6(1):14428 [FREE Full text] [doi: 10.2196/14428] [Medline: 32163036]
10. Heitzmann N, Seidel T, Hetmanek A, Wecker C, Fischer MR, Ufer S, et al. Facilitating diagnostic competences in simulations in higher education a framework and a research agenda. Frontline Learning Research 2019 Dec 3:1-24. [doi: 10.14786/flr.v7i4.384]
11. Kopp V, Stark R, Fischer MR. Fostering diagnostic knowledge through computer-supported, case-based worked examples: effects of erroneous examples and feedback. Med Educ 2008 Aug;42(8):823-829. [doi: 10.1111/j.1365-2923.2008.03122.x] [Medline: 18564096]
12. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Review of Educational Research 2020 Jun 15;90(4):499-541. [doi: 10.3102/0034654320933544]
13. Cook DA, Brydges R, Hamstra SJ, Zendejas B, Szostek JH, Wang AT, et al. Comparative effectiveness of technology-enhanced simulation versus other instructional methods: a systematic review and meta-analysis. Simul Healthc 2012 Oct;7(5):308-320. [doi: 10.1097/SIH.0b013e3182614f95] [Medline: 23032751]
14. Schubert T, Friedmann F, Regenbrecht H. The experience of presence: factor analytic insights. Presence: Teleoperators & Virtual Environments 2001 Jun;10(3):266-281. [doi: 10.1162/105474601300343603]
15. Shaffer DW, Resnick M. Thick authenticity: new media and authentic learning. J Interact Learn Res 1999;10(2):195-216 [FREE Full text]
16. Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA. Reconsidering fidelity in simulation-based training. Acad Med 2014 Mar;89(3):387-392 [FREE Full text] [doi: 10.1097/ACM.0000000000000130] [Medline: 24448038]
17. Hofer M. Presence und involvement. 1st ed. Baden-Baden: Nomos; 2016:978-973.
18. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Acad Med 1993 Jun;68(6):443-451. [doi: 10.1097/00001888-199306000-00002] [Medline: 8507309]
19. Luctkar-Flude M, Wilson-Keates B, Larocque M. Evaluating high-fidelity human simulators and standardized patients in an undergraduate nursing health assessment course. Nurse Educ Today 2012 May;32(4):448-452. [doi: 10.1016/j.nedt.2011.04.011] [Medline: 21565436]
20. Rethans JJ, Sturmans F, Drop R, van der Vleuten C. Assessment of the performance of general practitioners by the use of standardized (simulated) patients. Br J Gen Pract 1991 Mar;41(344):97-99 [FREE Full text] [Medline: 2031767]
21. Friedman CP, France CL, Drossman DD. A randomized comparison of alternative formats for clinical simulations. Med Decis Making 1991;11(4):265-272. [doi: 10.1177/0272989X9101100404] [Medline: 1766329]
22. Padgett J, Cristancho S, Lingard L, Cherry R, Haji F. Engagement: what is it good for? The role of learner engagement in healthcare simulation contexts. Adv Health Sci Educ Theory Pract 2019 Oct;24(4):811-825. [doi: 10.1007/s10459-018-9865-7] [Medline: 30456474]
23. Norman G, Dore K, Grierson L. The minimal relationship between simulation fidelity and transfer of learning. Med Educ 2012 Jul;46(7):636-647. [doi: 10.1111/j.1365-2923.2012.04243.x] [Medline: 22616789]
24. Schoenherr JR, Hamstra SJ. Beyond fidelity: deconstructing the seductive simplicity of fidelity in simulator-based education in the health care professions. Simul Healthc 2017 Apr;12(2):117-123. [doi: 10.1097/SIH.0000000000000226] [Medline: 28704289]
25. La Rochelle JS, Durning SJ, Pangaro LN, Artino AR, van der Vleuten CPM, Schuwirth L. Authenticity of instruction and student performance: a prospective randomised trial. Med Educ 2011 Aug;45(8):807-817. [doi: 10.1111/j.1365-2923.2011.03994.x] [Medline: 21752077]
26. Sweller J, van Merrienboer JJG, Paas FGWC. Cognitive architecture and instructional design. Educational Psychology Review 1998;10(3):251-296. [doi: 10.1023/a:1022193728205]
27. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. Med Teach 2014 May;36(5):371-384. [doi: 10.3109/0142159X.2014.889290] [Medline: 24593808]
28. Dankbaar MEW, Alsma J, Jansen EEH, van Merrienboer JJG, van Saase JLCM, Schuit SCE. An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. Adv Health Sci Educ Theory Pract 2016 Aug;21(3):505-521 [FREE Full text] [doi: 10.1007/s10459-015-9641-x] [Medline: 26433730]
29. Haji FA, Cheung JJH, Woods N, Regehr G, de Ribaupierre S, Dubrowski A. Thrive or overload? The effect of task complexity on novices' simulation-based learning. Med Educ 2016 Sep;50(9):955-968. [doi: 10.1111/medu.13086] [Medline: 27562895]
30. Young M, Thomas A, Lubarsky S, Ballard T, Gordon D, Gruppen LD, et al. Drawing boundaries: the difficulty in defining clinical reasoning. Acad Med 2018 Jul;93(7):990-995. [doi: 10.1097/ACM.0000000000002142] [Medline: 29369086]

31. Guagnano MT, Merlitti D, Manigrasso MR, Pace-Palitti V, Sensi S. New medical licensing examination using computer-based case simulations and standardized patients. Acad Med 2002 Jan;77(1):87-90. [doi: 10.1097/00001888-200201000-00020] [Medline: 11788331]
32. Hawkins R, MacKrell Gaglione M, LaDuca T, Leung C, Sample L, Gliva-McConvey G, et al. Assessment of patient management skills and clinical skills of practising doctors using computer-based case simulations and standardised patients. Med Educ 2004 Sep;38(9):958-968. [doi: 10.1111/j.1365-2929.2004.01907.x] [Medline: 15327677]
33. Hu L, Chen G, Li P, Huang J. Retracted article: multimedia effect in problem solving: a meta-analysis. Educ Psychol Rev 2019 Jul 11;32(3):901. [doi: 10.1007/s10648-019-09490-4]
34. Stillman PL, Swanson DB, Smee S, Stillman AE, Ebert TH, Emmel VS, et al. Assessing clinical skills of residents with standardized patients. Ann Intern Med 1986 Nov;105(5):762-771. [doi: 10.7326/0003-4819-105-5-762] [Medline: 3767153]
35. Woolliscroft JO, Calhoun JG, Billiu GA, Stross JK, MacDonald M, Templeton B. House officer interviewing techniques: impact on data elicitation and patient perceptions. J Gen Intern Med 1989;4(2):108-114. [doi: 10.1007/BF02602349] [Medline: 2709168]
36. Casus computer software. 2018. URL: https://www.instruct.eu/en/ [accessed 2021-01-30]
37. Bornemann BM. Documentation forms of internal medicine and surgery for history taking and the physical examination for the medical training of students in Germany: An analysis of content and structure. Diss. München: Institut für Didaktik und Ausbildungsforschung in der Medizin der Ludwig-Maximilians-Universität München; 2016. URL: https://edoc.ub.uni-muenchen.de/19166/1/Bornemann_Barbara.pdf [accessed 2021-02-16]
38. Seidel T, Stürmer K, Blomberg G, Kobarg M, Schwindt K. Teacher learning from analysis of videotaped classroom situations: does it make a difference whether teachers observe their own teaching or that of others? Teaching and Teacher Education 2011 Feb;27(2):259-267. [doi: 10.1016/j.tate.2010.08.009]
39. Vorderer P, Wirth W, Gouveia F, Biocca F, Saari T, Jäncke F, et al. MEC spatial presence questionnaire (MEC-SPQ): short documentation and instructions for application. Report to the European Community, Project Presence: MEC (IST--37661). 2001. URL: http://www.ijk.hmt-hannover.de/presence [accessed 2021-01-30]
40. Frank B. Validation. Measuring Presence in Laboratory-Based Research with Microworlds 2014:51-61. [doi: 10.1007/978-3-658-08148-5_6]
41. Opfermann M. There's more to it than instructional design: the role of individual learner characteristics for hypermedia learning. Berlin: Logos Verlag; 2008:1-295.
42. Rheinberg F, Vollmeyer R, Burns BD. FAM: Ein fragebogen zur erfassung aktueller motivation in lern- und leistungssituationen. Diagnostica 2001 Apr;47(2):57-66. [doi: 10.1026//0012-1924.47.2.57]
43. Wigfield A. Expectancy-value theory of achievement motivation: a developmental perspective. Educ Psychol Rev 1994 Mar;6(1):49-78. [doi: 10.1007/bf02209024]
44. Schmidmaier R, Eiber S, Ebersbach R, Schiller M, Hege I, Holzer M, et al. Learning the facts in medical school is not enough: which factors predict successful application of procedural knowledge in a laboratory setting? BMC Med Educ 2013 Mar 22;13(1):28 [FREE Full text] [doi: 10.1186/1472-6920-13-28] [Medline: 23433202]
45. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: a tutorial. Advances in Methods and Practices in Psychological Science 2018 Jun 01;1(2):259-269. [doi: 10.1177/2515245918770963]
46. Lakens D. Equivalence tests: a practical primer for tests, correlations, and meta-analyses. Soc Psychol Personal Sci 2017 May;8(4):355-362 [FREE Full text] [doi: 10.1177/1948550617697177] [Medline: 28736600]
47. R Foundation for statistical computing. R [Computer software]. Vienna, Austria: R Foundation for Statistical Computing; 2019. URL: https://www.r-project.org/ [accessed 2021-02-16]
48. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian J Stat. 1979. URL: https://www.jstor.org/stable/4615733?seq=1 [accessed 2021-02-16]
49. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale: Lawrence Erlbaum; 1988.
50. G*Power computer software. 2014. URL: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html [accessed 2021-01-30]
51. Sweller J, van Merriënboer JJG, Paas F. Cognitive architecture and instructional design: 20 years later. Educ Psychol Rev 2019 Jan 22;31(2):261-292. [doi: 10.1007/s10648-019-09465-5]
52. Lapkin S, Levett-Jones T. A cost-utility analysis of medium vs. high-fidelity human patient simulation manikins in nursing education. J Clin Nurs 2011 Dec;20(23-24):3543-3552. [doi: 10.1111/j.1365-2702.2011.03843.x] [Medline: 21917033]
53. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990 Sep;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
54. Lewis KL, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, et al. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). Adv Simul (Lond) 2017;2:10 [FREE Full text] [doi: 10.1186/s41077-017-0043-4] [Medline: 29450011]

Abbreviations

CG: case group

NHST: null hypothesis significance testing
SP: standardized patient
TOST: 2 separate 1-sided test
VP: virtual patient

Edited by G Eysenbach, R Kukafka; submitted 25.06.20; peer-reviewed by J Cheung, P Bergl, A Kononowicz, S Edelbring; comments to author 08.08.20; revised version received 01.10.20; accepted 27.12.20; published 04.03.21

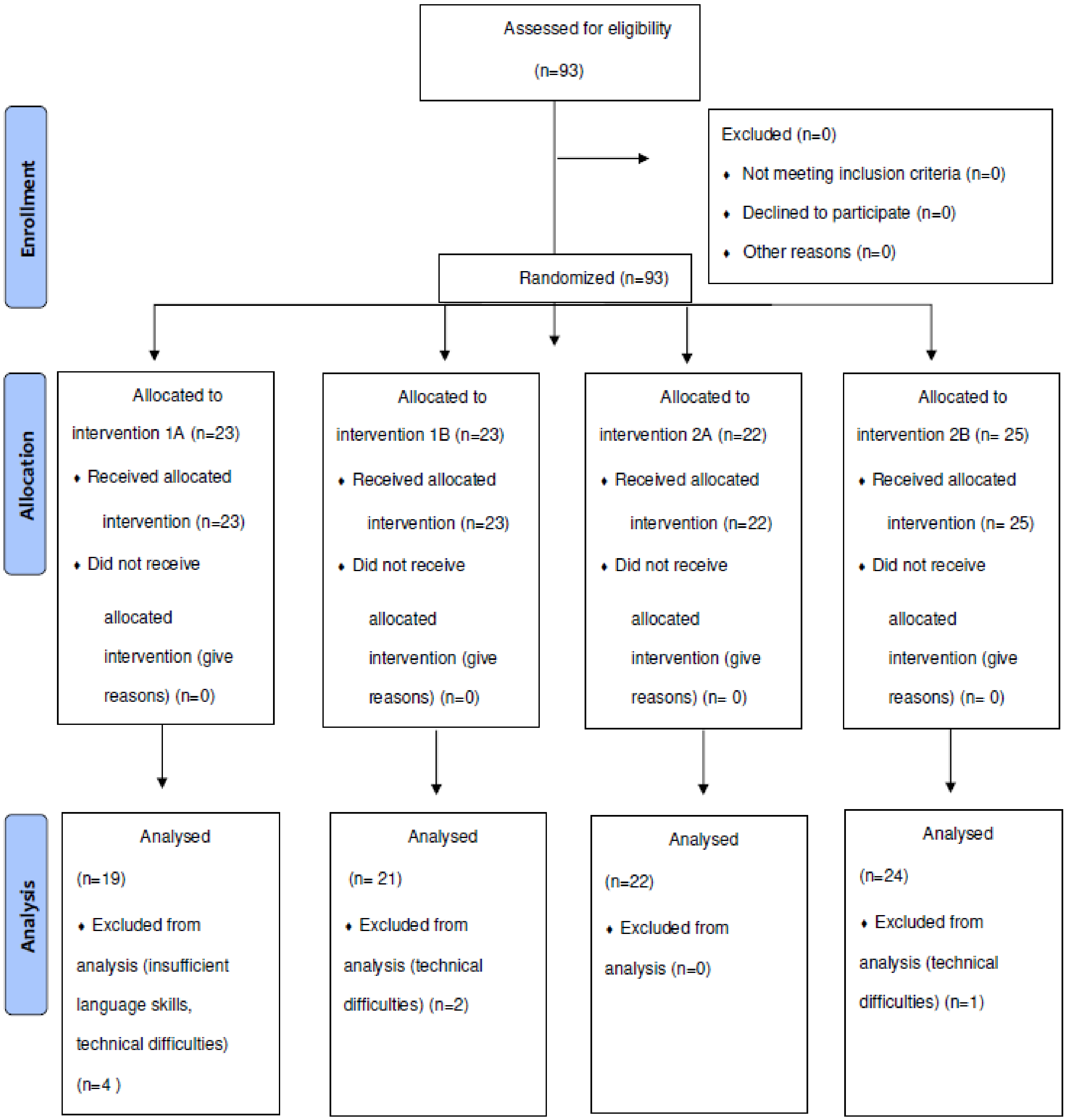
Please cite as:
Fink MC, Reitmeier V, Stadler M, Siebeck M, Fischer F, Fischer MR
Assessment of Diagnostic Competences With Standardized Patients Versus Virtual Patients: Experimental Study in the Context of History Taking
J Med Internet Res 2021;23(3):e21196
URL: https://www.jmir.org/2021/3/e21196
doi: 10.2196/21196
PMID: 33661122

Appendices

Multimedia Appendix 1. Participant characteristics across all conditions and CONSORT (Consolidated Standards of Reporting Trials)–style diagram of participant flow.

Table 1
Participant characteristics across all conditions.

	Condition 1A	Condition 1B	Condition 2A	Condition 2B
Age in years mean, (SD)	25.42 (3.63)	26.52 (4.45)	26.71 (2.92)	25.52 (6.63)
Sex, n (%)				
Females	11 (58)	11 (52)	14 (67)	18 (72)
Males	8 (42)	10 (48)	7 (33)	7 (28)
Expertise, n (%)				
Novices	10 (53)	12 (57)	11 (52)	14 (56)
Intermediates	9 (47)	9 (43)	10 (48)	11 (44)


Enrollment
Allocation
Analysis
Assessed for eligibility (n=93)
Excluded (n=0)
Not meeting inclusion criteria (n=0)
Declined to participate (n=0)
Other reasons (n=0)
Randomized (n=93)
Allocated to intervention 1A (n=23)
Received allocated intervention (n=23)
Did not receive allocated intervention (give reasons) (n=0)
Allocated to intervention 1B (n=23)
Received allocated intervention (n=23)
Did not receive allocated intervention (give reasons) (n=0)
Allocated to intervention 2A (n=22)
Received allocated intervention (n=22)
Did not receive allocated intervention (give reasons) (n= 0)
Allocated to intervention 2B (n= 25)
Received allocated intervention (n= 25)
Did not receive allocated intervention (give reasons) (n= 0)
Analysed (n=19)
Excluded from analysis (insufficient language skills, technical difficulties) (n=4)
Analysed (n= 21)
Excluded from analysis (technical difficulties) (n=2)
Analysed (n=22)
Excluded from analysis (n=0)
Analysed (n=24)
Excluded from analysis (technical difficulties) (n=1)

Multimedia Appendix 2. Overview of the experimental procedure and simulation phases.

Table 1

Overview of the Experiment

Part of the experiment	Activity / test	Duration in minutes
Pretest		
	Briefing	10
	Conceptual knowledge test	40
	Strategic knowledge test	40
Break		10
Simulation phase I (Case 1- 3)		70
Break and change of modality		5
Simulation phase II (Case 4-6)		70
Debriefing/posttest		
	Working memory test	20

Table 2

Overview of the Simulation Phases

Part of the experiment	Activity / test	Duration in minutes
Simulation briefing		10
	Fiction contract,	
	Familiarization with content	
	Familiarization with technical aspects	
	Interest scales	
	Motivation questionnaires	
Simulation phase (similar for all three cases)		60
	Presentation of chief complaint	
	Independent history taking	
	Diagnostic accuracy measurement	
	Authenticity scales	
	Cognitive load scales	
	Epistemic emotions scales	

Multimedia Appendix 3. Table containing the questions provided with all virtual patients. These questions were allocated to the five history-taking categories of main symptoms, prior history, allergies and medication, social and family history, and system review.

NR	Main symptoms[HS]	Prior history [MV]	Allergies and medication [AM]	Social and family history [SF]	Systems review [SUE]
01	Are you experiencing the complaints for the first time?	Do you know of any pre-existing conditions?	Do you have common or chronic infections for which you have to take antibiotics?	Did your parents or another one of your closer relatives die at a very young age?	Have you gained or lost any weight in recent weeks?
02	Do you suffer from pain?	Have you ever had surgery?	Do you take medication on a regular basis? Or do you maybe take special medication in specific situations? Something like medication for allergies or painkillers?	Has anyone in your family died a sudden cardiac death?	Do you have night sweats?
03	How are you doing right now?	Have you had surgery in recent weeks?	Have you noticed whether your eyes were twitching or your nose was running?	Do you smoke or did you use to smoke?	Have you eaten and drunk enough today?
04	Do you experience the complaints only during exertion or also at rest?	Has your mobility been limited, for instance by a plaster cast or through a disease that confined you to bed?	Do you have known allergies or asthma?	How much alcohol do you drink normally?	Have you noticed whether you were able to perform as well as usual?
05	Have you recently avoided leaving the house because you were afraid that something might happen?	Do you have a coagulation disorder?		Did you drink a lot of alcohol yesterday?	Are you sleeping well?
06	Did it occur suddenly or did you already experience dyspnea in recent days or weeks?	Have you had thrombosis before?		Have you completed occupational training? If so, for which occupation?	Do you have problems with your stool or with your urination?
07	Do you have a cough?	Do you suffer from high blood pressure?		Are you currently employed or do you own your own business?	Have you had bloody tarry stool or have you vomited blood recently?

08	Can you rest on a straight surface?	Have you had problems with your heart before?		Are you married?	Have you had fever or chills in the last few days?
09	Have you experienced occasional dizziness in the last few months? Or have you passed out?	Have you made use of psychotherapeutic treatment before?		Do you have children?	Do you have lip herpes at the moment?
10	Have you experienced fear of death?	Did you suffer from heart muscle inflammation as a child?		Do your parents or siblings have chronic diseases (e.g. high blood pressure, diabetes etc.)?	Do you have any pain in your arms or in the jaw area?
11	Have you experienced other symptoms? For instance, rapid heartbeat or dizziness? [Penultimate position within the category, excluded from analysis as it was only included in the “aimueller” case]	Do you suffer from a muscular disorder?		Have you taken a longer plane, bus, or car trip recently?	Do you feel any traction or tingling in your hands?
12	Were you chewing gum when it happened? [Very last position within the category, excluded from analysis as it was only included in the “aimueller” case]	Have your thyroid glands been checked for overactivity or underactivity?		Do you exercise regularly?	Have your legs gotten bigger?
13		Have you had an acute infection in the last few weeks? Have you had a cough or cold or something similar?		Have you taken drugs or an energy booster or something similar recently?	Did you had the feeling of a racing or stumbling heart in the past few days?
14		Do you go for regular check-ups with your GP?		Can you tell me how much you usually drink per day?	What does your urine look like? Have you noticed any unusual colour? Extremely bright, dark or brown or anything like that?

15		Do you know how high your blood pressure is usually?		Do you have siblings?	
16		Do you know whether any blood levels have been bad before? Cholesterol or similar values?		Are there any known hereditary diseases in your family?	
17		Have you had a malignant illness before (e.g. cancer or a tumor or something similar)?		Have you been under a lot of stress recently? [last position within the category, excluded from analysis as it was only included in some cases]	
18		Have you ever had a stroke?			
19		Have you been treated by a neurologist before?			
20		Have you ever had a pneumothorax or have you ever had lung surgery?			
21		Are you pregnant or have you given birth recently? [Last position within the category, excluded from analysis as it was only included in one case]			

Multimedia Appendix 4. Authenticity scales, cognitive load scales, coding scheme for diagnostic accuracy, coding scheme for the quality of evidence generation, motivation scales, and details of the diagnostic knowledge tests.

Table 1
Authenticity scales

Nr	Item
1	I consider the history-taking simulation as authentic.
2	The simulation of the medical interview seemed like a real professional demand.
3	The experience in the history-taking simulation resembled the experience of a real professional demand.
4	When I participated in history-taking, it seemed to me as if I was a real part of the simulated situation.
5	When I participated in history-taking, I felt like I was physically present in the clinical environment.
6	When I participated in history-taking, it seemed to me as if I could affect things, like in a real medical interview.
7	When I participated in history-taking, I focussed strongly on the
8	situation. When I participated in history-taking, I forgot intermittently that I
9	take part in a study.
10	When I participated in history-taking, I was immersed in the situation. When I participated in history-taking, I was fully engaged.

Note. Items 1-3 measure the subscale realness, items 4-6 spatial presence and items 7-10 presence. This measure was used as a 5-point scale, ranging from 1 (“Strongly disagree”) to 5 (“Strongly agree”).

Table 2
Cognitive load scales

Nr	Item
1	How easy or difficult do you consider „History taking for dyspnea" at this moment?
2	How easy or difficult is it for you to work with the simulation?
3	How easy or difficult is it for you to distinguish important and unimportant information in the simulation?
4	How easy or difficult is it for you to collect all the information that you need in the learning environment?
5	How easy or difficult was it to understand the last case?

Note. Item 1 measures intrinsic load, items 2-4 extraneous load and item 5 germane load. This measure was used as a 5-point scale, ranging from 1 (“easy”), 2 ("rather easy"), 3 ("neutral"), 4 ("rather hard"), and 5 ("hard").

Table 3
Coding Scheme for Diagnostic Accuracy for All Cases

Case	1 Point (Fully correct)	0.5 Points (Partially correct)	0 Points (Incorrect)
CGA_1	- Pulmonary embolism with lymphoma - Pulmonary embolism with prostate cancer	- Pulmonary embolism - Myelodysplastic syndrome	All other diagnoses from the long-menu
CGA_2	- Congestive heart failure with atrial fibrillation - Congestive heart failure with arrhythmia	- Left-sided heart failure - Acute decompensated heart failure - Right-sided heart failure - Dyspnea caused by pleural effusion - Acute coronary syndrome	All other diagnoses from the long-menu
CGA_3	- Hyperventilation tetany caused by panic attack - Hyperventilation tetany caused by panic disorder	- Panic attack - Panic disorder	All other diagnoses from the long-menu
CGB_1	- Pulmonary embolism with coagulation disorder - Pulmonary embolism with antiphospholipid syndrome - Pulmonary embolism with hereditary thrombophilia	- Pulmonary embolism	All other diagnoses from the long-menu
CGB_2	- Community-acquired pneumonia (CAP) -Bacterial pneumonia	- Acute bronchitis - Pneumonia	All other diagnoses from the long-menu
CGB_3	- Obstructive hypertrophic cardiomyopathy - Cardiac insufficiency with concentric left ventricular hypertrophy	- Cardiac insufficiency - Right-sided heart failure - Left-sided heart failure - Valvular heart disease - Dilated cardiomyopathy - Arrhythmogenic right ventricular cardiomyopathy - Myocarditis - Cardiomyopathy - Hypertrophic cardiomyopathy	All other diagnoses from the long-menu

Note. Points allocated were defined as follows. 1 Point: A solution determined ex-ante by the author was discovered (i.e., one type of the disease for which the case was created was listed). 0.5 Points: The listed diagnosis can be explained by the symptoms and clinical findings (e.g., from prior information and replies of the patient). 0 Points: An incorrect diagnosis is listed, a diagnosis with incorrect additional information is listed or no diagnosis is listed.

Table 4

Coding Scheme for the Quality of Evidence Generation

CGA_1	CGA_2	CGA_3	CGB_1	CGB_2	CGB_3
HS02	HS02	HS03	MV01	HS02	HS02
SUE01	HS03	HS05	MV05	HS03	HS09
SUE02	SUE06	HS09	SF16	MV13	MV07
SUE04	SUE13	HS10	HS02	SUE02	AM02
SUE05	SUE14	SUE01	HS03	SUE04	SF01
SUE06	SF14	SUE05		SUE08	SF02
SUE14		SUE11			SF12
		SUE13			SUE01
					SUE04
					SUE05
					SUE12
					SUE13

Note. Only items highly relevant for the case are marked with the corresponding code. The abbreviation CG corresponds to the case-group, the number to the case number.

Table 5

Motivation scales

Nr	Item
1	I believe I am up to the difficulty of this task.
2	I will probably not solve this task successfully (Reverse coded)
3	I believe everyone can solve this task.
4	Probably, I will not solve this task successfully (Reverse coded).
5	I believe it is important to be able to solve this task.
6	Even if this task is not part of examinations, it is important to be able to solve this task.
7	It would be useful to engage oneself with this task.
8	It would be useful to occupy oneself with this task, as it is generally useful to master this type of task.

Note. Items 1-4 measure the expectancy aspect of motivation, items 5-8 the value aspect of motivation. The scale for the expectancy aspect ranged from (1) strongly disagree to (7) strongly agree. The scale for the value aspect ranged from (1) strongly disagree to 5 strongly agree. The instruction used was: "We would like to know more about your current attitude towards the presented task. Please indicate the attitude that suits best to you."

Table 6
Details of the diagnostic knowledge tests

	Conceptual knowledge	Strategic knowledge
Content	40 multiple-choice questions on dyspnea with the answer formats single-choice and Pick-N	10 case vignettes on dyspnea with four single-choice questions each
Reliability	Cronbach α=.76.	Cronbach α=.81
Scoring	In single-choice questions, participants received 1.0 points for selecting the correct answer and 0 points for an incorrect answer. In Pick-N questions, participants were allocated 1.0 points for each entirely correct answer pattern and partial credit (0.50 points) if they provided at least 50 percent of the correct answers to a question.	
Total score	We divided the number of correct answers by the number of provided questions to calculate scores for the knowledge tests. Thus, both knowledge test scores ranged from 0 to 1.	

Multimedia Appendix 5. Boxplots and bee swarm plots for authenticity, cognitive load, and clinical reasoning variables for standardized patients and virtual patients.

Figure 1. Boxplots and beeswarmplots of authenticity variables in standardized patients and virtual patients. Means are highlighted with a dot in the box. Within beeswarmplots points visualize the distribution of the variable similar to a scatterplot.

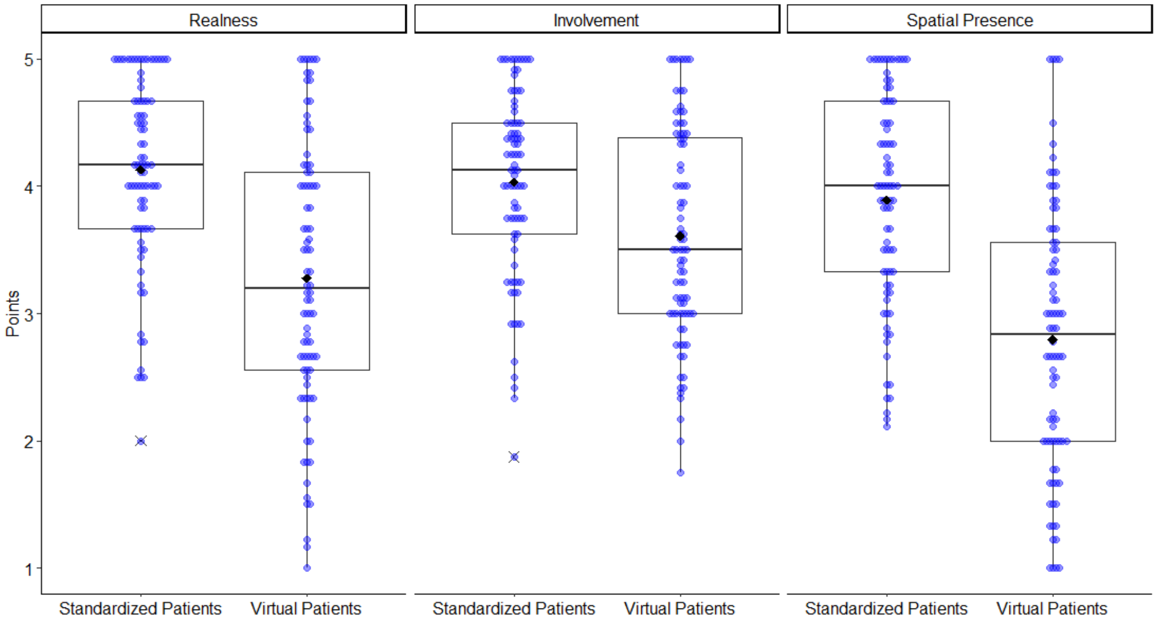


Figure 2. Boxplots and beeswarmplots of cognitive load variables in standardized patients and virtual patients. Means are highlighted with a dot in the box. Within beeswarmplots points visualize the distribution of the variable similar to a scatterplot.

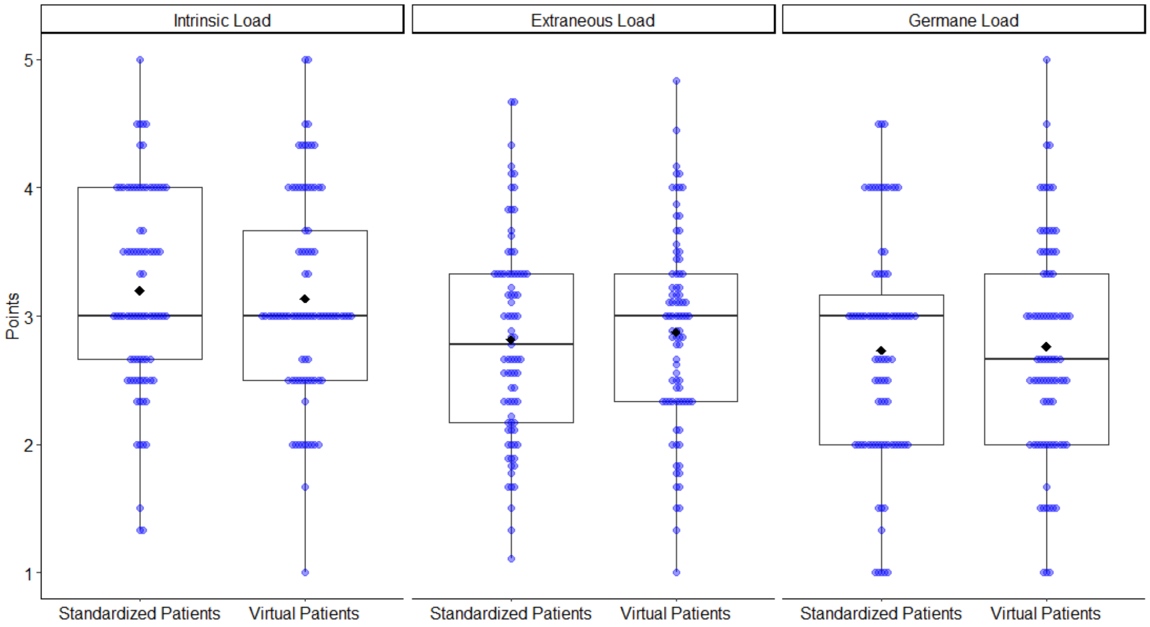
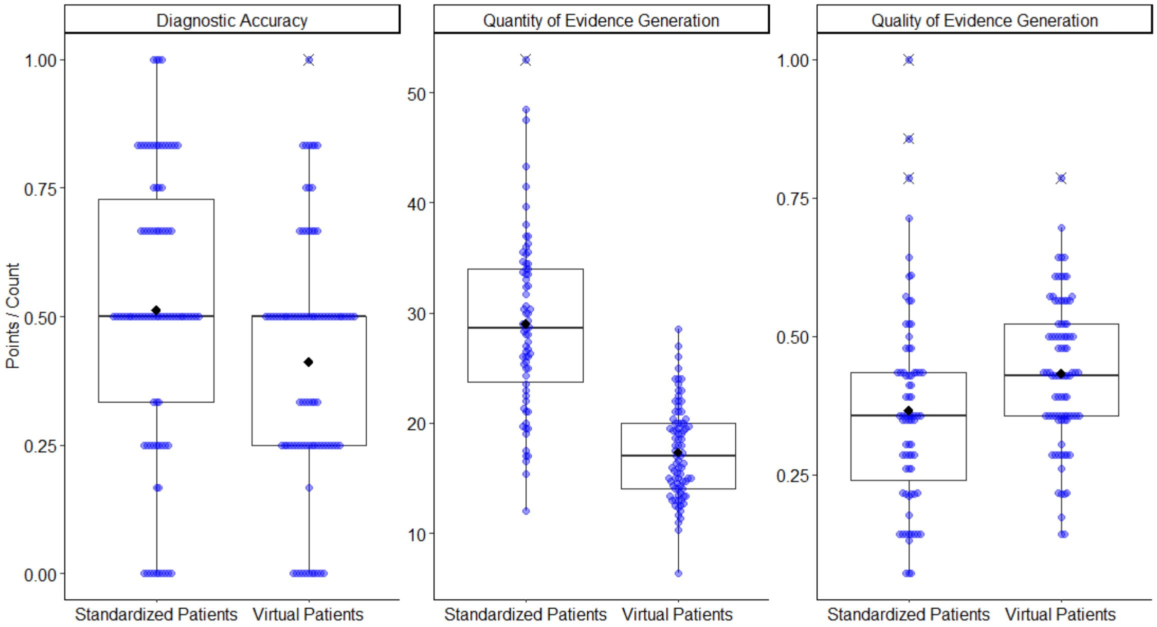


Figure 3. Boxplots and beeswarmplots of clinical reasoning variables in standardized patients and virtual patients. Means are highlighted with a dot in the box. Within beeswarmplots points visualize the distribution of the variable similar to a scatterplot.



Note
The R-packages ggplot (Wickham, 2016) and ggebeeswarm were utilized for data visualization.

3. Article 2: Diagnosing Virtual Patients with Partial Prior Knowledge: The Interplay between Knowledge and Diagnostic Activities

Reference

Fink, M. C., Heitzmann, N., Reitmeier, V., Siebeck, M., Fischer, F., & Fischer, M. R. (submitted). Diagnosing virtual patients with partial prior knowledge: The interplay between knowledge and diagnostic activities. Manuscript submitted for publication to Advances in Health Sciences Education, June 17, 2022.

Abstract

Theory: Clinical reasoning theories agree that knowledge and the diagnostic process are associated with diagnostic success. However, the exact contributions of these components of clinical reasoning to diagnostic success remain unclear. This is particularly the case when operationalizing the diagnostic process with diagnostic activities (i.e., teachable practices that generate knowledge) and for learners with partial prior knowledge. Therefore, we conducted a study investigating to what extent knowledge and diagnostic activities uniquely explain variance in diagnostic success with virtual patients among medical students with partial prior knowledge.

Method: The study sample consisted of *N* = 106 medical students in their third to fifth year of university studies in Germany (six-year curriculum). Participants completed professional knowledge tests before diagnosing virtual patients. Diagnostic success with the virtual patients was assessed with a comprehensive diagnostic score as well as diagnostic accuracy to answer the call for more extensive measurement of clinical reasoning outcomes. The three diagnostic activities hypothesis generation, evidence generation, and evidence evaluation were tracked.

Results: Professional knowledge predicted performance in terms of the comprehensive diagnostic score and displayed a small association with diagnostic accuracy. Diagnostic activities predicted comprehensive diagnostic score and diagnostic accuracy. Hierarchical regressions showed that the diagnostic activities made a unique contribution to diagnostic success, even when knowledge was taken into account.

Conclusions: Our results support the argument that the diagnostic process is more than merely an embodiment of knowledge. We discuss possible mechanisms explaining this finding, which may be restricted to diagnostic activities and learners with partial prior knowledge.

INTRODUCTION

Clinical reasoning is an extensive construct that consists of a variety of components (Young et al., 2018). Key components discussed in important clinical reasoning definitions and theories are knowledge, diagnostic processes, and outcome measures (Elstein, 2009; Heitzmann et al., 2019; Schmidt et al., 1990). Research on the relationships between the aforementioned components of clinical reasoning can contribute to improving our theoretical understanding of this construct. Moreover, research carried out on this topic can yield relevant insights for assessment methods and instructional support (Daniel et al., 2019; Heitzmann et al., 2017). This paper adds to research on the associations of key components of clinical reasoning. In doing so, it directs its attention to diagnostic activities (i.e., teachable practices that generate knowledge) as an operationalization of the diagnostic process for which empirical evidence is still largely lacking. Moreover, this paper focuses on medical students with partial prior knowledge. Such medical students were selected primarily because they likely do not yet possess rich knowledge networks that allow for diagnosing via pattern recognition (Schmidt and Rikers, 2007), and only a few findings on diagnostic processes are available for this group (Boulet and Durning, 2019).

Three main perspectives on clinical reasoning

Three main perspectives on clinical reasoning can be distinguished. First, knowledge-centered theories, such as illness script theory (Schmidt et al., 1990), assume that the amount, type, and structure of knowledge networks developed through formal training and practical medical experience are crucial for diagnosing in an automatic pattern recognition process (Charlin et al., 2007). Second, problem-solving theories, typically emphasizing the hypothetico-deductive method (Elstein et al., 1978; Elstein et al., 1990), argue that reasoning strategies (also called diagnostic processes), such as generating hypotheses, play an important role in diagnosing in a conscious way. Third, cognitive theories suppose that diagnosing is heavily influenced by biases and the interplay between different cognitive systems (Elstein and Schwartz, 2002). A popular example of cognitive theories are medical dual-process theories of diagnosing (Croskerry, 2009; Eva, 2004; Evans, 2008). These theories assume that a separate fast, unconscious system and slow, conscious cognitive system are both involved in diagnosing. However, the three described theoretical perspectives are no longer considered mutually exclusive. Despite different operationalizations, most researchers agree that clinical reasoning includes aspects of knowledge and diagnostic processes to some extent (Eva, 2004).

Assessing clinical reasoning with virtual patients

Virtual patients can be defined as digital simulations of important clinical situations such as the medical interview (Cook et al., 2010). Virtual patients can vary with regard to many characteristics, such as the underlying model, type of user input, and degree of authenticity (Huwendiek et al., 2009). Nevertheless, it can be said that virtual patients provide some kind of interactivity and contain audiovisual materials. Moreover, virtual patients are conducted in a highly standardized way and can offer detailed log data about participants' diagnostic processes. Perhaps for these reasons, virtual patients have become increasingly popular tools for formative and summative assessment in medical education in recent decades (Ryall et al., 2016) and are even used in national licensing examinations in medicine (Boulet and Durning, 2019). The aforementioned features and their widespread use highlight that virtual patients could be particularly suitable for investigating the relationships among the key components of clinical reasoning.

Focused and comprehensive outcome measures used in virtual patients

In the past, most virtual patient assessments used the focused outcome measure of diagnostic accuracy (Daniel et al., 2019), which can be defined as the correctness of the final diagnosis. Diagnostic accuracy has the advantage of being relatively easy to measure electronically (e.g., via single-choice questions regarding the participant's diagnosis) and can be scored relatively objectively. However, practitioners and researchers have repeatedly argued that virtual patients should capture diagnostic success more comprehensively (Daniel et al., 2019; Elder, 2018; Round et al., 2009). Comprehensive outcome measures for virtual patients can include but are not limited to additional diagnostic tests, treatment decisions, prognosis, and justifications for all of these aspects (Daniel et al., 2019). Incorporating aspects like these into virtual patient assessments could help to diminish overtreatment and undertreatment of patients (Mamede and Schmidt, 2014) and gain more detailed insights into students' specific errors in diagnosing. Moreover, the extent of the association between focused and comprehensive measures of diagnosing remains an open question.

Operationalization of the key components of clinical reasoning

Our study operationalizes the key components of clinical reasoning based on a framework by Heitzmann et al. (2019) and related literature (Förtsch et al., 2018; Stark et al., 2011). In terms of the three aforementioned perspectives on clinical reasoning, this framework provides a problem-solving theory that also incorporates knowledge-related aspects. This perspective and the framework's focus on technology-enhanced learning make it suitable investigating virtual patients. In this framework, knowledge is assessed as *professional knowledge,* consisting of conceptual and strategic knowledge (Heitzmann et al., 2019). *Conceptual knowledge* is knowledge about facts and constructs, termed "knowing what", whereas *strategic knowledge* refers to knowledge about possible paths and heuristics in diagnosing, termed "knowing how" (Förtsch et al., 2018; Stark et al., 2011). The diagnostic process is operationalized in our study via diagnostic activities. *Diagnostic activities* are "components of professional problem-solving" (Heitzmann et al., 2019, p. 3) and knowledge-generating practices that are learned through training.. They can occur in varying quantity, quality, and sequence – but it is mainly their quality that is assumed to be linked with diagnostic success (Heitzmann et al., 2019). These characteristics differentiate diagnostic activities from other diagnostic processes. For example, diagnostic activities can be distinguished from the notion of diagnostic steps – in which an ideal diagnostic sequence is determined based on experts' think-aloud protocols (Kassirer, 2010). The diagnostic activities of *hypothesis generation*, *evidence generation*, and *evidence evaluation* (Heitzmann et al., 2019) were selected because theoretical accounts and empirical studies indicate that they are related to diagnostic success in the context of medical history-taking (Fink et al., 2022; Ramsey et al., 1998; Roter and Hall, 1987). Please see Table 1 for definitions of these terms. Diagnostic success is measured in our study with a focused *diagnostic accuracy* score and a *comprehensive diagnostic score*. *Diagnostic accuracy* can be defined as the level of agreement between the student's diagnosis and the correct expert solution (Heitzmann et al., 2019). The *comprehensive diagnostic score* includes the diagnostic accuracy, treatment selected, diagnostic measures taken for clarification, and suspected findings in a physical examination. The latter operationalization aims to provide a more extensive measure of clinical reasoning, as called for in the literature (Daniel et al., 2019; Elder, 2018; Round et al., 2009).

Table 1 Definitions of the three diagnostic activities measured in this study

Term	**Definition**
Hypothesis generation	Hypothesis generation refers to creating a case diagnosis based on initial but sometimes unspecific key information about the patient (e.g., formulating a hypothesis upon initially encountering a case).
Evidence generation	Evidence generation refers to gathering and creating additional information for the diagnosis (e.g., asking questions in a medical interview).
Evidence evaluation	Evidence evaluation refers to interpreting the meaning and reliability of pieces of acquired information (e.g., analyzing information such as a lab test result).

See Heitzmann et al. (2019) for a comprehensive overview of all diagnostic activities.

The relationships among the key components of clinical reasoning

As previously mentioned, clinical reasoning is an extensive construct that can be viewed from various theoretical perspectives (Young et al., 2018). This study explores the relationships between key components of clinical reasoning as operationalized based on the framework by Heitzmann et al. (2019).

The relationship between prior professional knowledge and diagnostic success

Stark et al. (2011) investigated the associations of different knowledge components with diagnostic success in two experiments. Conceptual knowledge, strategic knowledge, and performance on text-based problem-solving tasks, focusing on diagnostic accuracy, were measured. In both experiments, medical students were the participants, and diagnostic success in the problem-solving tasks was positively correlated with conceptual and strategic prior knowledge. Adding to these results, a study by Schmidmaier et al. (2013) with medical students as participants examined associations between prior knowledge and performance in a text-based problem-solving task that required clinical decision-making. More specifically, students had to describe underlying pathophysiological processes and provide explanations for their decisions. The study found a high correlation with strategic knowledge and a medium correlation with conceptual knowledge for the aforementioned problem-solving task. Recently, associations between knowledge and diagnostic success have also been found in the context of virtual patients. In a study by Kiesewetter et al. (2020), medical students completed knowledge tests and virtual patient assessments. Participants with a high combined score for conceptual and strategic knowledge performed better in diagnostic accuracy than participants with low scores in the knowledge test. In sum, these studies indicate that conceptual and strategic knowledge are linked to both focused and comprehensive diagnostic success measures.

The relationship between diagnostic activities and diagnostic success

Associations between the quality of hypothesis generation and diagnostic success have been found in studies involving self-generated and externally suggested hypotheses in solving text-based cases (Coderre et al., 2010; Leblanc et al., 2001; Leblanc et al., 2002). Moreover, correlations between hypothesis generation and diagnostic success measures have been discovered with standardized patients (Barrows et al., 1982; Neufeld et al., 1981). Taken together, these studies suggest that the quality of hypothesis generation is positively associated with diagnostic success in other contexts as well, such as with virtual patients.

Correlations between the quality of evidence generation and diagnostic success have also been reported. Woolliscroft et al. (1989) investigated physicians' history-taking with standardized patients and found an association between specific questions asked and the percentage of critical features obtained. In a study by Stillman et al. (1991), physicians took part in standardized patient evaluations. Performance on a history-taking checklist filled out by the standardized patients had a small but significant positive correlation with achieved diagnostic accuracy. Moreover, Fink et al. (2021b) discovered a medium positive association between the quality of evidence generation and diagnostic accuracy in virtual patients.

A relationship between the quality of evidence evaluation and diagnostic success can also be presumed. The data interpretation process that takes place within the script concordance test (Charlin et al., 2000), a valid and reliable test of clinical reasoning, shares similarities with the definition of evidence evaluation by Heitzmann et al. (2019) that is applied in this study (see Table 1). Investigating such a data interpretation process in virtual patients rather than the text-based cases included in the script concordance test seems particularly promising.

Up to now, the contribution of diagnostic activities to diagnostic success has not been sufficiently researched by studies investigating multiple predictors together -- with one notable exception. Groves et al. (2003) examined failures in three diagnostic processes when working on text-based cases in medicine, two of which were similar to the diagnostic activities of hypothesis generation and evidence evaluation. The study found that failures in these diagnostic processes predicted lack of diagnostic success (Groves et al., 2003).

Nevertheless, despite repeatedly finding a relationship between diagnostic activities and diagnostic success, the reported studies did not systematically control for prior knowledge. Thus, we cannot rule out that the observed diagnostic activities are only epiphenomena, fully determined by prior knowledge.

Are diagnostic activities an embodiment of knowledge?

As previously mentioned, an analysis of whether diagnostic activities make a unique contribution to explaining diagnostic success over and above knowledge seems warranted. Norman et al. (2005) theorized that the diagnostic process depends strongly on available knowledge and is a strategy "to access and apply different kinds of specific knowledge" (p. 424). Moreover, this research question is particularly important for students with partial knowledge. For this group of students, deep and rich knowledge networks like illness scripts, which would enable them to make diagnoses via a quick and automatic pattern recognition process, are frequently not yet available (Schmidt and Rikers, 2007). Also, students with partial knowledge may fail to activate available knowledge networks sufficiently when diagnostic problems possess a challenging level of difficulty. Nevertheless, this group of students knows how to perform certain diagnostic activities, such as generating evidence (e.g., through asking questions in a medical interview), that have been taught in medical school, which could help them activate their knowledge and make a correct diagnosis. Against this background, it is an open question whether diagnostic activities mainly represent an embodiment of knowledge or are an important component of clinical reasoning, building on but not entirely determined by accessible knowledge, particularly among students with partial knowledge.

Research question and hypotheses

This study investigates to what extent diagnostic activities and prior professional knowledge uniquely explain variance in diagnostic success. This research question is examined for two indicators of diagnostic success: a comprehensive diagnostic score and diagnostic accuracy. Concerning comprehensive diagnostic score, we hypothesize that three diagnostic activities (H1.1), namely hypothesis generation, evidence generation, and evidence evaluation, as well as prior professional knowledge (H1.2), consisting of conceptual and strategic knowledge, both explain variance. Moreover, we assume that the diagnostic activities increase the amount of explained variance over and above prior professional knowledge (H1.3). For diagnostic accuracy, we propose the same hypotheses as for comprehensive diagnostic score (H2.1 - H2.3).

METHOD

Procedure and participants

The participants began the experiment by completing a conceptual and a strategic knowledge test. Then, the participants underwent a familiarization procedure explaining how to work with the virtual patients. Afterward, the participants diagnosed multiple virtual patients on the topic of history-taking for dyspnea.

Altogether, *N* = 121 medical students took part in the study. These participants were studying medicine in years three to five of a six-year program. Due to using hierarchical regression analyses and for consistency reasons, all participants with missing values were dropped, resulting in a final sample of *N* = 106 participants, with a mean age of *M* = 24.76 years, *SD* = 3.83. This final sample included *n* = 70 females (66.0%), *n* = 9 males (8.5%) and *n* = 27 (25.5%) participants without gender information. This high percentage of participants without gender information was probably primarily caused by an electronic form that allowed participants to skip this question.

Knowledge tests

Conceptual knowledge test

The conceptual knowledge test contained 20 items relevant to dyspnea and history-taking. This test consisted of previously validated exam questions and thus used two popular question formats: single-choice questions and multiple-response questions. In single-choice questions, 1.0 points were allocated for each correct answer. In multiple-response questions, points were awarded as follows: 1.0 points were given for an entirely correct answer pattern, and 0.50 points were allocated if more than 50% of the participant's answers were correct (Bauer et al., 2011). To build a scale, the number of points achieved was divided by the number of questions posed. This scale ranged from 0 (*low knowledge*) to 1 (*high knowledge*). The test reached acceptable reliability of $\alpha = .66$.

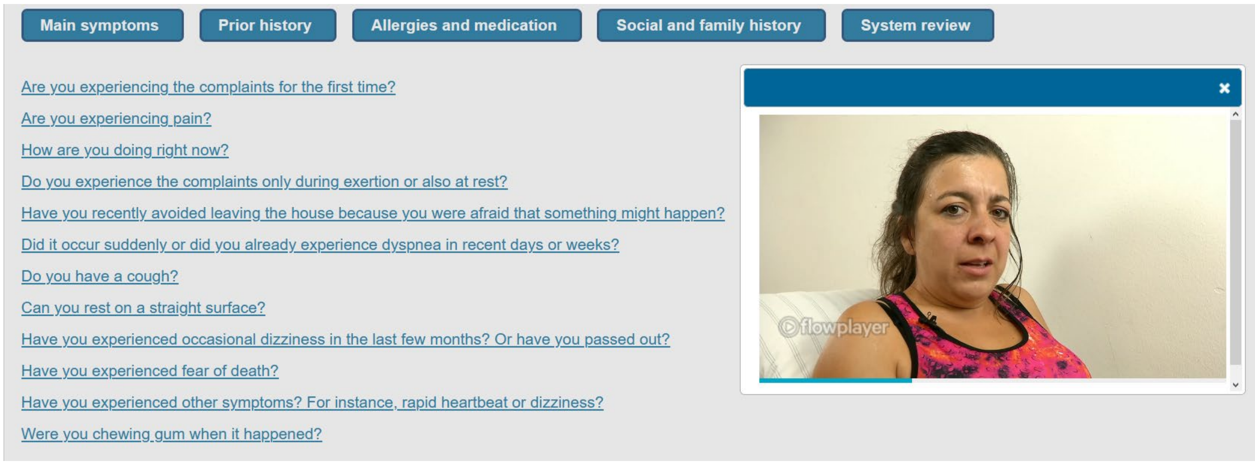
Strategic knowledge test

Strategic knowledge of dyspnea and history-taking was measured with four key feature cases (Hrynchak et al., 2014) validated in a prior study (Fink et al., 2021b). Each key feature case contained four single-choice questions. These four single-choice questions focused on the diagnosis, treatment, symptoms, and further diagnostic measures. 1.0 points were allocated for each correct answer. The scale for the strategic knowledge test was built by dividing the number of points achieved by the number of questions posed. This scale ranged from 0 (*low knowledge*) to 1 (*high knowledge*). The scale's reliability was acceptable, with $\alpha = .65$.

Virtual patients

Topic and simulation scenario

The participants encountered multiple virtual patients representing different causes of dyspnea and engaged in history-taking for diagnosing. The simulation scenario for the virtual patients was as follows. The simulation began with the presentation of prior information (e.g., lab results) and the patient's chief complaint. Next, participants selected questions to ask the virtual patient from a menu of history-taking questions. This menu included up to 69 standardized questions for each case and was subdivided into the categories *main symptoms*, *prior history*, *allergies and medication*, *social and family history*, and *system review*. The history-taking questions and menu had been validated in previous studies (Fink et al., 2021b; Fink et al., 2021a), and examples of the history-taking questions are listed in Appendix S1. After the participant selected a question from the menu, the corresponding answer was streamed as a video. Each virtual patient encounter lasted between a minimum of five minutes and a maximum of ten minutes. Before each virtual patient, participants were instructed to spend at least the minimum amount of time working with the simulation. They were then notified by prompts when the minimum and maximum time had been reached. A screenshot of a virtual patient at the point of selecting questions from the menu is provided in Fig. 1.



Are you experiencing the complaints for the first time?

Are you experiencing pain?

How are you doing right now?

Do you experience the complaints only during exertion or also at rest?

Have you recently avoided leaving the house because you were afraid that something might happen?

Did it occur suddenly or did you already experience dyspnea in recent days or weeks?

Do you have a cough?

Can you rest on a straight surface?

Have you experienced occasional dizziness in the last few months? Or have you passed out?

Have you experienced fear of death?

Have you experienced other symptoms? For instance, rapid heartbeat or dizziness?

Were you chewing gum when it happened?

Fig. 1 Screenshot of a virtual patient

Creation of the virtual patients and electronic assessment environment

As a first step to creating the virtual patients, professional actors were hired and then trained for their role by a physician and an acting coach. When filming the videos, the professional actors exhibited the patients' symptoms according to their script. After editing, the videos were integrated with additional case information to create the virtual patients in the electronic assessment environment CASUS (Instruct, 2021).

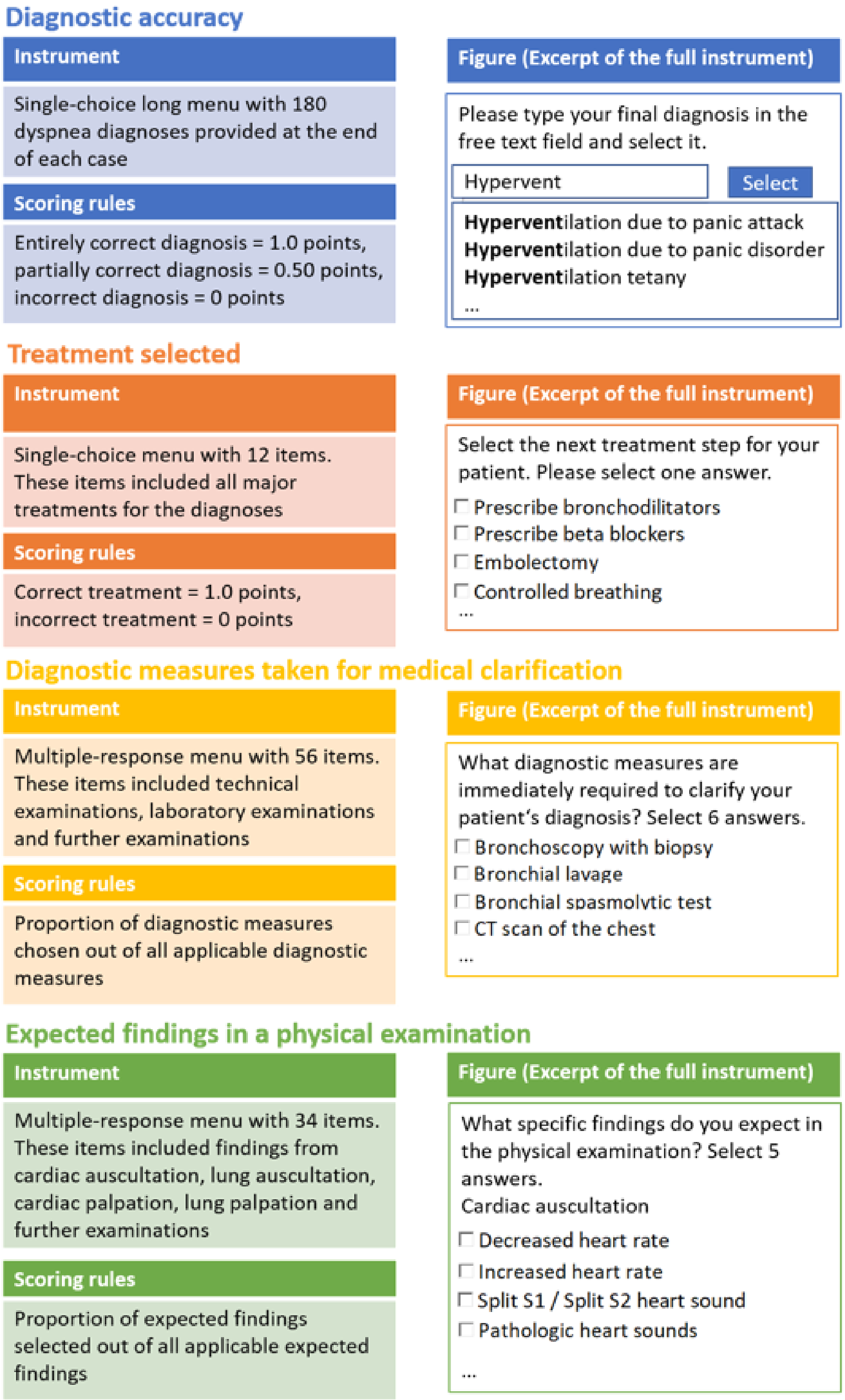
Diagnostic success measures

Diagnostic success was assessed with *diagnostic accuracy* and a *comprehensive diagnostic score*.

Diagnostic accuracy was operationalized as the correspondence between the participant's diagnosis and the case's solution. This variable was assessed with a long menu

– a free text field with a concealed list of answers and an autocomplete feature. The solutions used to score the answers were determined by a licensed physician and a specialist in general medicine and previously used in a study by Fink et al. (2021b). More information on the instrument is available in Fig. 2.

The comprehensive diagnostic score encompassed four equally-weighted variables: 1) *diagnostic accuracy*, 2) *treatment selected*, 3) *diagnostic measures taken for medical clarification,* and 4) *expected findings in a physical examination*. Diagnostic accuracy was operationalized and measured as previously described. Treatment selected was defined as the most important, next treatment for the patient. Diagnostic measures taken for clarification refer to all technical/diagnostic measures immediately necessary to investigate the diagnosis further. Expected findings in the physical examination denote the specific signs and symptoms expected to be observed in a physical exam following history-taking. More details on the instruments and the scoring are provided in Fig. 2. Participants' responses to the described variables were compared to a sample solution jointly developed by a licensed physician and a specialist in general medicine using *R* scripts. A principal component analysis with varimax rotation as well as corresponding Eigenvalue and scree plot analyses indicated that all four variables belong to one comprehensive diagnostic score factor and explained 61.1% of the variance in comprehensive diagnostic score (see Appendix S2). Due to the different answer formats and points allocated, scores on the four variables were standardized before calculating the average comprehensive diagnostic score.



Diagnostic accuracy

Instrument

Single-choice long menu with 180 dyspnea diagnoses provided at the end of each case

Scoring rules

Entirely correct diagnosis = 1.0 points, partially correct diagnosis = 0.50 points, incorrect diagnosis = 0 points

Figure (Excerpt of the full instrument)

Please type your final diagnosis in the free text field and select it.

Hypervent Select

Hyperventilation due to panic attack
Hyperventilation due to panic disorder
Hyperventilation tetany
…

Treatment selected

Instrument

Single-choice menu with 12 items. These items included all major treatments for the diagnoses

Scoring rules

Correct treatment = 1.0 points, incorrect treatment = 0 points

Figure (Excerpt of the full instrument)

Select the next treatment step for your patient. Please select one answer.

□ Prescribe bronchodilitators
□ Prescribe beta blockers
□ Embolectomy
□ Controlled breathing
…

Diagnostic measures taken for medical clarification

Instrument

Multiple-response menu with 56 items. These items included technical examinations, laboratory examinations and further examinations

Scoring rules

Proportion of diagnostic measures chosen out of all applicable diagnostic measures

Figure (Excerpt of the full instrument)

What diagnostic measures are immediately required to clarify your patient's diagnosis? Select 6 answers.

□ Bronchoscopy with biopsy
□ Bronchial lavage
□ Bronchial spasmolytic test
□ CT scan of the chest
…

Expected findings in a physical examination

Instrument

Multiple-response menu with 34 items. These items included findings from cardiac auscultation, lung auscultation, cardiac palpation, lung palpation and further examinations

Scoring rules

Proportion of expected findings selected out of all applicable expected findings

Figure (Excerpt of the full instrument)

What specific findings do you expect in the physical examination? Select 5 answers.

Cardiac auscultation

□ Decreased heart rate
□ Increased heart rate
□ Split S1 / Split S2 heart sound
□ Pathologic heart sounds
…

Fig. 2 Diagnostic success measures

Case selection and preliminary analyses

The diagnoses for the four virtual patient cases included in our study are reported in Table 2. Moreover, the respective descriptive statistics for these cases are reported in Table 3. However, two other cases had to be excluded from our study due to floor effects on diagnostic success measures. Please see Appendix S3 for the diagnoses and descriptive statistics for these excluded cases.

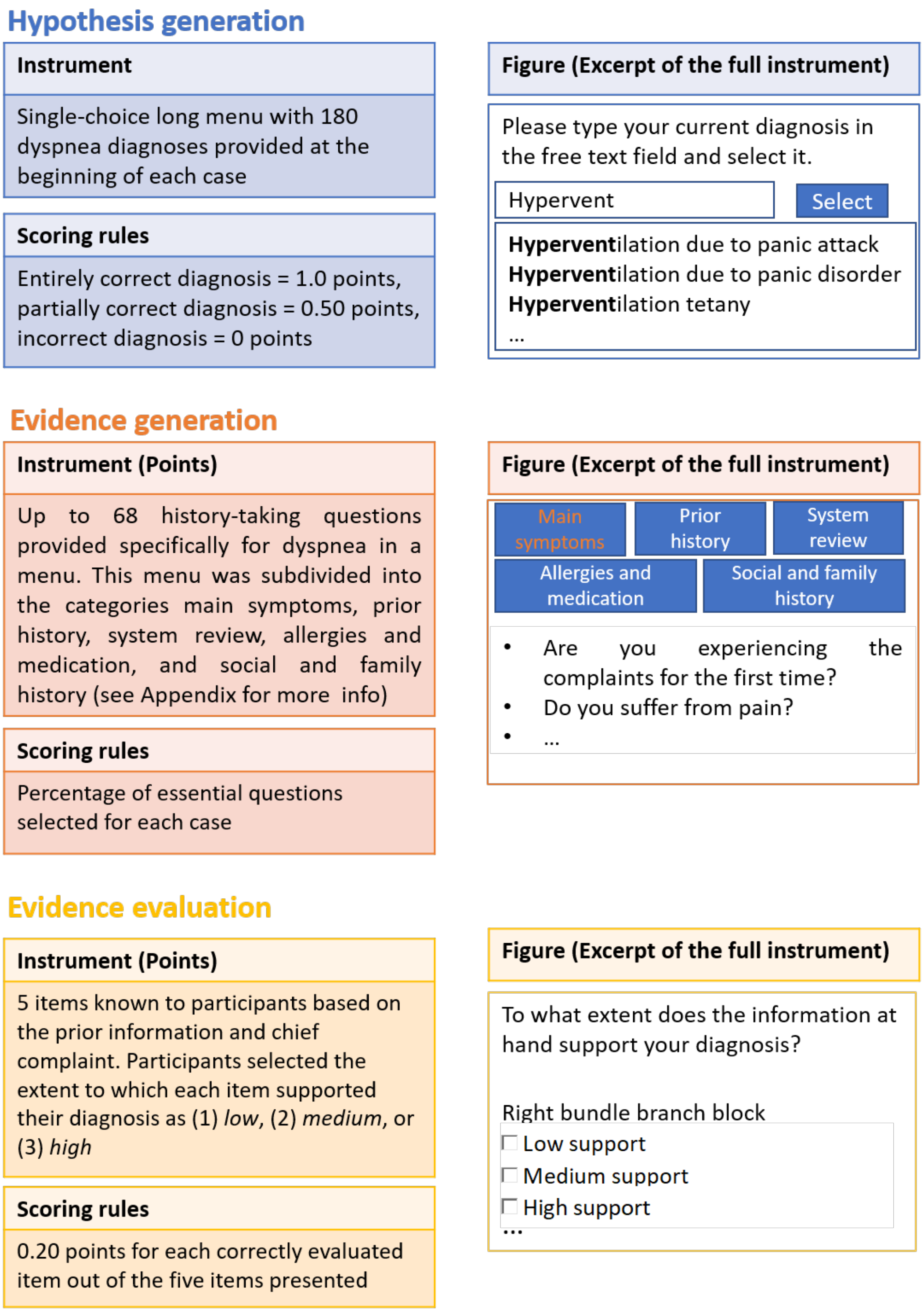
Table 2 Diagnoses of the virtual patient cases included in the study

Case number	Diagnosis	Patient characteristics	Patient name
1	Hypertrophic cardiomyopathy	25 years, male	Mr. Albrecht
2	Pneumonia	55 years, female	Ms. Klein
3	Pulmonary embolism with a coagulation disorder	35 years, female	Ms. Aimüller
4	Panic attack	45 years, male	Mr. Lehner

Diagnostic activities

Based on Heitzmann et al. (2019), we also assessed three diagnostic activities. We measured the quality of *hypothesis generation* using the same long menu previously described as an instrument for measuring diagnostic accuracy. In contrast to diagnostic accuracy, the measure of hypothesis generation occurred at the beginning of each virtual patient encounter.

The quality of *evidence generation* was assessed based on the questions selected during history-taking. Participants selected these questions from the menu described in Appendix S1, and all questions were specific to dyspnea and standardized across the virtual patients. To score this variable, we used a coding scheme previously utilized for the same history-taking questions on the same virtual patients by Fink et al. (2021b). This coding scheme was a joint, common solution developed by one licensed physician and one specialist in general medicine that specified the essential questions for each case. The quality of *evidence evaluation* was measured retrospectively after the participant diagnosed each virtual patient. In completing this instrument, participants judged to what extent aspects known from the prior information and chief complaint supported their final diagnosis. This instrument and the corresponding sample solutions were developed by a licensed physician and reviewed by a specialist in general medicine. Additional information on all three diagnostic activities is provided in Fig. 3. It should be added that the participants' diagnostic activities were automatically compared to the sample solutions via *R* scripts.



Hypothesis generation

Instrument

Single-choice long menu with 180 dyspnea diagnoses provided at the beginning of each case

Scoring rules

Entirely correct diagnosis = 1.0 points, partially correct diagnosis = 0.50 points, incorrect diagnosis = 0 points

Figure (Excerpt of the full instrument)

Please type your current diagnosis in the free text field and select it.

Hypervent

Hyperventilation due to panic attack
Hyperventilation due to panic disorder
Hyperventilation tetany
…

Evidence generation

Instrument (Points)

Up to 68 history-taking questions provided specifically for dyspnea in a menu. This menu was subdivided into the categories main symptoms, prior history, system review, allergies and medication, and social and family history (see Appendix for more info)

Scoring rules

Percentage of essential questions selected for each case

Figure (Excerpt of the full instrument)

- Are you experiencing the complaints for the first time?
- Do you suffer from pain?
- …

Evidence evaluation

Instrument (Points)

5 items known to participants based on the prior information and chief complaint. Participants selected the extent to which each item supported their diagnosis as (1) *low*, (2) *medium*, or (3) *high*

Scoring rules

0.20 points for each correctly evaluated item out of the five items presented

Figure (Excerpt of the full instrument)

To what extent does the information at hand support your diagnosis?

Right bundle branch block
□ Low support
□ Medium support
□ High support
…

Fig. 3 Diagnostic activities measures

Data collection method and statistical and power analyses

The study's data was gathered from October 2019 until February 2021 at the University Hospital, LMU Munich, in Germany. Due to the COVID-19 pandemic, the data collection method had to be changed while the study was running. Until March 2020, data from $n = 30$ participants included in the final sample was gathered on-site in a computer lab. After March 2020, data from $n = 76$ participants in the final sample was collected web-based. A control analysis reported in Appendix S4 showed that the lab-based and web-based participants differed in terms of knowledge, diagnostic activities, and diagnostic success variables. Therefore, we ran statistical tests for effects of the data collection method in Appendix S5 by repeating the regression analyses reported in the results section while including the data collection method as a factor. In these analyses, we modeled interaction effects between the data collection method and all relevant predictors and found that the effect of the predictors did not depend on the data collection method.

We used R version 4.0.2 (R Core Team, 2020) for our statistical analyses. Multiple regression and hierarchical regression analyses were conducted to investigate our research questions. Frequently used assumptions checks for regression models, including residuals vs. fitted values plots, Q-Q plots, and scale-location plots, confirmed that these regression models were a good fit. In all statistical analyses, the significance level was set to $\alpha = .05$.

Post hoc power analyses were conducted with G*Power version 3.1 (Faul et al., 2009). For the power analyses, we set the error probability to $\alpha = .05$ and the sample size to N = 106. Our analyses were based on a medium effect of Cohen's $f^2 = 0.15$ and revealed power of at least $\beta = 0.87$ for each analysis.

RESULTS

Descriptive statistics and intercorrelations

Participants reached a medium score on the conceptual knowledge (*M* = 0.54, *SD* = 0.14) and strategic knowledge (*M* = 0.50, *SD* = 0.14) tests preceding the virtual patient cases. As reported in Table 3, performance in case 1 to case 4 on the diagnostic activities and diagnostic success measures was medium and can be considered suitable.

Intercorrelations for professional knowledge, the three diagnostic activities, and diagnostic success measures are reported in Table 4. The relationships between these variables are examined in more detail in the following regression models. It should be added that we found a medium correlation between conceptual and strategic knowledge (*r* = .55). This correlation was examined more closely for multicollinearity issues using the variance inflation index. As collinearity between the two knowledge types was slight to moderate (VIF = 1.44), both variables were included together in the regression models.

Table 3 Descriptive statistics for diagnostic activities and diagnostic success measures

	Case 1 (Albrecht)	**Case 2 (Klein)**	**Case 3 (Aimüller)**	**Case 4 (Lehner)**	**Total**
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Diagnostic activities					
Hypothesis generation	0.35 (0.26)	0.73 (0.41)	0.08 (0.20)	0.23 (0.39)	0.35 (0.19)
Evidence generation	0.40 (0.16)	0.37 (0.20)	0.67 (0.24)	0.32 (0.20)	0.44 (0.13)
Evidence evaluation	0.59 (0.21)	0.48 (0.19)	0.37 (0.25)	0.50 (0.23)	0.49 (0.11)
Diagnostic success measures					
Diagnostic accuracy	0.31 (0.33)	0.64 (0.46)	0.49 (0.47)	0.32 (0.41)	0.43 (0.23)
Treatment selected	0.60 (0.49)	0.69 (0.47)	0.50 (0.50)	0.44 (0.50)	0.56 (0.25)
DM	0.56 (0.17)	0.59 (0.37)	0.42 (0.22)	0.36 (0.30)	0.48 (0.17)
EF	0.52 (0.16)	0.56 (0.21)	0.47 (0.15)	0.78 (0.35)	0.58 (0.14)
Comprehensive diagnostic score	—	—	—	—	.04 (1.00)

The comprehensive diagnostic score was normalized with z-scores ranging from -3 to +3 and only calculated for the total score. Range of all other variables: (0) *low* to (1) *high*. Abbreviations: DM = Diagnostic measures taken for medical clarification, EF = Expected findings in a physical examination

Table 4 Intercorrelations of knowledge, diagnostic activities, and diagnostic success measures

	1.	2.	3.	4.	5.	6.	7.
1. Conceptual knowledge	—						
2. Strategic knowledge	.55***	—					
3. Hypothesis generation	.00	-.07	—				
4. Evidence generation	.31**	.47***	-.02	—			
5. Evidence evaluation	.01	.17	.18	.11	—		
6. Comprehensive diagnostic score	.36***	.41***	.30**	.42***	.35***	—	
7. Diagnostic accuracy	.23*	.21*	.41***	.22*	.18	.76***	—

Two-tailed Pearson correlations. Note that the scores for hypothesis generation, evidence generation, evidence evaluation, the comprehensive diagnostic score, and diagnostic accuracy were aggregated over four virtual patients.
* p < .05, ** p < .01, *** p < .001

The contribution of diagnostic activities and professional knowledge to the comprehensive diagnostic score

Regression analyses for the comprehensive diagnostic score as criterion were conducted. Model 1, containing diagnostic activities as predictors, was significant. As expected in H1.1, the three diagnostic activities together explained a substantial amount of variance in the comprehensive diagnostic score. Model 2a, encompassing the two aspects of professional knowledge as predictors, was also significant. In line with H1.2, professional knowledge accounted for substantial amounts of variance in the comprehensive diagnostic score. Model 2b consisted of the predictors in Model 2a plus the three diagnostic activities added in a second step; this model was also significant. A comparison of the two models indicated that Model 2b explained substantially more variance than Model 2a ($F(3, 100) = 12.30$, $p < .001$, $\Delta \mathrm{R}^2 = .21$, $\Delta \mathrm{Adj.\,R}^2 = .20$). This finding supports H1.3, that the diagnostic activities increase the amount of explained variance in the comprehensive diagnostic score over and above professional knowledge. Table 5 contains the details of the multiple and hierarchical regression models used.

Table 5 Regression analyses for comprehensive diagnostic score as outcome

Predictor	***b***	***ß***	***ß* 95% CI**	***p***	***Model test and fit***
Model 1					$F(3, 102) = 17.21, p < .001$
Intercept	-2.30***			< .001	$R^2 = .34$
Hypothesis generation	1.09**	0.26	[0.10, 0.42]	.002	Adj. $R^2 = .32$
Evidence generation	2.40***	0.40	[0.24, 0.56]	< .001	
Evidence evaluation	1.76**	0.26	[0.10, 0.42]	.002	

Model 2a					$F(2, 103) = 12.55, p < .001$
Intercept	-1.42***			<.001	$R^2 = .20$
Conceptual knowledge	1.07	0.19	[-0.02, 0.40]	.074	Adj. $R^2 = .18$
Strategic knowledge	1.66**	0.31	[0.10, 0.52]	.005	
Model 2b					$F(5, 100) = 14.05, p < . 001$
Intercept	-2.92***			<.001	$R^2 =.41$
Conceptual knowledge	1.07*	0.19	[0.01, 0.37]	.043	Adj. $R^2 = .38$
Strategic knowledge	0.86	0.16	[-0.04, 0.36]	.121	
Hypothesis generation	1.14***	0.27	[0.12, 0.43]	<.001	
Evidence generation	1.60**	0.27	[0.09, 0.44]	.003	
Evidence evaluation	1.65**	0.24	[0.08, 0.40]	.003	

Model 1 is a multiple regression containing diagnostic activities variables. Model 2 is a hierarchical regression consisting of knowledge variables in Model 2a and knowledge and diagnostic activities in Model 2b. *b* represents unstandardized regression weights. *ß* represents standardized regression weights. CI = confidence interval. * $p < .05$, ** $p < .01$, *** p < .001

The contribution of diagnostic activities and professional knowledge to diagnostic accuracy

Regression analyses for diagnostic accuracy as the criterion were also conducted. Model 3, containing diagnostic activities as predictors, was significant. As expected in H2.1, the three diagnostic activities together explained a substantial amount of variance in diagnostic accuracy. Model 4a, encompassing conceptual and strategic knowledge as predictors, was also significant, but this was only due to a significant intercept term. However, the bivariate relations between conceptual and strategic knowledge and diagnostic accuracy were significant (see Table 2). These findings can be seen as mixed evidence for H2.2 that professional knowledge is associated with diagnostic accuracy. Model 4b consisted of the predictors in Model 4a plus the three diagnostic activities added in a second step; this model was also significant. A comparison of the two models indicated that Model 4b explained substantially more variance than Model 4a ($F(3, 100) = 8.85$, $p < .001$, $\Delta \mathrm{R}^2 = .20$, $\Delta \mathrm{Adj.\,R}^2 = .18$). This finding supports H2.3, that the diagnostic activities increase the amount of explained variance in diagnostic accuracy over and above professional knowledge. Table 6 contains the details of the multiple and hierarchical regression models used.

Table 6 Regression analyses for diagnostic accuracy as outcome

Predictor	***b***	***ß***	***ß* 95% CI**	***p***	***Model test and fit***
Model 3					$F(3, 102) = 10.00, p < .001$
Intercept	0.01			.981	$R^2 = .23$
Hypothesis generation	0.48**	0.40	[0.22, 0.57]	<.001	Adj. $R^2 = .20$
Evidence generation	0.40*	0.22	[0.05, 0.40]	.012	
Evidence evaluation	0.17	0.09	[-0.09, 0.26]	.331	
Model 4a					$F(2, 103) = 3.42, p = .037$
Intercept	0.19*			.040	$R^2 = .06$
Conceptual knowledge	0.27	0.16	[-0.06, 0.39]	.157	Adj. $R^2 = .04$
Strategic knowledge	0.19	0.12	[-0.11, 0.35]	.302	
Model 4b					$F(5, 100) = 6.99, p < .001$
Intercept	-0.12			.335	$R^2 = .26$

Conceptual knowledge	0.24	0.14	[-.006, 0.35]	.172	Adj. R^2 = .22
Strategic knowledge	0.12	0.07	[-0.15, 0.30]	.516	
Hypothesis generation	0.49	0.40	[0.23, 0.58]	< .001	
Evidence generation	0.26	0.15	[-0.05, 0.34]	.142	
Evidence evaluation	0.16	0.08	[-0.10, 0.26]	.368	

Model 3 is a multiple regression containing diagnostic activities variables. Model 4 is a hierarchical regression, consisting of knowledge variables in Model 4a and knowledge and diagnostic activities in Model 4b. *b* represents unstandardized regression weights. *ß* represents standardized regression weights. CI = confidence interval. * *p* < .05, ** *p* < .01, *** p < .001

DISCUSSION

Principal findings

We investigated to what extent diagnostic activities and professional knowledge contribute to making a successful diagnosis in the context of virtual patients presenting with dyspnea. Our analyses were conducted with respect to both a more extensive comprehensive diagnostic score and a more focused diagnostic accuracy score.

The contribution of the diagnostic activities to diagnostic success

The diagnostic activities (Heitzmann et al., 2019) of hypothesis generation, evidence generation, and evidence evaluation were used to operationalize the diagnostic process. These three diagnostic activities together accounted for a substantial amount of the variance in the comprehensive diagnostic score and the focused diagnostic accuracy score (Model 1 R^2 = .34 resp. Model 3 R^2 = .23).

Next, we will discuss the contribution of the individual diagnostic activities. Hypothesis generation was a strong predictor of the comprehensive diagnostic score and diagnostic accuracy in both regression models (Model 1 and Model 3). This finding concurs with research highlighting the strong associations between hypotheses and diagnostic success in solving text-based cases (Coderre et al., 2010; Leblanc et al., 2001; Leblanc et al., 2002), as well as in standardized patients (Barrows et al., 1982; Neufeld et al., 1981). Likewise, evidence generation predicted the comprehensive diagnostic score and focused diagnostic accuracy score. This result is in line with correlational results gathered in virtual patients, standardized patients, and real-life professional contexts (Fink et al., 2021b; Stillman et al., 1991; Woolliscroft et al., 1989). Evidence evaluation, however, was only a significant predictor of the comprehensive diagnostic score, not of the diagnostic accuracy score. This unexpected result may be explained by looking at the information upon which the evidence evaluation instrument was based. In our evidence evaluation instrument, participants retrospectively assessed the extent to which five key pieces of information supported their final hypothesis. Competence in interpreting the meaning of key pieces of information and the information itself may have helped participants request the treatments and diagnostic measures included in the comprehensive diagnostic score. However, competence in interpreting the meaning of key information and the information itself may not have substantially assisted participants in selecting the correct final diagnosis.

Overall, our results demonstrate that diagnostic activities account for variance in diagnostic success measures. This result is consistent with theories that view clinical reasoning as a problem-solving process (Elstein et al., 1978; Elstein et al., 1990), and adds to the study by Groves et al. (2003), which found that failures in diagnostic processes relatively similar to diagnostic activities predicted lack of diagnostic success. Moreover, our results suggest that diagnostic activities could serve as a fruitful starting point for providing instructional support. Instructional support in the form of prompts and other cognitively-stimulating interventions (Chernikova et al., 2019; Chernikova et al., 2020) that target diagnostic activities could potentially be effective due to the observed association between diagnostic activities and diagnostic success. Once again, it must be emphasized that the investigated diagnostic activities differ from other conceptualizations of the diagnostic process. Diagnostic activities are knowledge-generating practices that can occur in varying sequences and quality and are taught and applied in medical education and medical practice. Other notions of diagnostic processes, such as diagnostic steps, regard diagnosing as a relatively linear process with an ideal sequence (Kassirer, 2010). Generalizations of our findings on diagnostic activities to other diagnostic processes should therefore be made only if the conceptualizations are sufficiently similar.

The contribution of professional knowledge to diagnostic success

In this study, we assessed professional knowledge (Heitzmann et al., 2019), consisting of conceptual and strategic knowledge. Professional knowledge explained a substantial amount of variance in the comprehensive diagnostic score and little variance in the diagnostic accuracy score (Model 2a R^2 = .20 resp. Model 4a R^2 = .06). Positive relationships with both diagnostic success measures were expected because Heitzmann et al.'s framework (2019) and other clinical reasoning theories, such as the illness script theory (Schmidt et al., 1990), assume that knowledge plays an important role in diagnosing.

The result that professional knowledge is predictive of comprehensive diagnostic score is in line with several empirical studies that found associations between knowledge and diagnosing in text-based problem-solving tasks and diagnosing virtual patients (Kiesewetter et al., 2020; Schmidmaier et al., 2013; Stark et al., 2011). To be more specific, we found in Model 2a that only strategic knowledge and not conceptual knowledge was a statistically significant predictor of the comprehensive diagnostic score. However, in bivariate correlation analyses, both types of knowledge displayed a medium correlation with the comprehensive diagnostic score and a medium correlation with each other. Thus, the non-significance of conceptual knowledge as a predictor might be due to its medium-level correlation with strategic knowledge (r = .55) and the shared variance of both variables. However, the amount of shared variance was completely acceptable, as highlighted by the reported variance inflation index.

Contrary to our expectations, there was mixed evidence for the relationship between professional knowledge and diagnostic accuracy. For one thing, there were significant bivariate correlations between conceptual and strategic knowledge and diagnostic accuracy (see Table 2). For another thing, both types of professional knowledge together did not predict the narrow diagnostic accuracy score in a regression and explained little variance (Model 4a). The non-significance of both knowledge types as predictors in the regression model could potentially be caused by their medium-level correlation. However, as previously mentioned, the shared variance between conceptual and strategic knowledge was fully acceptable, as highlighted by the reported variance inflation index. Therefore, it is more likely that both predictors did not become significant in the regression model because they were not associated strongly enough with the outcome. The small amount of explained variance discovered in the reported correlations and regressions for diagnostic accuracy can also be explained by looking at expertise development theory. According to the illness script theory (Schmidt et al., 1990), medical students mainly acquire extensive biomedical and clinical knowledge networks in the initial stages of expertise development (Boshuizen and Schmidt, 1992; Evans and Patel, 1989). Through repeated engagement with patients, medical students then acquire knowledge networks called illness scripts, which contain symptoms and underlying conditions (Schmidt and Rikers, 2007), in later stages of expertise development. The aforementioned knowledge networks are integrated and reorganized as expertise develops until they can be used efficiently in a largely automatic pattern recognition process of diagnosing (Charlin et al., 2007). Because the participants in our study were in their third to fifth year of medical school, it is reasonable to assume that they were still in or at the end of the initial stage of expertise development. As the participants also had little experience in treating patients, it is likely that they possessed only a few illness scripts, and the process of knowledge integration and reorganization was not yet advanced. This lacking integration and reorganization of knowledge could have impeded participants' application of their knowledge in diagnosing. The low observed associations between professional knowledge and diagnostic accuracy would be consistent with this explanation.

Are the diagnostic activities an embodiment of knowledge?

We also analyzed whether the diagnostic activities can be considered merely an embodiment of knowledge – or whether diagnostic activities can contribute to diagnostic success beyond prior knowledge. For the comprehensive diagnostic score and diagnostic accuracy score, hierarchical regressions demonstrated that the diagnostic activities added a significant amount of explained variance to that explained by participants' professional knowledge ($\Delta R^2 = .21$ resp. $\Delta R^2 = .20$). This result provides preliminary evidence that diagnostic activities make a unique contribution to diagnostic success and are thus more than merely an embodiment of knowledge. There are two major possible mechanisms explaining this finding. First, the quality with which the diagnostic activities (i.e., hypothesis generation, evidence generation, and evidence evaluation in this study) were performed may have increased the medical students' diagnostic success. Second, engagement in diagnostic activities with virtual patients may have helped the medical students access, activate or even generate relevant knowledge (i.e., learn) that they then implicitly applied in diagnosing. We would like to highlight that these findings were obtained for medical students with partial expertise who diagnosed relatively challenging cases, as shown by the reported diagnostic accuracy rates. The medical students in our study's sample probably did not yet possess deep and rich knowledge networks or could not apply these due to the challenging cases used. With increasing competence development, however, knowledge networks become more elaborated, and a quick and automatic pattern recognition process for diagnosing becomes crucial (Schmidt and Rikers, 2007). Therefore, we believe that the described findings mainly hold for medical students in earlier phases of competence development. In addition, the findings could potentially apply to advanced medical students and experts when they are presented with particularly difficult or novel cases.

Limitations

One limitation of the study is the scope and reliabilities of the used professional knowledge tests. Our study mainly focused on professional knowledge (Heitzmann et al., 2019) and measured its two main aspects of conceptual and strategic knowledge. However, the literature includes a multitude of other knowledge classifications which contain additional types of knowledge that could perhaps also be relevant for success in diagnosing (Förtsch et al., 2018). As mentioned earlier, the reliabilities of the two knowledge tests used were acceptable. It is a well-known fact that particularly low reliabilities place a constraint on associations between variables. Due to achieving only acceptable reliabilities, it is possible that the strength of associations between knowledge and the other variables reported in our study were weakened to some extent.

Another limitation of the study concerns the data collection method. The first part of the data was collected in the lab. Due to the COVID-19 pandemic, however, the second part of the data was gathered online. We acknowledge that web-based data collection does not allow for the same level of control and monitoring as lab-based data collection (Reips, 2000). As mentioned, a control analysis in Appendix S4 demonstrated that the lab-based and web-based participants differed with respect to conceptual knowledge, strategic knowledge, evidence generation, the comprehensive diagnostic score, and diagnostic accuracy. Thus, we screened for effects of the data collection method with another control analysis by recalculating the reported regressions with the data collection method as an additional factor (Appendix S5). We found that the predictors in our regression models did not depend on the data collection method. In addition to this statistical finding, the mentioned change in data collection method should not have affected this study's regression analyses because these focused on relating variables within and not across participants.

Conclusions

We conducted a study assessing medical students' clinical reasoning with virtual patients to examine to what extent knowledge and the diagnostic process, as operationalized by diagnostic activities, contribute to successful diagnosing. Our results provide support for clinical reasoning theories that conceptualize clinical reasoning as encompassing both process-related and knowledge-related aspects. Moreover, we found that the diagnostic activities learners engaged in made a unique contribution to diagnostic success, even when knowledge was considered. This result supports the view that the diagnostic process is – or can be – more than merely an embodiment of knowledge. There were two major possible mechanisms explaining this finding. First, the quality with which the diagnostic activities were performed may have increased the medical students' diagnostic success. Second, engaging in diagnostic activities like generating hypotheses and evidence may have helped the medical students access, activate or generate relevant knowledge. Also, the reported findings suggest that diagnostic activities could potentially serve as a starting point for providing effective instructional support with cognitively-stimulating interventions. Finally, we would like to highlight that our findings were obtained for third- to fifth-year medical students who probably possessed partial knowledge when facing cases with relatively challenging difficulty. Future research should extend our findings by measuring knowledge more extensively than in this study, recruiting participants in later stages of competence development with better integrated and organized knowledge, and tracking knowledge access, application and generation specifically during the diagnostic process.

Contributors: MCF contributed to the conceptualization and design of the study, gathered and analyzed the data, and created the first draft. NH contributed to the conceptualization and design of the study, advised on statistical analyses, and critically revised the article. VR contributed to the conceptualization and design of the study, advised on statistical analyses, and critically revised the article. MS conceptualized and designed the study, acquired funding, advised on statistical analyses, and critically revised the article. FF conceptualized and designed the study, acquired funding, advised on statistical analyses, and critically revised the article. MRF conceptualized and designed the study, acquired funding, advised on statistical analyses, and critically revised the article. All authors approved the final manuscript for publication and agreed to be accountable for all aspects of the work.

Acknowledgments: We thank Hannah Gerstenkorn for developing the materials and Keri Hartman for language editing. We would also like to thank Johannes Kissel, who helped to carry out the study. The first author thanks Larissa Kaltefleiter for her advice.

Funding: The study received funding from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) (Research Unit COSIMA, Project Number FOR2385).

Conflicts of interests: None.

Ethical approval: The study obtained ethical approval (No. 18-302) from the Ethics Committee of the Medical Faculty of LMU Munich. Participation in the study was voluntary and participants gave informed consent.

References

Barrows, H. S., Norman, G. R., Neufeld, V. R. & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medical practice. Clinical and Investigative Medicine, 5, 49–55

Bauer, D., Holzer, M., Kopp, V. & Fischer, M. R. (2011). Pick-N multiple choice-exams: A comparison of scoring algorithms. Advances in Health Sciences Education, 16, 211–221. DOI 10.1007/s10459-010-9256-1

Boshuizen, H. P. A. & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. Cognitive Science, 16, 153–184. DOI 10.1016/0364-0213(92)90022-M

Boulet, J. R. & Durning, S. J. (2019). What we measure … and what we should measure in medical education. Medical Education, 53, 86–94. DOI 10.1111/medu.13652

Charlin, B., Boshuizen, H. P. A., Custers, E. J. & Feltovich, P. J. (2007). Scripts and clinical reasoning. Medical Education, 41, 1178–1184. DOI 10.1111/j.1365-2923.2007.02924.x

Charlin, B., Roy, L., Brailovsky, C., Goulet, F. & van der Vleuten, C. (2000). The Script Concordance test: A tool to assess the reflective clinician. Teaching and Learning in Medicine, 12, 189–195. DOI 10.1207/s15328015tlm1204_5

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T. & Fischer, F. (2019). Facilitating diagnostic competences in higher education - A meta-analysis in medical and teacher education. Educational Psychology Review, 68, 157–196. DOI 10.1007/s10648-019-09492-2

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T. & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. Review of Educational Research, 90, 499–541. DOI 10.3102/0034654320933544

Coderre, S., Wright, B. & McLaughlin, K. (2010). To think is good: Querying an initial hypothesis reduces diagnostic error in medical students. Academic Medicine, 85, 1125–1129. DOI 10.1097/ACM.0b013e3181e1b229

Cook, D. A., Erwin, P. J. & Triola, M. M. (2010). Computerized virtual patients in health professions education: A systematic review and meta-analysis. Academic Medicine, 85, 1589–1602. DOI 10.1097/ACM.0b013e3181edfe13

Croskerry, P. (2009). A universal model of diagnostic reasoning. Academic Medicine, 84, 1022–1028. DOI 10.1097/ACM.0b013e3181ace703

Daniel, M., Rencic, J., Durning, S. J., Holmboe, E., Santen, S. A., Lang, V., Ratcliffe, T., Gordon, D., Heist, B., Lubarsky, S., Estrada, C. A., Ballard, T., Artino, A. R., Sergio Da Silva, A., Cleary, T., Stojan, J. & Gruppen, L. D. (2019). Clinical reasoning assessment methods: A scoping review and practical guidance. Academic Medicine, 94, 902–912. DOI 10.1097/ACM.0000000000002618

Elder, A. (2018). Clinical skills assessment in the twenty-first century. Medical Clinics of North America, 102, 545–558. DOI 10.1016/j.mcna.2017.12.014

Elstein, A. S. (2009). Thinking about diagnostic thinking: A 30-year perspective. Advances in Health Sciences Education, 14, 7–18. DOI 10.1007/s10459-009-9184-0

Elstein, A. S. & Schwartz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. British Medical Journal, 324, 729–732. DOI 10.1136/bmj.324.7339.729

Elstein, A. S., Shulman, L. S. & Sprafka, S. A. (1978). Medical problem solving: An analysis of clinical reasoning. (Harvard University Press: Cambridge, MA)

Elstein, A. S., Shulman, L. S. & Sprafka, S. A. (1990). Medical problem solving: A ten-year retrospective. Evaluation & the Health Professions, 13, 5–36. DOI 10.1177/016327879001300102

Eva, K. W. (2004). What every teacher needs to know about clinical reasoning. Medical Education, 39, 98–106. DOI 10.1111/j.1365-2929.2004.01972.x

Evans, D. A. & Patel, V. L. (Eds.) (1989). Cognitive science in medicine: Biomedical modeling. (MIT Press: Cambridge, MA)

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. Annual Review of Psychology, 59, 255–278. DOI 10.1146/annurev.psych.59.103006.093629

Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149–1160. DOI 10.3758/BRM.41.4.1149

Fink, M. C., Heitzmann, N., Siebeck, M., Fischer, F. & Fischer, M. R. (2021a). Learning to diagnose accurately through virtual patients: Do reflection phases have an added benefit? BMC Medical Education, 21, 523. DOI 10.1186/s12909-021-02937-9

Fink, M. C., Reitmeier, V., Siebeck, M., Fischer, F. & Fischer, M. R. (2022). Live and video simulations of medical history-taking: Theoretical background, design, development and validation of a learning environment. (In F. Fischer & A. Opitz (Eds.), Learning to diagnose with simulations: Examples from teacher education and medical education (pp. 109-122). Springer: Cham)

Fink, M. C., Reitmeier, V., Stadler, M., Siebeck, M., Fischer, F. & Fischer, M. R. (2021b). Assessment of diagnostic competences with standardized patients versus virtual patients: Experimental study in the context of history taking. Journal of Medical Internet Research, 23, e21196. DOI 10.2196/21196

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M., Girwidz, R., Obersteiner, A., Reiss, K., Stürmer, K., Siebeck, M., Schmidmaier, R., Seidel, T., Ufer, S., Wecker, C. & Neuhaus, B. J. (2018). Systematizing professional knowledge of medical doctors and teachers: Development of an interdisciplinary framework in the context of diagnostic competences. Education Sciences, 8, 207. DOI 10.3390/educsci8040207

Groves, M., O'Rourke, P. & Alexander, H. (2003). Clinical reasoning: The relative contribution of identification, interpretation and hypothesis errors to misdiagnosis. Medical Teacher, 25, 621–625. DOI 10.1080/01421590310001605688

Heitzmann, N., Fischer, M. R. & Fischer, F. (2017). Towards more systematic and better theorised research on simulations. Medical Education, 51, 129–131. DOI 10.1111/medu.13239

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., Schmidmaier, R., Neuhaus, B. J., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R. & Fischer, F. (2019). Facilitating diagnostic competences in simulations in higher education: A framework and a research agenda. Frontline Learning Research, 7, 1–24. DOI 10.14786/flr.v7i4.384

Hrynchak, P., Glover Takahashi, S. & Nayer, M. (2014). Key-feature questions for assessment of clinical reasoning: A literature review. Medical Education, 48, 870–883. DOI 10.1111/medu.12509

Huwendiek, S., de Leng, B. A., Zary, N., Fischer, M. R., Ruiz, J. G. & Ellaway, R. (2009). Towards a typology of virtual patients. Medical Teacher, 31, 743–748. DOI 10.1080/01421590903124708

Instruct. (2021). CASUS. Retrieved June 16, 2022, from https://www.instruct.eu/

Kassirer, J. P. (2010). Teaching clinical reasoning: Case-based and coached. Academic Medicine, 85, 1118–1124. DOI 10.1097/acm.0b013e3181d5dd0d

Kiesewetter, J., Sailer, M., Jung, V. M., Schönberger, R., Bauer, E., Zottmann, J. M., Hege, I., Zimmermann, H., Fischer, F. & Fischer, M. R. (2020). Learning clinical reasoning: How virtual patient case format and prior knowledge interact. BMC Medical Education, 20, 73. DOI 10.1186/s12909-020-1987-y

Leblanc, V. R., Brooks, L. R. & Norman, G. R. (2002). Believing is seeing: The influence of a diagnostic hypothesis on the interpretation of clinical features. Academic Medicine, 77, S67-S69. DOI 10.1097/00001888-200210001-00022

Leblanc, V. R., Norman, G. R. & Brooks, L. R. (2001). Effect of a diagnostic suggestion on diagnostic accuracy and identification of clinical features. Academic Medicine, 76, S18-S20. DOI 10.1097/00001888-200110001-00007

Mamede, S. & Schmidt, H. G. (2014). The twin traps of overtreatment and therapeutic nihilism in clinical practice. Medical Education, 48, 34–43. DOI 10.1111/medu.12264

Neufeld, V. R., Norman, G. R., Feightner, J. W. & Barrows, H. S. (1981). Clinical problem-solving by medical students: A cross-sectional and longitudinal analysis. Medical Education, 15, 315–322. DOI 10.1111/j.1365-2923.1981.tb02495.x

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. Medical Education, 39, 418–427. DOI 10.1111/j.1365-2929.2005.02127.x

R Core Team. (2020). R: A language and environment for statistical computing. Retrieved June 16, 2022, from https://www.R-project.org/

Ramsey, P. G., Curtis, J. R., Paauw, D. S., Carline, J. D. & Wenrich, M. D. (1998). History-taking and preventive medicine skills among primary care physicians: An assessment using standardized patients. The American Journal of Medicine, 104, 152–158. DOI 10.1016/S0002-9343(97)00310-0

Reips, U. D. (2000). The web experiment method: Advantages, disadvantages, and solutions. (In M. H. Birnbaum (Ed.), Psychological experiments on the internet (pp. 89-118). Academic Press: San Diego, CA)

Roter, D. L. & Hall, J. A. (1987). Physicians' interviewing styles and medical information obtained from patients. Journal of General Internal Medicine, 2, 325–329. DOI 10.1007/BF02596168

Round, J., Conradi, E. & Poulton, T. (2009). Improving assessment with virtual patients. Medical Teacher, 31, 759–763. DOI 10.1080/01421590903134152

Ryall, T., Judd, B. K. & Gordon, C. J. (2016). Simulation-based assessments in health professional education: A systematic review. Journal of Multidisciplinary Healthcare, 9, 69–82. DOI 10.2147/JMDH.S92695

Schmidmaier, R., Eiber, S., Ebersbach, R., Schiller, M., Hege, I., Holzer, M. & Fischer, M. R. (2013). Learning the facts in medical school is not enough: Which factors predict successful application of procedural knowledge in a laboratory setting? BMC Medical Education, 13, 28. DOI 10.1186/1472-6920-13-28

Schmidt, H. G., Norman, G. & Boshuizen, H. (1990). A cognitive perspective on medical expertise: Theory and implications. Academic Medicine, 65, 611–621. DOI 10.1097/00001888-199010000-00001

Schmidt, H. G. & Rikers, R. M. J. P. (2007). How expertise develops in medicine: Knowledge encapsulation and illness script formation. Medical Education, 41, 1133–1139. DOI 10.1111/j.1365-2923.2007.02915.x

Stark, R., Kopp, V. & Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: Two experimental studies in undergraduate medical education. Learning and Instruction, 21, 22–33. DOI 10.1016/j.learninstruc.2009.10.001

Stillman, P. L., Swanson, D. B., Regan, M. B., Philbin, M. M., Nelson, V., Ebert, T., Ley, B., Parrino, T., Shorey, J., Stillman, A., Alpert, E., Caslowitz, J., Clive, D., Florek, J., Hamolsky, M., Hatem, C., Kizirian, J., Kopelman, R., Levenson, D., Levinson, G., McCue, J., Pohl, H., Schiffman, F., Schwartz, J., Thane, M. & Wolf, M. (1991). Assessing clinical skills of residents utilizing standardized patients. Annals of Internal Medicine, 114, 393–401. DOI 10.7326/0003-4819-105-5-762

Woolliscroft, J. O., Calhoun, J. G., Billiu, G. A., Stross, J. K., MacDonald, M. & Templeton, B. (1989). House officer interviewing techniques. Journal of General Internal Medicine, 4, 108–114. DOI 10.1007/BF02602349

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, 990–995. DOI 10.1097/ACM.0000000000002142

Appendices

Appendix S1. History-taking questions

The history-taking menu has the following categories *main symptoms* (MS), *prior history* (PH), *allergies and medication* (AM), *social and family history* (SF), and *system review* (SR). The categories were adapted from Bornemann (2016).

Table 1 History-taking questions

Category	**Code**	**Example**
Main symptoms	MS	Do you experience the complaints for the first time?
Prior history	PH	Do you know of any pre-existing conditions?
Allergies and medication	AM	Do you frequently have infections against which you take antibiotics?
Social and family history	SF	Have your parents or other relatives of your family passed away at a rather young age?
System review	SR	Has your weight changed within the last weeks?

An overview of all 69 included questions is available in Fink et al. (2021).

Sources

Bornemann, B. (2016). *Dokumentationsbögen der Inneren Medizin und der Chirurgie für Anamnese und körperliche Untersuchung für die studentische Lehre in Deutschland* (Diss., Institut für Didaktik und Ausbildungsforschung in der Medizin der Ludwig-Maximilians-Universität München). Retrieved from https://edoc.ub.uni-muenchen.de/19166/

Fink, M. C., Reitmeier, V., Stadler, M., Siebeck, M., Fischer, F., Fischer, M. R. (2021). Assessment of diagnostic competences with standardized patients versus virtual patients: Experimental study in the context of history taking. *Journal of Medical Internet Research*, *23*(3), e21196. https://doi.org/10.2196/21196

Appendix S2. Principal component analysis for the comprehensive diagnostic score

Table 1 Component loadings

	Component	
	1	Uniqueness
Diagnostic accuracy	0.75	0.44
Treatment selected	0.78	0.39
Expected findings in a physical examination	0.74	0.45
Diagnostic measures taken for medical clarification	0.85	0.28

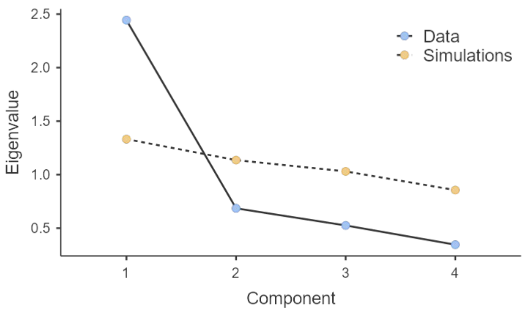
'varimax' rotation was used

Table 2 Component statistics summary

Component	SS Loadings	% of Variance	Cumulative %
1	2.44	61.09	61.09

Table 3 Initial Eigenvalues

Component	Eigenvalue	% of Variance	Cumulative %
1	2.44	61.09	61.09
2	0.69	17.14	78.22
3	0.53	13.13	91.36
4	0.35	8.64	100.00



Scree Plot

Appendix S3. Diagnoses of the virtual patients

Table 1 Diagnoses of the virtual patients

Case number in the study	Case number in the paper	Diagnosis	Patient characteristics	Patient name
1	1	Hypertrophic cardiomyopathy	25 years, male	Mr. Albrecht
2	2	Pneumonia	55 years, female	Ms. Klein
3	_	Pulmonary embolism in case of prostate cancer	70 years, male	Mr. Wagner
4	3	Pulmonary embolism with coagulation disorder	35 years, female	Ms. Aimüller
5	_	Heart insufficiency with thoracic aortic aneurysm	65 years, female	Ms. Bircher
6	4	Panic attack	45 years, male	Mr. Lehner

We discovered floor effects on the diagnostic success measures diagnostic accuracy, and treatment selected (see Table 2). Therefore, these cases were excluded from our study.

Table 2 Descriptive statistics of the diagnostic success measures of the excluded cases

	Mr. Wagner	Ms. Bircher
Diagnostic accuracy	0.04 (0.14)	0.08 (0.23)
Treatment selected	0.01 (0.10)	0.09 (0.28)
Expected findings in a physical examination	0.38 (0.16)	0.42 (0.20)
Diagnostic measures taken for medical clarification	0.12 (0.11)	0.43 (0.19)

Appendix S4. Control analysis for data collection

Table 1 Descriptives for professional knowledge, diagnostic activities and diagnostic success measures across both data collection groups and results of independent samples t-tests

Variable	**Web-based data collection *M* (*SD*)**	**Lab-based data collection *M* (*SD*)**	***t***	***df***	***p***
Conceptual knowledge	0.50 (0.13)	0.62 (0.13)	-4.36	104	< .001
Strategic knowledge	0.47 (0.14)	0.58 (0.13)	-3.85	104	< .001
Hypothesis generation	0.36 (0.19)	0.34 (0.18)	0.59	104	0.554
Evidence generation	0.41 (0.12)	0.51 (0.11)	-3.53	104	< .001
Evidence evaluation	0.48 (0.11)	0.50 (0.12)	-0.89	104	0.374
Diagnostic accuracy	0.39 (0.21)	0.54 (0.23)	-3.11	104	0.002
Comprehensive diagnostic score	-0.14 (0.75)	0.34 (0.73)	-2.99	104	0.004

Diagnostic quality was a normalized z-score ranging from -3 to +3. Scale range for all other variables: 0-1. Note that the scores for diagnostic quality, hypothesis generation, and evidence evaluation were aggregated over three virtual patients.

Appendix S5. Regression analyses including data collection as factor

Table 1 Regression analyses for the comprehensive diagnostic score as outcome including the data collection method as categorical variable

Predictor	b	ß	p
Intercept	-3.10		< .001
Data collection	0.96	0.22	.303
Conceptual knowledge	1.01	0.18	.111
Strategic knowledge	0.77	0.14	.240
Hypothesis generation	1.21	0.29	.002
Evidence evaluation	1.88	0.28	.006
Evidence generation	1.79	0.30	.006
Conceptual knowledge ✻ Data collection	-0.04	-0.01	.978
Strategic knowledge ✻ Data collection	-0.02	-0.00	.986
Hypothesis generation ✻ Data collection	-0.07	-0.02	.934
Evidence generation ✻ Data collection	-0.95	-0.16	.460
Evidence evaluation ✻ Data collection	-0.64	-0.09	.629

F(11, 94) = 6.30, p < .001, R^2 = .42

Table 2 Regression analyses for diagnostic accuracy as outcome including the data collection method as categorical variable

Predictor	b	ß	p
Intercept	-0.03		.832
Data collection	0.23	0.62	.433
Conceptual knowledge	0.17	0.11	.387
Strategic knowledge	0.05	0.03	.803
Hypothesis generation	0.41	0.34	.001
Evidence evaluation	0.12	0.06	.566
Evidence generation	0.25	0.14	.213
Conceptual knowledge ✻ Data collection	-0.25	-0.15	.564
Strategic knowledge ✻ Data collection	0.14	0.09	.745
Hypothesis generation ✻ Data collection	0.43	0.35	.116
Evidence generation ✻ Data collection	-0.27	-0.15	.508
Evidence evaluation ✻ Data collection	-0.13	-0.06	.761

F(11, 94) = 4.09, p < .001, R^2 = .32

4. Article 3: Learning to Diagnose Accurately through Virtual Patients: Do Reflection Phases Have an Added Benefit?

Reference

Fink, M. C., Heitzmann, N., Siebeck, M., Fischer, F., & Fischer, M. R. (2021). Learning to diagnose accurately through virtual patients: Do reflection phases have an added benefit*? BMC Medical Education*, *21*, 523. https://doi.org/10.1186/s12909-021-02937-9

Fink *et al. BMC Medical Education* (2021) 21:523
https://doi.org/10.1186/s12909-021-02937-9

BMC Medical Education

RESEARCH Open Access

Learning to diagnose accurately through virtual patients: do reflection phases have an added benefit?

Maximilian C. Fink[1,2*], Nicole Heitzmann[3,4], Matthias Siebeck[1,4], Frank Fischer[3,4] and Martin R. Fischer[1,4]

Abstract

Background: Simulation-based learning with virtual patients is a highly effective method that could potentially be further enhanced by including reflection phases. The effectiveness of reflection phases for learning to diagnose has mainly been demonstrated for problem-centered instruction with text-based cases, not for simulation-based learning. To close this research gap, we conducted a study on learning history-taking using virtual patients. In this study, we examined the added benefit of including reflection phases on learning to diagnose accurately, the associations between knowledge and learning, and the diagnostic process.

Methods: A sample of $N = 121$ medical students completed a three-group experiment with a control group and pre- and posttests. The pretest consisted of a conceptual and strategic knowledge test and virtual patients to be diagnosed. In the learning phase, two intervention groups worked with virtual patients and completed different types of reflection phases, while the control group learned with virtual patients but without reflection phases. The posttest again involved virtual patients. For all virtual patients, diagnostic accuracy was assessed as the primary outcome. Current hypotheses were tracked during reflection phases and in simulation-based learning to measure diagnostic process.

Results: Regarding the added benefit of reflection phases, an ANCOVA controlling for pretest performance found no difference in diagnostic accuracy at posttest between the three conditions, $F(2, 114) = 0.93$, $p = .398$. Concerning knowledge and learning, both pretest conceptual knowledge and strategic knowledge were not associated with learning to diagnose accurately through reflection phases. Learners' diagnostic process improved during simulation-based learning and the reflection phases.

Conclusions: Reflection phases did not have an added benefit for learning to diagnose accurately in virtual patients. This finding indicates that reflection phases may not be as effective in simulation-based learning as in problem-centered instruction with text-based cases and can be explained with two contextual differences. First, information processing in simulation-based learning uses the verbal channel and the visual channel, while text-based learning only draws on the verbal channel. Second, in simulation-based learning, serial cue cases are used to gather information step-wise, whereas, in text-based learning, whole cases are used that present all data at once.

Keywords: Reflection Phases, Diagnostic Competences, Simulation, Medical Education

* Correspondence: Maximilian.Fink@med.uni-muenchen.de
[1]Institute of Medical Education, University Hospital, LMU Munich, Pettenkoferstr. 8a, 80336 Munich, Germany
[2]Institute of Education, Universität der Bundeswehr München, Neubiberg, Germany
Full list of author information is available at the end of the article

Introduction

A recent meta-analysis revealed that simulation-based learning has a large positive effect on learning complex skills, including diagnostic competences in medicine [1]. Moreover, there is evidence that the positive effects of simulation-based learning may be enhanced by combining it with instructional support measures [2–4]. Indeed, numerous studies have confirmed that reflection phases are a particularly effective type of instructional support [5–7]. However, a closer inspection of these studies shows that reflection phases were primarily investigated for text-based cases and not for simulation-based learning with virtual patients. Therefore, to what extent reflection phases can foster learning to diagnose accurately in simulation-based learning is an open question. . Below, we summarize our study's underlying conceptual framework, define virtual patients and text-based cases, and discuss the potential effect of reflection on facilitating diagnostic competences.

Underlying conceptual framework

Our study is based on the conceptual framework for acquiring diagnostic competences in simulations with instructional support by Heitzmann et al. [8]. They define simulations as models of diagnostic situations that can be manipulated and sometimes even controlled by participants. The instructional support provided can include, for instance, examples, prompts, or reflection phases. The effectiveness of simulation-based learning with instructional support depends on individuals' diagnostic process and prerequisites such as prior knowledge. The diagnostic process can be operationalized through eight diagnostic activities, including the *current hypothesis* (preliminary diagnosis) learners form in the course of diagnosing. Knowledge encompasses the two types *conceptual knowledge* and *strategic knowledge*. Conceptual knowledge refers to knowledge about constructs and their relations, while strategic knowledge is defined as knowledge about heuristics and strategies in diagnosing. The primary outcome measure of simulation-based learning in this framework is *diagnostic accuracy* - the agreement between the participant's diagnosis and a correct sample solution [8]. Next, we will define virtual patients and text-based cases and briefly describe possible differences in information processing while learning from them.

Virtual patients and text-based cases

Virtual patients are a special type of computer simulation representing clinical situations such as history-taking or physical examinations [9]. Moreover, virtual patients frequently include audio-visual material as well as text-based information [10, 11]. The *text-based cases* used in studies on reflection phases typically consist of a description of the patient's main symptoms, as well as relevant findings from history-taking, the physical examination, and lab investigations [5–7]. Two theoretical perspectives suggest that information processing during learning from virtual patients and text-based cases may differ. According to the cognitive theory of multimedia learning [12], humans possess two separate channels for visual and verbal information processing that are used during learning [13]. Consequently, learners will process virtual patients using both channels, while text-based cases will only be processed in the channel for verbal information. Moreover, differences in the case formats could determine how information is processed [11, 14]. Virtual patients typically represent *serial cue* cases, in which information is obtained step-wise by navigating through a digital environment. In text-based cases, information is typically presented in the *whole case* format, in which all relevant information is displayed at once. In the following section, we will discuss the potential effect of reflection on facilitating diagnostic competences.

Reflection and facilitating diagnostic competences

Reflection is defined as a cognitive and metacognitive process in which learners deal with their thoughts and actions, as well as their bases, intending to modify them [15]. On the one hand, reflective processes can implicitly occur in virtual patients containing design features that provide opportunities for this. On the other hand, *reflection phases* as instructional support can explicitly induce beneficial reflective processes by providing specific instructions and a dedicated phase of time for this activity. In medical diagnosing interventions, the instructions for reflection phases typically include questions on the initial hypothesis, alternative hypotheses, and reasons for and against these hypotheses [5, 7, 16]. The effectiveness of reflection has been primarily tied to dual-process theory [17], which claims that two cognitive systems are used in diagnosing: a fast, heuristic, and a slow, reflective system. In line with Mamede et al. [18], reflection phases induce slow cognitive processes that could be particularly beneficial for correcting mistakes caused by faulty heuristic diagnosing. Current research on reflection phases centers around (1) the effectiveness of reflection, (2) the associations between prior knowledge and learning from reflection, and (3) the quality of the diagnostic process.

Concerning (1) the effectiveness of reflection, a meta-analysis on instructional support in problem-centered instruction in the domains of medical education and teacher education reported a medium positive effect ($g = 0.58$) of including reflection phases on promoting diagnostic competences [2]. In addition, a literature review for medical education by Mamede et al. [18] found that reflection phases facilitated diagnostic competences in most studies that used them to validate diagnoses with specific reasoning instructions. At this point, it should be noted that the medical education literature primarily investigated the effect of reflection phases for learning from text-based cases, while results for simulation-based learning are lacking. In contrast, a cross-domain meta-analysis focused on simulation-based learning discovered no added benefit of including reflection phases on fostering complex skills [1]. In sum, there is more evidence that reflection phases are effective than not effective at fostering diagnostic competences in medicine. Despite opposing findings from other domains, we currently assume that this is also true in the context of simulations.

The (2) associations between prior knowledge and learning from reflection should also be examined. Support for this association comes from the two aforementioned meta-analyses on instructional support in problem-centered instruction and simulation-based learning [1, 2], which both showed that reflection phases were more beneficial for college students with high prior knowledge than with low prior knowledge. In these meta-analyses, learners' prior knowledge was measured dichotomously (low vs. high) based on years of academic training and content familiarity. In partial contrast to these results, an experiment by Mamede et al. [19] demonstrated that physicians in specialty training but not undergraduate medical students benefitted from conscious, slow diagnostic thinking when solving complex problems. The authors argued that the undergraduate college students in their study did not possess the necessary knowledge foundation to experience improvement through reflective processes. In short, the literature indicates that learners with higher prior knowledge benefit more from reflection phases than learners with lower prior knowledge. However, further research on the level of expertise required to profit from reflection phases is necessary.

Two topics concerning the quality of the diagnostic process (3) should be investigated further. First, the diagnostic process during reflection phases should be examined by inspecting learners' hypotheses. Mamede et al. [20] showed that hypotheses improved from a first point in the diagnostic process before reflection to a second point in the diagnostic process after reflection. In their study, four different types of reflection phases (no specific instructions, arguments for the diagnosis, arguments against the diagnosis, and arguments for and against the diagnosis) were applied to text-based cases. As is the case during reflection phases, learners might also be able to enhance their hypotheses over the course of simulation-based learning without reflection phases by gathering and interpreting additional data [8]. Second, the optimal timing of reflection phases within the diagnostic process should be analyzed. Initial evidence highlights that reflection phases are particularly effective *during* rather than before or after diagnosing [18]. However, two different operationalizations of reflection phases during diagnosing are conceivable: In *accompanying reflection*, learners reflect in the middle of a case and then continue working on it before providing a final diagnosis. In *concluding reflection*, learners reflect after completing a case, right before providing a final diagnosis. Each type of reflection phases could have specific benefits. Accompanying reflection could primarily help learners plan and monitor their ongoing diagnostic process in the sense of improved self-regulated learning [21]. Concluding reflection could offer learners more case information to reconsider in the sense of self-generated feedback to be used in problem-solving [22]. In light of the potential benefits of accompanying reflection over concluding reflection for the diagnostic process, we assume that this type of instructional support is particularly effective for virtual patients with serial cue cases.

Research questions and hypotheses

To investigate reflection phases in the context of simulations, we address the following research questions: To what extent do reflection phases affect learning to diagnose accurately in virtual patients? (RQ1) We hypothesize that the inclusion of reflection phases in simulation-based learning has an added benefit for learning to diagnose accurately (H1.1). Furthermore, we assume that accompanying reflection is more beneficial for learning to diagnose accurately than concluding reflection (H1.2). To what extent is prior knowledge associated with learning to diagnose accurately through reflection phases? (RQ2) We expect that learners with higher conceptual (H2.1) and strategic (H2.2) knowledge would experience greater improvement in diagnostic accuracy than learners with lower prior knowledge of these types. To what extent does the diagnostic process improve during simulation-based learning with virtual patients and during reflection phases, in the sense of enhancements in current hypotheses and diagnostic accuracy over the course of cases? (RQ3) We assume that the diagnostic process improves both during simulation-based learning (H3.1) and reflection phases (H3.2).

Method

Sampling procedure, participants, and research design

Data collection for the study ran from October 2019 to February 2021. Recruitment took place on-campus and through online advertising. Medical students from LMU Munich with high German language proficiency in their third to fifth year of medical school were eligible. The final sample consisted of $N = 121$ participants with an average age of $M = 24.90$ years, $SD = 4.01$ years. The gender of participants was distributed as follows: $n = 82$

(67.7 %) female, *n* = 10 (8.3 %) male, and *n* = 29 (24.0 %) no answer. The high proportion of participants with no answer on gender was likely caused by the use of an electronic form that allowed skipping this question without selecting an option. The final sample represents about 5 % of the enrolled third to fifth year medical students from LMU Munich and is representative in age for this population. We report more details on the sampling and participants in Additional file 1: Appendix S1 and S2.

The study used a pretest-posttest design, varying the type of reflection. Participants were randomly assigned to one of three conditions: (1) concluding reflection (*n* = 42), (2) accompanying reflection (*n* = 39), and (3) control group (*n* = 40). Data collection moved from the lab to the web in the middle of the study due to the COVID-19 pandemic. In both types of data collection, an identical learning environment was used. In lab-based data collection, an experimenter was present in the computer room at the university hospital. In web-based data collection, an experimenter was connected via video chat. The proportion of participants experiencing each data collection method across conditions are provided in Additional file 1: Appendix S2. A chi-square test showed that the proportions participants experiencing each data collection method did not differ across the conditions, $\chi^2(2, N = 121) = 0.01, p = .994$.

Procedure

We provide a visualization of the procedure for the different conditions in Fig. 1. Participants began the pretest by completing the conceptual and strategic knowledge tests to assess their prior knowledge. The conceptual and strategic knowledge tests are described in more detail later. Next, participants completed a familiarization with the simulation-based learning environment and then diagnosed three virtual pretest patients. During the learning phase, all participants solved three other virtual patients. In all conditions, participants were reminded via prompts to spend a minimum of 5 min on each simulation and had to stop working on the simulation after a time limit of 10 min. We selected the time limit of 10 min based on a prior study using similar cases [23]. Our goal was to provide sufficient time for diagnosing with an efficiency mindset but without inducing severe time pressure. In the accompanying reflection condition, a reflection phase took place halfway through each case. In the concluding reflection condition, a reflection phase was conducted after completing each case but before providing a final diagnosis. Moreover, only during the learning phase and in all conditions, including the control group, a video-based expert solution was presented after fully completing and diagnosing each virtual patient. The expert solution contained the correct diagnosis and strategic knowledge on the correct diagnostic process. In the posttest, participants completed three additional virtual patients.

Materials

Virtual patients

Participants diagnosed nine virtual patients suffering from different causes of dyspnea. The virtual patients were validated in a study by Fink et al. [23]. In Additional file 1: Appendix S3, we provide an overview of the diagnoses and characteristics of the virtual patients. The virtual patients of the learning phase were selected so that a transfer to the virtual patients of the pre- and posttest was possible. In fact, the learning phase contained various cardiopulmonary perfusion and diffusion problems that shared a common hypothesis space with the pre- and posttest. The (semi)-professional actors playing the patients were selected based on the virtual patients' characteristics and trained for their role by an acting coach and a physician. The created virtual patients were then embedded into the digital learning environment CASUS [24]. We present a screenshot of one of the virtual patients in Fig. 2.

At the beginning of each virtual patient encounter, prior diagnostic information (e.g., lab results) and the chief complaint were presented in an introductory video. Then, participants took the patient´s history by selecting from a menu of 69 questions (cf. the questions on the left of Fig. 2). The answer to each selected history-taking question was streamed as a short video. Additional file 1: Appendix S3 provides examples of the history-taking questions used and a source for the complete list of history-taking questions.

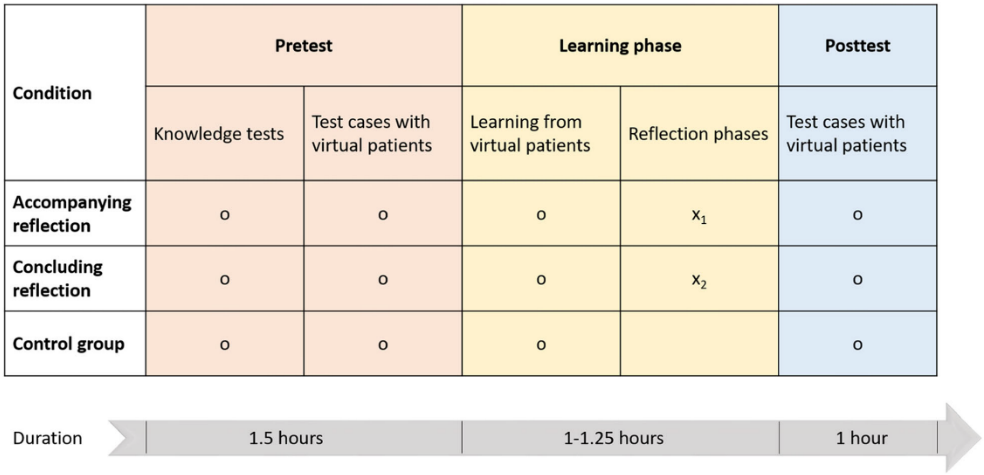
Reflection phases

The content for the accompanying and concluding reflection conditions were based on scripts developed by Mamede et al. [5, 7, 16]. As previously mentioned, in *accompanying reflection*, learners reflected after 5 min equaling halfway through working on a case. In *concluding reflection*, learners reflected after completing the case, right before offering their final diagnosis. The scripts for both types of reflection consisted of nine questions and are documented in Additional file 1: Appendix S4. Participants received 4 min and 20 s within each case to engage in reflection.

Instruments

Diagnostic accuracy

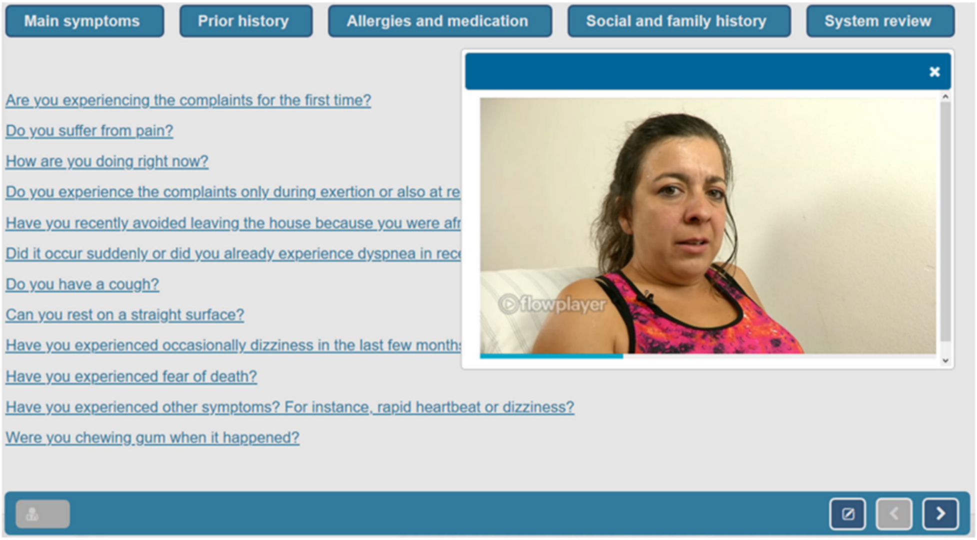
Diagnostic accuracy was measured in each virtual patient with a long menu consisting of 180 possible diagnoses related to dyspnea. Participants selected one diagnosis per case, which was compared to a solution. One point was



Condition	Pretest		Learning phase		Posttest
	Knowledge tests	Test cases with virtual patients	Learning from virtual patients	Reflection phases	Test cases with virtual patients
Accompanying reflection	o	o	o	x_1	o
Concluding reflection	o	o	o	x_2	o
Control group	o	o	o		o
Duration	1.5 hours		1-1.25 hours		1 hour

Fig. 1 Illustration of the study procedure, including approximate durations. Note on the symbols: o Indicates a measurement, x Indicates a treatment. Details on the intervention: X_1: Reflection halfway through each case, X_2: Reflection after completing each case

awarded for the designated correct answer, 0.50 points for a partially correct answer, and 0 points for all other diagnoses. The learners' answers were compared using *R* scripts to the common sample solution of two expert physicians validated in Fink et al. [23]. Mean scores for diagnostic accuracy were calculated for the pretest, posttest, and the learning phase and ranged from 0 (*low*) to 1 (*high*). The third case in the pretest (diagnostic accuracy $M = 0.05$, $SD = 0.14$) and the second case in the posttest ($M = 0.08$, $SD = 0.23$) were excluded from our analyses because of floor effects (see Additional file 1: Appendix S3 for the diagnoses in these cases).



Are you experiencing the complaints for the first time?
Do you suffer from pain?
How are you doing right now?
Do you experience the complaints only during exertion or also at re
Have you recently avoided leaving the house because you were afr
Did it occur suddenly or did you already experience dyspnea in rec
Do you have a cough?
Can you rest on a straight surface?
Have you experienced occasionally dizziness in the last few months
Have you experienced fear of death?
Have you experienced other symptoms? For instance, rapid heartbeat or dizziness?
Were you chewing gum when it happened?

Fig. 2 Virtual patient by Fink et al. [23] licensed under CC BY 4.0

Current hypothesis in the diagnostic process

To assess participants' current hypothesis in the diagnostic process, we proceeded as follows. We asked participants in every condition to select their current hypothesis for each patient from the same long menu described for diagnostic accuracy directly after reading the prior diagnostic information and watching the chief complaint on video. Moreover, participants' current hypothesis was additionally measured at the start and the end of each type of reflection.

Conceptual knowledge test

The conceptual knowledge tests focused on dyspnea and history-taking. The test consisted of 20 items and contained single-choice and pick-N multiple-choice questions. In single-choice questions, participants received one point for the correct answer. In pick-N multiple-choice questions, participants received one point if their entire answer pattern was correct. If participants selected more than 50 % correct answers in a pick-N multiple-choice question, they were awarded 0.50 points, in line with Bauer et al. [25]. Conceptual knowledge scores were determined by dividing the number of points achieved by the number of questions posed. Thus, conceptual knowledge scores ranged from 0 (*low knowledge*) to 1 (*high knowledge*). The time limit for the test was set to 20 min. The reliability was acceptable, with Cronbach´s α = 0.61.

Strategic knowledge test

Strategic knowledge on dyspnea and history-taking was assessed with four key feature cases [26]. Each case consisted of four single-choice questions regarding the diagnosis, treatment, symptoms, and further diagnostic measures. One point was awarded for each correct answer. Strategic knowledge test scores were calculated by dividing the number of points achieved by the number of questions posed. Therefore, strategic knowledge scores ranged from 0 (*low knowledge*) to 1 (*high knowledge*). Testing time was set to 20 min. The test's reliability was acceptable, with Cronbach´s α = 0.65.

Cognitive load

Cognitive load was assessed as a control variable once directly after the end of the learning phase. We measured this variable as a control variable because a negative association between cognitive load and performance in medical skills, such as diagnosing, has been shown repeatedly [27]. Moreover, reflection phases could affect the cognitive load present in the different experimental conditions. We used for the assessment of cognitive load a five-item, five-point scale by Opfermann [28]. The scale differentiates between germane, extraneous, and intrinsic cognitive load and lets participants rate their mental effort from (1) very low to (5) very high.

Manipulation checks

One manipulation check on duration showed that, as intended, participants in the intervention groups spent about four additional minutes on the reflection phase for each case (see Additional file 1: Appendix S5). We consider this sufficient time for reflection in cases with a time limit of 10 min. Another manipulation check confirmed that participants successfully engaged in reflection by writing a sufficient amount of notes in our digital environment (see Additional file 1: Appendix S5).

Statistical analyses and sample size

We used R (Version 4.0.2) [29] for the statistical analyses. We investigated RQ1 with an analysis of covariance. RQ2 was examined with one-tailed Pearson correlations. For RQ3, we used one-tailed paired sample t-tests. In all statistical analyses, the significance level was set to α = 0.05.

An a priori-power analysis was conducted with G*Power (Version 3.1) [30], assuming an error probability of α = 0.05 and a power of β = 0.80. For the main analysis of RQ1, we hypothesized that the effect of reflection phases on learning to diagnose accurately would be medium-sized, with $g = 0.58$, based on the meta-analysis by Chernikova et al. [2]. Based on this assumed effect size, the power analysis yielded a required sample size of $N = 118$ participants with 39 participants per group.

Results

Preliminary analyses

We report descriptive statistics and results from a one-way analysis of variance for knowledge, diagnostic accuracy, and cognitive load in Table 1. These results show that knowledge and diagnostic accuracy did not differ across the experimental conditions in the different phases of the experiment. Similarly, cognitive load control variables did not differ across the experimental conditions when they were measured directly after the learning phase.

The effect of reflection phases on learning to diagnose accurately (RQ1)

To answer RQ1, we conducted an analysis of covariance using the diagnostic accuracy score from the posttest as the outcome. After adjustment for pretest diagnostic accuracy, there was no statistically significant difference in posttest diagnostic accuracy between the conditions, $F(2, 114) = 0.93$, $p = .398$, $\eta p^2 = 0.02$. Thus, H1.1, an added benefit of reflection phases on learning to diagnose accurately, could not be confirmed. A pairwise comparison

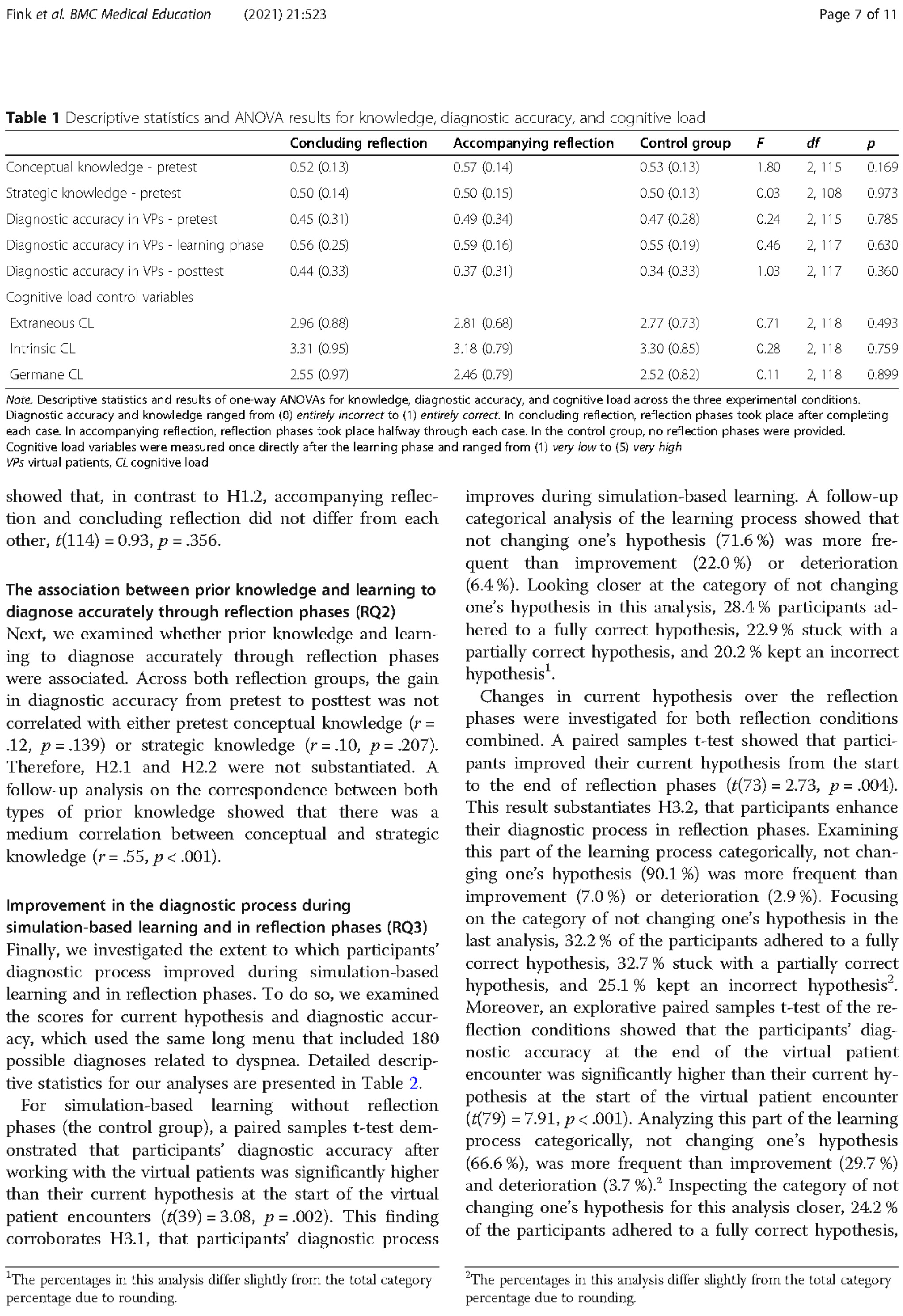


Table 1 Descriptive statistics and ANOVA results for knowledge, diagnostic accuracy, and cognitive load

	Concluding reflection	Accompanying reflection	Control group	*F*	*df*	*p*
Conceptual knowledge - pretest	0.52 (0.13)	0.57 (0.14)	0.53 (0.13)	1.80	2, 115	0.169
Strategic knowledge - pretest	0.50 (0.14)	0.50 (0.15)	0.50 (0.13)	0.03	2, 108	0.973
Diagnostic accuracy in VPs - pretest	0.45 (0.31)	0.49 (0.34)	0.47 (0.28)	0.24	2, 115	0.785
Diagnostic accuracy in VPs - learning phase	0.56 (0.25)	0.59 (0.16)	0.55 (0.19)	0.46	2, 117	0.630
Diagnostic accuracy in VPs - posttest	0.44 (0.33)	0.37 (0.31)	0.34 (0.33)	1.03	2, 117	0.360
Cognitive load control variables						
Extraneous CL	2.96 (0.88)	2.81 (0.68)	2.77 (0.73)	0.71	2, 118	0.493
Intrinsic CL	3.31 (0.95)	3.18 (0.79)	3.30 (0.85)	0.28	2, 118	0.759
Germane CL	2.55 (0.97)	2.46 (0.79)	2.52 (0.82)	0.11	2, 118	0.899

Note. Descriptive statistics and results of one-way ANOVAs for knowledge, diagnostic accuracy, and cognitive load across the three experimental conditions. Diagnostic accuracy and knowledge ranged from (0) *entirely incorrect* to (1) *entirely correct*. In concluding reflection, reflection phases took place after completing each case. In accompanying reflection, reflection phases took place halfway through each case. In the control group, no reflection phases were provided. Cognitive load variables were measured once directly after the learning phase and ranged from (1) *very low* to (5) *very high*
VPs virtual patients, *CL* cognitive load

showed that, in contrast to H1.2, accompanying reflection and concluding reflection did not differ from each other, $t(114) = 0.93$, $p = .356$.

The association between prior knowledge and learning to diagnose accurately through reflection phases (RQ2)

Next, we examined whether prior knowledge and learning to diagnose accurately through reflection phases were associated. Across both reflection groups, the gain in diagnostic accuracy from pretest to posttest was not correlated with either pretest conceptual knowledge ($r = .12$, $p = .139$) or strategic knowledge ($r = .10$, $p = .207$). Therefore, H2.1 and H2.2 were not substantiated. A follow-up analysis on the correspondence between both types of prior knowledge showed that there was a medium correlation between conceptual and strategic knowledge ($r = .55$, $p < .001$).

Improvement in the diagnostic process during simulation-based learning and in reflection phases (RQ3)

Finally, we investigated the extent to which participants' diagnostic process improved during simulation-based learning and in reflection phases. To do so, we examined the scores for current hypothesis and diagnostic accuracy, which used the same long menu that included 180 possible diagnoses related to dyspnea. Detailed descriptive statistics for our analyses are presented in Table 2.

For simulation-based learning without reflection phases (the control group), a paired samples t-test demonstrated that participants' diagnostic accuracy after working with the virtual patients was significantly higher than their current hypothesis at the start of the virtual patient encounters ($t(39) = 3.08$, $p = .002$). This finding corroborates H3.1, that participants' diagnostic process improves during simulation-based learning. A follow-up categorical analysis of the learning process showed that not changing one's hypothesis (71.6 %) was more frequent than improvement (22.0 %) or deterioration (6.4 %). Looking closer at the category of not changing one's hypothesis in this analysis, 28.4 % participants adhered to a fully correct hypothesis, 22.9 % stuck with a partially correct hypothesis, and 20.2 % kept an incorrect hypothesis[1].

Changes in current hypothesis over the reflection phases were investigated for both reflection conditions combined. A paired samples t-test showed that participants improved their current hypothesis from the start to the end of reflection phases ($t(73) = 2.73$, $p = .004$). This result substantiates H3.2, that participants enhance their diagnostic process in reflection phases. Examining this part of the learning process categorically, not changing one's hypothesis (90.1 %) was more frequent than improvement (7.0 %) or deterioration (2.9 %). Focusing on the category of not changing one's hypothesis in the last analysis, 32.2 % of the participants adhered to a fully correct hypothesis, 32.7 % stuck with a partially correct hypothesis, and 25.1 % kept an incorrect hypothesis[2]. Moreover, an explorative paired samples t-test of the reflection conditions showed that the participants' diagnostic accuracy at the end of the virtual patient encounter was significantly higher than their current hypothesis at the start of the virtual patient encounter ($t(79) = 7.91$, $p < .001$). Analyzing this part of the learning process categorically, not changing one's hypothesis (66.6 %), was more frequent than improvement (29.7 %) and deterioration (3.7 %).[2] Inspecting the category of not changing one's hypothesis for this analysis closer, 24.2 % of the participants adhered to a fully correct hypothesis,

[1] The percentages in this analysis differ slightly from the total category percentage due to rounding.

[2] The percentages in this analysis differ slightly from the total category percentage due to rounding.

Table 2 Descriptive statistics for the diagnostic process during the learning phase

	Current hypothesis – start of VPs	**Current hypothesis – start of reflection phase**	**Current hypothesis – end of reflection phase**	**Diagnostic accuracy - end of VPs**
Control group	0.45 (0.21)	-	-	0.55 (0.19)
Reflection conditions	0.39 (0.21)	0.50 (0.21)	0.57 (0.25)	0.58 (0.21)

Note. Current hypothesis and diagnostic accuracy were measured as indicators of the diagnostic process several times during the experiment. Both variables were assessed with the same instrument and ranged from (0) *entirely incorrect* to (1) *entirely correct*. Means and SDs are provided in the table above separately for the control group and both reflection conditions (accompanying and concluding reflection) combined. For the control group, the current hypothesis at the start of the VPs and diagnostic accuracy at the end of the VPs are given. In the reflection conditions, the current hypothesis at the start and the end of the reflection phases is also reported

VPs virtual patients

21.5 % stuck with a partially correct hypothesis, and 21.0 % kept an incorrect hypothesis.

Discussion

Principal findings

Regarding the first research question (RQ1), we observed no added benefit of reflection phases for learning to diagnose accurately. This finding is not in line with the medium effects of reflection phases and other instructional supports on cognitive outcomes in problem-centered instruction [2, 4]. However, our finding corresponds to new meta-analytic results that reflection has no additional benefit for complex skills in simulation-based learning [1].

One difference between simulation-based learning and problem-centered instruction that could explain the differential effects is their average effectiveness. Simulation-based learning has a large effect on learning [1], while the effect of problem-centered instruction is moderate [2, 4]. Consequently, adding reflection to simulation-based learning might not lead to a further increase in the highly beneficial effect of simulation-based learning itself. This explanation, however, is not supported by the fact that other instructional supports and particular combinations of instructional supports demonstrated added benefits in simulation-based learning [1].

Another difference between simulation-based learning and problem-centered instruction that could influence reflection phases' effectiveness could be cognitive load. However, our control analysis on cognitive load showed that cognitive load in the virtual patients reached medium values comparable to problem-centered instruction with text-based cases [14]. Our results can be compared to the results for the text-based cases because exactly the same cognitive load scale was used in these two studies. Therefore, we can infer that cognitive load was not excessively high in our virtual patients. Moreover, cognitive load did not differ across the experimental conditions, suggesting that reflection phases did not manipulate cognitive load.

A more plausible explanation for the discovered differential effectiveness of reflection phases in simulation-based learning and problem-centered instruction concerns the case format. In simulation-based learning, serial cue cases are typically utilized, which was also true in our experiment. Serial cue cases present data in a step-wise fashion and involve interactive case construction and interpretation [11, 14]. In problem-centered instruction, text-based whole cases are typically used. Whole cases require the learner to remember and interpret all of the information that is presented [11, 14]. Comparing both case formats, it can also be argued, that serial cue cases may perhaps provide by their very nature more room for implicit reflective processes than whole cases. The lack of effect of reflection phases in our study could be explained by the differences between these case formats as follows. Reflection phases might be less effective in serial cue cases when cases are interactively constructed, and there is room for implicit reflective processes. However, reflection phases might be more effective in whole cases when interpreting the full case information is essential, and there is little room for implicit reflective processes.

Another plausible explanation for the difference in the effectiveness of reflection phases in simulation-based learning and problem-centered instruction is based on the theory of multimedia learning [12]. According to this theory, information processing differs during simulation-based learning and problem-centered instruction using text-based cases. The finding that reflection phases had no effect on learning to diagnose accurately in our study but are generally effective in problem-centered instruction can be explained according to this theory as follows. In simulation-based learning with virtual patients, the visual and the verbal channels are used simultaneously, and the largest benefit for learning may arise from integrating both channels [31]. Reflection phases might not support this integration process. In problem-centered instruction based on text cases, however, only the verbal channel is used. Reflection phases might particularly support the cognitive processes of selecting and organizing words that are important for creating an elaborate verbal representation [31].

Moreover, to complement our main research question, we examined the optimal timing of reflection phases. We initially assumed that accompanying reflection would outperform concluding reflection due to improved planning and monitoring of the diagnostic process [21]. Nevertheless, we also acknowledged that the concluding reflection condition might be associated with creating better self-generated feedback to be used in problem-solving [22]. However, the two reflection conditions had no effect on learning to diagnose accurately and did not differ from each other. Our findings suggest that in simulation-based learning, the two types of reflection phases do not differ in their effectiveness, and none of the described mechanisms is highly beneficial.

In the second research question (RQ2), we examined the associations between prior knowledge and learning to diagnose accurately through reflection phases. Neither conceptual nor strategic prior knowledge was correlated with improvements in diagnostic accuracy through reflection phases. This finding contradicts results from meta-analyses that learners with high prior knowledge benefit more from reflection than learners with low prior knowledge [1, 2]. However, there is a convincing explanation for this finding. In the described meta-analyses, knowledge was mainly operationalized as expertise determined by years of training. From an expertise development perspective, we investigated third to fifth year undergraduate medical students in our study, a cohort of learners with low to medium expertise. This cohort of learners was not able to learn through reflection phases in the context of virtual patients. This finding corresponds to an experiment by Mamede et al. [19], which showed that only postgraduate students and not undergraduate students, benefited from conscious, slow thinking when solving complex text-based cases. Together, our study and, even more convincingly, the experiment by Mamede et al. [19] indicate that reflection phases' effectiveness for learning to diagnose accurately might depend more on large differences in expertise than on smaller, context-specific differences in knowledge.

In the third research question (RQ3), we analyzed the extent to which participants' diagnostic process improves during simulation-based learning and reflection phases.

It is important to note that the improvements in the diagnostic process we reported probably depend to some extent on case difficulty. On the one hand, greater improvements during simulation-based learning and reflection phases are possible with more difficult cases. On the other hand, improvements are presumably impossible with overly difficult cases. The separately reported proportions of not changing one's hypothesis (specifying the proportion of fully correct, partially correct, and incorrect unchanged hypotheses), improvement, and deterioration suggest sufficient room for improvement during virtual patients and reflection phases.

The analysis of the simulation-based learning phase without reflection phases (the control group) demonstrated that participants improved their diagnoses from the start of each case to the end. A categorical follow-up analysis showed that a substantial number of participants improved. This improvement in the diagnostic process might, on the one hand, be explained by the step-wise gathering and interpretation of additional data while working with the virtual patients [8]. On the other hand, the expert sample solutions provided after participants gave their final diagnosis in each case might have also had a positive transfer effect on participants' diagnoses in the subsequent virtual patients.

The analysis of the diagnostic process during the reflection phases (both intervention groups) revealed that participants also improved their current hypotheses from the start of the reflection phase to the end. A categorical follow-up analysis showed that a smaller proportion of participants improved their current hypotheses during the reflection phases than while working with the virtual patients.

Together, these findings indicate that simulation-based learning with the virtual patients contributed more substantially to participants' improvements in the diagnostic process than reflection phases. Furthermore, improvements in the diagnostic process during the virtual patients and in the reflection phases we discovered in the learning phase did not transfer to an improved diagnostic accuracy in the posttest. There are two explanations we suggest for this finding. First, the reflection phases might not have been as effective as expected due to differences in case format and information processing (please see discussion for RQ1). Second, the expert solutions we included in all three experimental conditions during the learning phase could have affected posttest performance concerning diagnostic accuracy more strongly than the reflection phases [32]. More specifically, the expert solutions included strategic knowledge on the correct diagnostic process that may have contributed to reducing the differences between the control and reflection groups. However, providing feedback in the form of expert solutions is frequently considered a necessary part of simulation-based learning [33]. Therefore, we argue that it made sense to include expert solutions in all conditions.

To link our findings more closely to other research, we would like to briefly highlight similarities and differences between debriefing and the expert solutions and reflection phases used in our study. Debriefing can stimulate reflection processes and include solutions to the diagnostic process or performance [34]. In contrast to reflection phases and expert solutions, however, debriefing is more interactive and dialogic [34]. Thus, our findings cannot be generalized to debriefing, for which much further research seems necessary and valuable.

In conclusion, instructional support in the form of reflection phases had no added benefit for learning to diagnose accurately for undergraduate students with low to medium expertise in simulation-based learning with virtual patients. If our findings are replicated, this would suggest that other instructional supports might be more beneficial in this context and similar settings. Combinations of selective instructional support (such as examples and prompts) and adaptive instructional support could be promising alternatives to reflection phases, as both have been found to be beneficial in simulation-based learning and for learners with relatively little expertise [1, 4, 35, 36].

Limitations

One limitation of the study is that we switched data collection from the lab to the internet in the middle of the data collection period due to the COVID-19 pandemic. The drawback of web-based data collection is that it is considered less controlled than lab-based data collection [37]. However, this limitation should not be considered too severe in this study for two reasons. First, we conducted detailed manipulation checks that showed that the experiment was conducted as intended. Second, the proportions of web-based and lab-based data collection were similar in all conditions, as the Chi-squared test reported in the methods section showed.

Another limitation of the study could be the relatively low number of virtual patient cases we used. Other studies on reflection phases have used a larger number of text-based cases to assess diagnostic competences [6, 16]. The advantages of using a larger number of cases are that case specificity can be mitigated and reliability can be further increased [38, 39]. However, the benefits of using fewer virtual patient cases with a realistic duration, as we did in this study, are that more contextual information is conveyed and participants encounter a more interactive, realistic situation and task with higher validity [40].

A third limitation of the study could be the use of an immediate posttest. Even though positive effects of reflection on diagnostic accuracy have been reported on more immediate measures [20], most studies discovered positive effects on delayed posttests [5–7]. Therefore, it is possible that using a delayed posttest instead of an immediate posttest may have resulted in a positive effect of reflection phases on knowledge organization and retention, which were not assessed in the immediate posttest used.

Conclusions

We conducted a study on diagnosing in virtual patients with and without reflection phases. Our results showed that reflection phases did not have an added benefit on learning to diagnose accurately. This finding may be limited to the context of virtual patients and undergraduate medical students with low to medium expertise and needs replication. However, the results could have two important implications. First, reflection phases may not be as effective in simulation-based learning as in regular problem-centered instruction using text-based cases. This implication is substantiated by differences in case format and information processing between simulation-based learning and problem-centered instruction with text-based cases. Second, instructional supports other than reflection phases could be more beneficial for medical students with low to medium expertise in the context of simulation-based learning.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-021-02937-9.

Additional file 1: Appendix S1. Diagram of participant flow. **Appendix S2.** Participant characteristics. **Appendix S3.** Cases in the virtual patients and history-taking questions. **Appendix S4.** Reflection phases. **Appendix S5.** Manipulation checks.

Acknowledgements

We would like to thank Hannah Gerstenkorn and Victoria Reitmeier for their support in the development of the virtual patients. Also, we would like to thank Johannes Kissel who helped conduct the study and Keri Hartman for proofreading. The first author would like to thank Larissa Kaltefleiter for her advice.

Authors' contributions

MCF contributed to the conceptualization and design of the study, gathered and analyzed the data and created the first draft. NH contributed to the conceptualization and design of the study and revised the article. MS conceptualized and designed the study, acquired funding, and revised the article. FF conceptualized and designed the study, acquired funding, and revised the article. MRF conceptualized and designed the study, acquired funding, and revised the article. All authors approved the final manuscript for publication.

Funding

This study was funded by the German Research Association (Deutsche Forschungsgemeinschaft, DFG, Research Unit FOR2385 Cosima, Project 4). Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

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

Declarations

Ethics approval and consent to participate

The Ethics Committee of the Medical Faculty of LMU Munich granted ethical approval (No. 18–302) to the study. Participation in the study was voluntary. All participants gave informed consent. Data handling and data privacy protection regulations from the Ethics Committee of the LMU Munich Medical Faculty were followed. All procedures performed in the study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Consent for publication

Consent for publication was granted from the actor displayed in Fig. 1.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute of Medical Education, University Hospital, LMU Munich, Pettenkoferstr. 8a, 80336 Munich, Germany. [2]Institute of Education, Universität der Bundeswehr München, Neubiberg, Germany. [3]Department of Psychology, LMU Munich, Munich, Germany. [4]Munich Center of the Learning Sciences, LMU Munich, Munich, Germany.

Received: 10 May 2021 Accepted: 4 September 2021
Published online: 07 October 2021

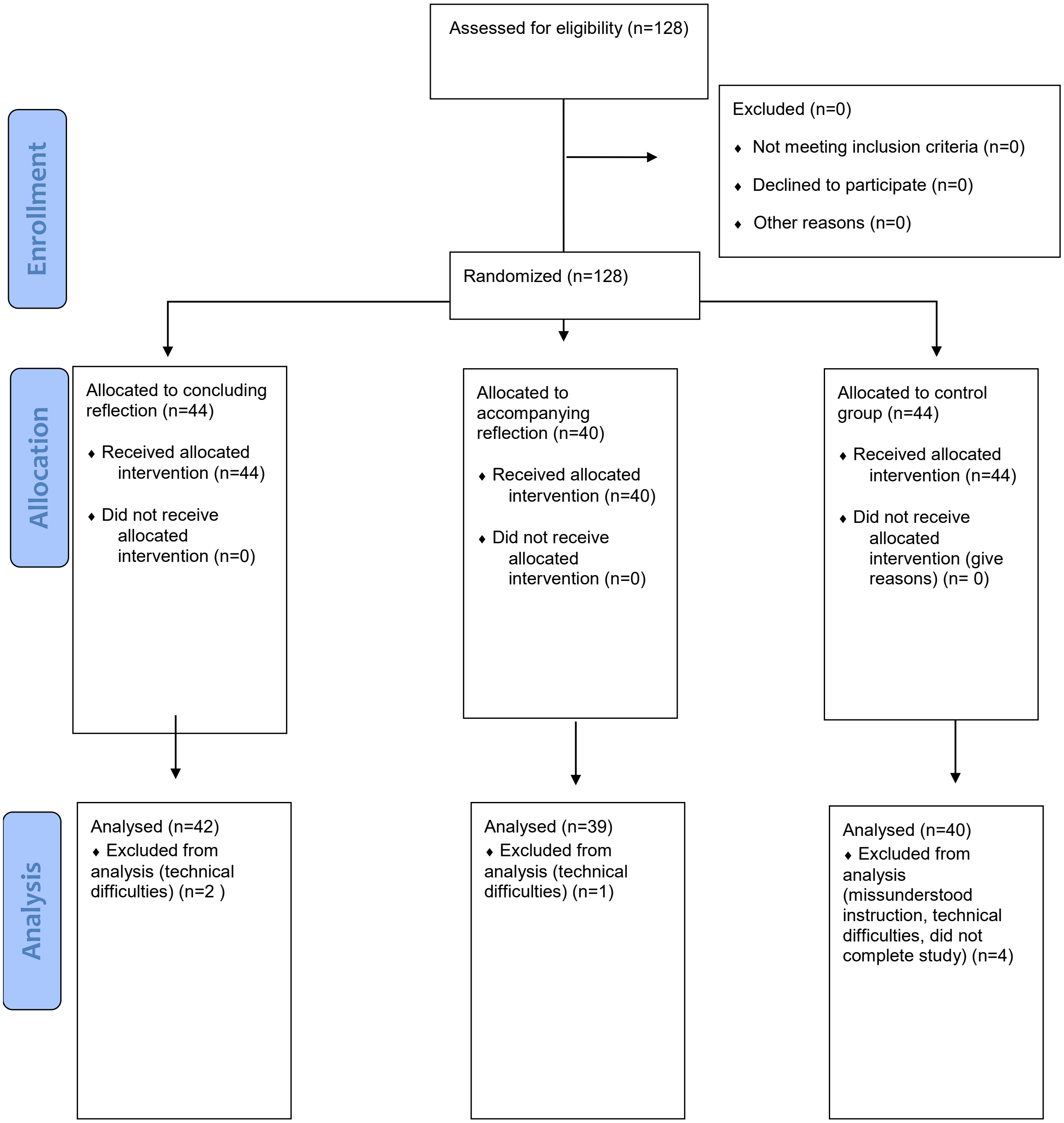
References

1. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Rev Educ Res. 2020;90:499–541. https://doi.org/10.3102/0034654320933544.
2. Chernikova O, Heitzmann N, Fink MC, Timothy V, Seidel T, Fischer F. Facilitating diagnostic competences in higher education: a meta-analysis in medical and teacher education. Educ Psychol Rev. 2019;68:157–96. https://doi.org/10.1007/s10648-019-09492-2.
3. Cook DA, Hamstra SJ, Brydges R, Zendejas B, Szostek JH, Wang AT, et al. Comparative effectiveness of instructional design features in simulation-based education: systematic review and meta-analysis. Med Teach. 2013;35: e867–98. https://doi.org/10.3109/0142159X.2012.714886.
4. Belland BR, Walker AE, Kim NJ, Lefler M. Synthesizing results from empirical research on computer-based scaffolding in STEM education: a meta-analysis. Rev Educ Res. 2017;87:309–44. https://doi.org/10.3102/0034654316670999.
5. Mamede S, van Gog T, Sampaio AM, Delbone de Faria RMD, Maria JP, Schmidt HG. How can students' diagnostic competence benefit most from practice with clinical cases? The effects of structured reflection on future diagnosis of the same and novel diseases. Acad Med. 2014;89:121–7. https://doi.org/10.1097/ACM.0000000000000076.
6. Ibiapina C, Mamede S, Moura A, Eloi-Santos S, van Gog T. Effects of free, cued and modelled reflection on medical students' diagnostic competence. Med Educ. 2014;48:796–805. https://doi.org/10.1111/medu.12435.
7. Mamede S, van Gog T, Moura AS, de Faria RMD, Peixoto JM, Rikers RMJP, Schmidt HG. Reflection as a strategy to foster medical students' acquisition of diagnostic competence. Med Educ. 2012;46:464–72. https://doi.org/10.1111/j.1365-2923.2012.04217.x .
8. Heitzmann N, Seidel T, Opitz A, Hetmanek A, Wecker C, Fischer MR, et al. Facilitating diagnostic competences in simulations in higher education: a framework and a research agenda. FLR. 2019;7:1–24. https://doi.org/10.14786/flr.v7i4.384.
9. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. Acad Med. 2010;85:1589–602. https://doi.org/10.1097/ACM.0b013e3181edfe13.
10. Hirumi A, Kleinsmith A, Johnsen K, Kubovec S, Eakins M, Bogert K, et al. Advancing virtual patient simulations through design research and interPLAY: part I: design and development. Educ Technol Res Dev. 2016;64: 763–85. https://doi.org/10.1007/s11423-016-9429-6.
11. Huwendiek S, de Leng BA, Zary N, Fischer MR, Ruiz JG, Ellaway R. Towards a typology of virtual patients. Med Teach. 2009;31:743–8. https://doi.org/10.1080/01421590903124708.
12. Mayer RE, Moreno R. A cognitive theory of multimedia learning: implications for design principles. J Educ Psychol. 1998;91:358–68.
13. Low R, Sweller J. The modality principle in multimedia learning. In: Mayer RE, editor. The Cambridge handbook of multimedia learning. Cambridge: Cambridge University Press; 2014. p. 227–46. https://doi.org/10.1017/CBO9781139547369.012.
14. Kiesewetter J, Sailer M, Jung VM, Schönberger R, Bauer E, Zottmann JM, et al. Learning clinical reasoning: how virtual patient case format and prior knowledge interact. BMC Med Educ. 2020;20:73. https://doi.org/10.1186/s12909-020-1987-y.
15. Nguyen QD, Fernandez N, Karsenti T, Charlin B. What is reflection? A conceptual analysis of major definitions and a proposal of a five-component model. Med Educ. 2014;48:1176–89.https://doi.org/10.1111/medu.12583.
16. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Med Educ. 2008;42:468–75. https://doi.org/10.1111/j.1365-2923.2008.03030.x.
17. Kahneman D. Thinking, fast and slow. 1st ed. New York: Macmillan; 2011.
18. Mamede S, Schmidt HG. Reflection in medical diagnosis: a literature review. Health Prof Educ. 2017;3:15–25. https://doi.org/10.1016/j.hpe.2017.01.003.
19. Mamede S, Schmidt HG, Rikers RMJP, Custers EJFM, Splinter TAW, van Saase JLCM. Conscious thought beats deliberation without attention in diagnostic decision-making: at least when you are an expert. Psychol Res. 2010;74:586–92. https://doi.org/10.1007/s00426-010-0281-8.
20. Mamede S, Hautz WE, Berendonk C, Hautz SC, Sauter TC, Rotgans J, et al. Think twice: effects on diagnostic accuracy of returning to the case to reflect upon the initial diagnosis. Acad Med. 2020;95:1223–9. https://doi.org/10.1097/ACM.0000000000003153.
21. Zimmerman BJ. Becoming a self-regulated learner: which are the key subprocesses? Contemp Educ Psychol. 1986;11:307–13. https://doi.org/10.1016/0361-476X(86)90027-5.
22. Butler DL, Winne PH. Feedback and self-regulated learning: a theoretical synthesis. Rev Educ Res. 1995;65:245–81.
23. Fink MC, Reitmeier V, Stadler M, Siebeck M, Fischer F, Fischer MR. Assessment of diagnostic competences with standardized patients versus virtual patients: experimental study in the context of history taking. J Med Internet Res. 2021;23:e21196. https://doi.org/10.2196/21196.
24. Instruct. CASUS. 2021. https://www.instruct.eu/. Accessed 8 May 2021.
25. Bauer D, Holzer M, Kopp V, Fischer MR. Pick-N multiple choice-exams: a comparison of scoring algorithms. Adv Health Sci Educ Theory Pract. 2011; 16:211–21. https://doi.org/10.1007/s10459-010-9256-1.
26. Hrynchak P, Glover Takahashi S, Nayer M. Key-feature questions for assessment of clinical reasoning: a literature review. Med Educ. 2014;48:870–83. https://doi.org/10.1111/medu.12509.
27. Cook DA, Brydges R, Hamstra SJ, Zendejas B, Szostek JH, Wang AT, et al. Comparative effectiveness of technology-enhanced simulation versus other instructional methods: a systematic review and meta-analysis. Simul Healthc. 2012;7:308–20. https://doi.org/10.1097/SIH.0b013e3182614f95.
28. Opfermann M. There's more to it than instructional design: the role of individual learner characteristics for hypermedia learning. Berlin: Logos; 2008.
29. R Core Team. R: a language and environment for statistical computing. 2021. https://www.R-project.org/. Accessed 10 May 2021.
30. Faul F, Buchner A, Erdfelder E, Lang AG. G*Power. 2014. https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html. Accessed 9 May 2021.
31. Mayer RE. Cognitive theory of multimedia learning. In: Mayer RE, editor. The Cambridge handbook of multimedia learning. Cambridge: Cambridge University Press; 2014. p. 43–71. https://doi.org/10.1017/CBO9781139547369.005.
32. Hattie J, Timperley H. The power of feedback. Rev Educ Res. 2007;77:81–112. https://doi.org/10.3102/003465430298487.
33. 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;27:10–28. https://doi.org/10.1080/01421590500046924.
34. Cheng A, Eppich W, Grant V, Sherbino J, Zendejas B, Cook DA. Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. Med Educ. 2014;48:657–66. https://doi.org/10.1111/medu.12432.
35. Steenbergen-Hu S, Cooper H. A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. J Educ Psychol. 2014;106:331–47. https://doi.org/10.1037/a0034752.
36. van Lehn K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychologist. 2011;46: 197–221.
37. Reips UD. The web experiment method: advantages, disadvantages, and solutions. In: Birnbaum MH, editor. Psychological experiments on the internet. San Diego: Academic Press; 2000. https://doi.org/10.5167/uzh-19760.
38. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: an analysis of clinical reasoning. Cambridge: Harvard University Press; 1978.
39. Swanson DB, Roberts TE. Trends in national licensing examinations in medicine. Med Educ. 2016;50:101–14. https://doi.org/10.1111/medu.12810.
40. van der Vleuten CPM, Schuwirth LWT. Assessment in the context of problem-based learning. Adv Health Sci Educ Theory Pract. 2019;24:903–14. https://doi.org/10.1007/s10459-019-09909-1.

Publisher's Note

APPENDICES

Appendix S1: Diagram of participant flow


Enrollment
Assessed for eligibility (n=128)
Excluded (n=0)
♦ Not meeting inclusion criteria (n=0)
♦ Declined to participate (n=0)
♦ Other reasons (n=0)
Randomized (n=128)
Allocation
Allocated to concluding reflection (n=44)
♦ Received allocated intervention (n=44)
♦ Did not receive allocated intervention (n=0)
Allocated to accompanying reflection (n=40)
♦ Received allocated intervention (n=40)
♦ Did not receive allocated intervention (n=0)
Allocated to control group (n=44)
♦ Received allocated intervention (n=44)
♦ Did not receive allocated intervention (give reasons) (n= 0)
Analysis
Analysed (n=42)
♦ Excluded from analysis (technical difficulties) (n=2)
Analysed (n=39)
♦ Excluded from analysis (technical difficulties) (n=1)
Analysed (n=40)
♦ Excluded from analysis (missunderstood instruction, technical difficulties, did not complete study) (n=4)

Appendix S2: Participant characteristics

Participant characteristics across all conditions

	Concluding reflection	Accompanying reflection	Control group	All groups
Age in years mean, (SD)	25.12 (3.76)	25.29 (3.90)	24.30 (4.39)	24.90 (4.01)
Participants	42	39	40	121
Sex, n (%)				
Females	30 (72)	28 (72)	24 (60)	82 (68)
Males	1 (2)	3 (8)	6 (15)	10 (8)
No answer	11 (26)	8 (20)	10 (25)	29 (24)
Data collection, n (%)				
Lab-based	12 (29)	11 (28)	11 (27)	34 (28)
Web-based	30 (71)	28 (72)	29 (73)	87 (72)

Appendix S3: Cases in the virtual patients and history-taking questions

Cases in the virtual patients

Phase of the experiment	Diagnosis	Patient characteristics
Pretest	Hypertrophic cardiomyopathy	25 years, male
	Pneumonia	55 years, female
	Pulmonary embolism in case of prostrate cancer	70 years, male
Learning phase	Acute posterior myocardial infarction	55 years, female
	Pulmonary embolism due to heparin induced thrombocytopenia	70 years, male
	Lung cancer	60 years, female
Posttest	Pulmonary embolism due to coagulation disorder	35 years, female
	Congestive heart failure with atrial fibrillation	65 years, female
	Hyperventilation tetany	45 years, male

The questions provided in the menu (history-taking questions) came from the categories *main symptoms*, *prior history*, *allergies and medication*, *social and family history*, and *system review* by Bornemann (2016) and were evaluated in Fink et al (2021).

History-taking Questions

Category	Code	Example
Main symptoms	MS	Do you experience the complaints for the first time?
Prior history	PH	Do you know of any pre-existing conditions?
Allergies and medication	AM	Do you frequently have infections against which you take antibiotics?
Social and family history	SF	Have your parents or other relatives of your family passed away at a rather young age?
System review	SR	Has your weight changed within the last weeks?

Bornemann, B. (2016). *Dokumentationsbögen der inneren Medizin und der Chirurgie für Anamnese und körperliche Untersuchung für die studentische Lehre in Deutschland* (Diss., Institut für Didaktik und Ausbildungsforschung in der Medizin der Ludwig-Maximilians-Universität München). Retrieved from https://edoc.ub.uni-muenchen.de/19166/

Fink, M. C., Reitmeier, V., Stadler, M., Siebeck, M., Fischer, F., Fischer, M. R., . . . Fischer, M. R. (2021). Assessment of diagnostic competences with standardized patients versus virtual patients: Experimental study in the context of history taking. *Journal of Medical Internet Research*, *23*(3), e21196. https://doi.org/10.2196/21196

Appendix S4: Reflection phases

Accompanying reflection

Nr	Question
1.	Please name your current diagnosis.
2.	What symptoms and findings agree with your current diagnosis?
3.	What symptoms and findings disagree with your current diagnosis?
4.	What questions will you ask the patient to examine your current diagnosis?
5.	Please name alternative diagnosis in case your current diagnosis is incorrect.
6.	What symptoms and findings agree with your alternative diagnoses?
7.	What symptoms and findings disagree with your alternative diagnoses?
8.	What questions will you ask the patient to examine your alternative diagnoses?
9.	After reflecting on your current diagnosis and alternative diagnoses: What do you consider now the most probable diagnosis?

Concluding reflection

Nr	Question
1.	Please name your current diagnosis.
2.	What symptoms and findings support your current diagnosis?
3.	What symptoms and findings disagree with your current diagnosis?
4.	What questions you asked the patient were important to examine your current diagnosis?
5.	Please name alternative diagnosis in case your current diagnosis is incorrect.
6.	What symptoms and findings agree with your alternative diagnoses?
7.	What symptoms and findings disagree with your alternative diagnoses?
8.	What questions you asked the patient were important to examine your alternative diagnoses?
9.	After reflecting on your current diagnosis and alternative diagnoses: What do you consider now the most probable diagnosis?

Further information
Question 1, 5, and 9 used a long-menu format (see the section on diagnostic accuracy), all other questions free text input.

Appendix S5: Manipulation checks

Table 1 reports the duration participants spent working on the cases in different phases of the experiment and statistics from a one-way ANOVA comparing the groups. As expected, the duration of participants did not differ across groups in the pretest and posttest. As intended, participants in the experimental groups spent about four additional minutes on reflection.

Table 1

Duration Spent in the Conditions

Phase	Concluding reflection	Accompanying reflection	Control group	*df*	*F*	*p*
Pretest	10.11 (1.81)	10.27 (1.74)	10.15 (1.56)	2, 118	0.10	.909
Learning Phase	13.71 (2.82)	14.25 (2.13)	8.97 (1.79)	2, 117	64.01	<.001
Posttest	8.71 (2.41)	9.32 (2.07)	9.31 (1.73)	2, 118	1.14	.322

Note. Means and (SDs) of the duration that participants spent in the conditions in minutes. One-way ANOVA results are reported across the three experimental conditions.

For another manipulation check, the number of words during reflection phases was calculated. Participants wrote in the accompanying reflection group on average $M = 22.94$, $SD = 12.06$ words per case during their reflections. Participants in the concluding reflection group wrote on average $M = 30.11$, $SD = 21.04$ words per case. There are two reasons that explain these relatively low word counts. First, participants summarized their reflections, as log-data inspection showed, primarily in bullet points. Second, participants spent during reflection only about half the time on free text questions and the other half on long-menu questions that did not contribute to the reported word count.

5. General Discussion

I begin this chapter with a summary of the results of the enclosed three empirical articles.[2] Afterwards, I discuss the three central research questions making up this dissertation: 1) What differences and similarities emerge in the assessment of diagnostic competences with the two assessment methods of standardized patients and virtual patients? 2) How are process variables related to diagnostic quality in simulation-based assessment? and 3) To what extent can the simulation-based learning of diagnostic competences be facilitated with scaffolding? I address the first two central research questions in the section on *implications for assessing diagnostic competences with simulations* (5.2). I elaborate on the third central research question in the section on *implications for facilitating diagnostic competences with simulations* (5.3). Finally, I discuss the limitations of the conducted research, provide some directions for future research, and offer a conclusion.

5.1 Summary of the Results of the Articles

Article 1 (Fink, Reitmeier et al., 2021) contrasted the assessment of diagnostic competences with standardized patients and virtual patients in a sample of $N = 86$ medical students. In a repeated-measures design, every participant engaged with both assessment methods. The article compared to what extent key variables associated with diagnostic competences (i.e., perceived authenticity, cognitive load, and diagnostic accuracy) differ or are equivalent in standardized patients and virtual patients. Moreover, the article examined to what extent these key variables are associated with diagnostic accuracy. With respect to perceived authenticity, all three facets were higher in standardized patients than in virtual patients but only displayed small correlations with diagnostic accuracy. Regarding cognitive load, all three facets were equivalent in standardized patients and virtual patients. Moreover, all facets of cognitive load correlated negatively with diagnostic accuracy. With respect to diagnostic competences, we assessed the two variables quality of evidence generation and diagnostic accuracy. The quality of evidence generation and diagnostic accuracy were positively associated in virtual patients only, not in standardized patients. Diagnostic accuracy was substantially higher in standardized patients than

[2] An overview of the empirical articles, including their specific research questions, was provided in Section 1.6. The three empirical articles are enclosed in Chapters 2 to 4.

in virtual patients, indicating that the use of standardized patients could positively affect examinees' scores.

Article 2 (Fink et al., submitted) focused on determining to what extent the diagnostic process and knowledge contribute to diagnostic quality. N = 106 medical students completed a conceptual and strategic professional knowledge test before diagnosing multiple virtual patients. The diagnostic activities of hypothesis generation, evidence generation, and evidence evaluation served as measures of the diagnostic process. Diagnostic quality was measured with a separate diagnostic accuracy score and a comprehensive diagnostic score. The comprehensive diagnostic score encompassed the four components of diagnostic accuracy, treatment selected, diagnostic measures for clarification, and expected findings in a physical examination. Professional knowledge predicted performance in the comprehensive diagnostic score and was associated with diagnostic accuracy in bivariate correlations. The diagnostic activities explained a medium amount of variance in the comprehensive diagnostic score and the diagnostic accuracy score. Furthermore, it was examined to what extent diagnostic activities add to the predictive value of professional knowledge. For both diagnostic quality scores as criterions, diagnostic activities added a substantial amount of explained variance to the effects of professional knowledge.

Article 3 (Fink, Heitzmann, Siebeck, Fischer, & Fischer, 2021) investigated the effect of reflection phases on learning to diagnose accurately through virtual patients. Moreover, it examined relationships between prior knowledge and learning, and improvements in the diagnostic process in reflection phases and simulation-based learning. In the pretest, N = 121 medical students filled out a conceptual and strategic prior knowledge test and then diagnosed virtual patients. In the intervention, a control group learned from virtual patients only while two experimental groups additionally completed different types of reflection phases. In the posttest, participants in all conditions diagnosed virtual patients. For all virtual patients, diagnostic accuracy served as the primary outcome. To measure the diagnostic process, participants' current hypothesis was tracked in the virtual patients and reflection phases. The results demonstrated that reflection phases did not have an added benefit for learning to diagnose accurately from pre- to post-test. Analyses of the relationship between prior knowledge and learning revealed that neither conceptual nor strategic knowledge was correlated with improvement in diagnostic accuracy. Regarding the diagnostic process in the learning phase, participants improved their hypotheses both during reflection phases and while working with virtual patients.

5.2 Implications for Assessing Diagnostic Competences with Simulations

This section focuses on the assessment of diagnostic competences with simulations. Therefore, the results from Article 1 and Article 2, in which diagnostic competences were assessed without including scaffolding, are discussed. Concurrently, theoretical and practical implications are provided. The chapter is further divided into two sections that contribute to this dissertation's first and second central research questions, respectively.

5.2.1 Differences between simulation modalities in the assessment of diagnostic competences

This dissertation aimed to enhance our understanding of the differences between simulation modalities in assessing diagnostic competences. First of all, it should be acknowledged that numerous simulation modalities exist (see Section 1.3). Because of their growing popularity and high relevance for the assessment of diagnostic competences, this dissertation focused on the two assessment methods of standardized patients and virtual patients, which in terms of modality correspond to live simulations and digital simulations, respectively (see Section 1.4). My contribution to comparing standardized patients and virtual patients centers on diagnostic accuracy, perceived authenticity, cognitive load, and further process variables noted as particularly important in the literature (see Sections 1.3 and 1.4).

Concerning diagnostic accuracy, there are only a few studies in medical education that have directly compared standardized patients and virtual patients (Edelstein et al., 2000; Guagnano et al., 2002; Hawkins et al., 2004). These studies reported correlations between diagnostic accuracy in standardized patients and virtual patients, thus indicating a correspondence between the two assessment methods. Nevertheless, the specific design features and characteristics included in standardized patients and virtual patients (see Section 1.3 and Section 1.4) could create a modality effect and be associated with performance differences. Article 1 (Fink, Reitmeier et al., 2021) adds to this literature by finding that diagnostic accuracy was higher for standardized patients than for virtual patients. This finding provides evidence for a modality effect and shows that the correspondence between virtual patients and standardized patients is lower than expected. Indeed, an additional equivalence test revealed that the obtained grades would not have been equivalent, and students would have received better fictitious grades in an assessment with standardized patients compared to virtual patients. Moreover, I expanded the literature by identifying two theoretical explanations for the performance differences encountered between the two assessment methods. On the one hand, assessment with standardized patients could have overestimated participants' true performance. A possible reason for this might be that the actors playing the standardized patients provided additional support. On the other hand, assessment with virtual patients could have underestimated the participants' true performance. This explanation was substantiated by the finding that participants displayed a lower quantity of evidence generation when working with the virtual patients than in the live simulations. Most likely, selecting history-taking questions from a menu for virtual patients took more time than formulating questions orally for standardized patients. I believe that both theoretical explanations can also shed light on performance differences in future comparisons of standardized patients and virtual patients and other simulation modalities.

Another important variable in the assessment of diagnostic competences is the perceived authenticity of the simulation modalities. The literature reports that standardized patients are among the most authentic types of simulations (Luctkar-Flude et al., 2012; Rethans et al., 1991). However, high values on perceived authenticity have also been reported for virtual patients (Friedman et al., 1991). Article 1 (Fink, Reitmeier et al., 2021) expands upon the literature by showing that standardized patients are given higher ratings on the three perceived authenticity facets of realness, presence, and spatial presence than virtual patients. The largest difference in perceived authenticity between standardized patients and virtual patients was encountered in the facet spatial presence. This finding is not surprising and indicates that virtual patients cannot provide the same sense of a physical environment that standardized patients can offer. With the rise of new simulation technologies, it seems likely that some of the differences in perceived authenticity between standardized patients and virtual patients encountered in this dissertation may fade away in the future. For instance, it has been shown that virtual reality simulations provide a higher sense of perceived authenticity than their regular digital counterparts (Makransky, Terkildsen, & Mayer, 2019). Therefore, differences in spatial presence between virtual patients incorporating virtual reality and standardized patients could be smaller than between the regular virtual patients and standardized patients investigated in this dissertation.

Cognitive load (Sweller et al., 1998) is another important variable that should be compared across simulation modalities because of its strong link to instructional design. Currently, however, there are only a few studies examining cognitive load in different simulation modalities. One study showed that cognitive load varied between simulation types with differing levels of task complexity (Haji et al., 2016). Another study indicated that the authenticity of different simulation types affected cognitive load (Dankbaar et al., 2016). Article 1 (Fink, Reitmeier et al., 2021) demonstrated that the three facets of cognitive load, namely intrinsic load, extraneous load, and germane load, were equivalent in standardized patients and virtual patients. Task complexity was highly comparable across simulations due to use of the same clinical cases, while perceived authenticity differed. Therefore, the results corroborate the argument that cognitive load is primarily determined by task complexity rather than by perceived authenticity. Moreover, the results underscore that cognitive load can reach a similar level in digital and live simulations. These findings could be relevant for instructional designers seeking to replicate the level of cognitive load inherent in a real-world situation or task in a simulation, or to align cognitive load levels across different simulation modalities.

Furthermore, the diagnostic process variable of evidence generation was examined. Previous direct comparisons of standardized patients and virtual patients had not investigated this variable (Edelstein et al., 2000; Guagnano et al., 2002; Hawkins et al., 2004). Article 1 (Fink, Reitmeier et al., 2021) discovered that the quantity of evidence generation was considerably higher in standardized patients than in virtual patients, but the quality of evidence generation was higher in virtual patients than in standardized patients. In standardized patients, participants formulated questions orally without guidance, leading to a higher number of questions being uttered that were not necessarily relevant to the current case. In contrast, in virtual patients, participants selected questions from a menu, which may have provided some scaffolding but also meant that a lower number of questions could be posed in the same available amount of time. The varying user input across the two assessment methods may have resulted in the reported differences in the quantity and quality of evidence generation. These findings and explanations highlight that assessors should always consider user input when designing simulations for assessment.

The comparison of standardized patients and virtual patients revealed differences in diagnostic accuracy, evidence generation, and perceived authenticity between the two assessment methods. These results illustrate that, particularly in summative assessment, the assessment methods should not be substituted one-for-one for each other but be adapted sensibly to the specific context. However, it is an open question to what extent these differences encountered for standardized patients and virtual patients generalize to other simulation modalities apart from live simulations and digital simulations.

5.2.2 Relationships between process variables and diagnostic quality in simulation-based assessment

Another central research question for this dissertation was how process variables are associated with diagnostic quality in simulation-based assessment. This analysis is important because the diagnostic process and other process variables have largely been neglected in simulation-based education so far (Heitzmann et al., 2017). As described in Section 1.2, there are many different conceptualizations of the diagnostic process. The present work builds upon the variables described in the conceptual framework by Heitzmann et al. (2019). Therefore, the variables that I discuss include diagnostic activities, professional knowledge, perceived authenticity, and cognitive load.

Concerning diagnostic activities, a number of empirical studies have reported correlations between diagnostic quality and hypothesis generation (Coderre et al., 2010; LeBlanc et al., 2001; LeBlanc et al., 2002) as well as evidence generation (Stillman et al., 1986; Woolliscroft et al., 1989). Moreover, there was theoretical support for an association between diagnostic quality and evidence evaluation stemming from the script concordance literature (Charlin et al., 2000). Studies from teacher education and medical education examining multiple diagnostic processes in tandem also indicated that the diagnostic process explained a substantial amount of variance in diagnostic quality (Groves et al., 2003; Kramer et al., 2021). In line with the summarized literature, Article 2 (Fink et al., submitted) showed that hypothesis generation and evidence generation were predictors of the comprehensive diagnostic score and the diagnostic accuracy score, explaining medium amounts of variance. This finding underscores that these diagnostic activities are related to diagnostic quality (Heitzmann et al., 2019) and aligns with prior results from teacher education and medical education (Groves et al., 2003; Kramer et al., 2021). Evidence evaluation was only a significant predictor of the comprehensive diagnostic score but not of diagnostic accuracy. The latter result is opposed to the assumption of the script concordance literature (Charlin et al., 2000) that the interpretation of various pieces of information is associated with diagnostic quality and may perhaps be explained by peculiarities of the instrument used (see Article 2 for more details). Article 1 (Fink, Reitmeier et al., 2021) provided mixed findings on the relationship between evidence generation and diagnostic accuracy. In this article, evidence generation correlated positively with diagnostic accuracy in virtual patients. In standardized patients, however, the correlation between the two variables was close to zero. The finding for virtual patients is in line with empirical studies and the assumption that diagnostic activities are related to diagnostic accuracy (Heitzmann et al., 2019; Stillman et al., 1986; Woolliscroft et al., 1989). The unexpected finding for standardized patients could have been caused by the lower total quality of evidence generation in this condition and unintended additional support provided by the actors to struggling participants. In sum, my findings concur with the literature that diagnostic activities are positively related to diagnostic quality. I extend the literature primarily with the finding that some of the diagnostic activities considered particularly relevant for history-taking explain a medium amount of variance in diagnostic quality in this context. One important implication of these findings is that diagnostic process measures should be included more frequently in simulation-based assessments in the future. On the one hand, educational theory has argued that the inclusion of diagnostic process measures allows for a broader assessment of diagnostic competences than purely focusing on knowledge prerequisites or performance in tasks (Blömeke et al., 2015). On the other hand, both my own and other research have shown that diagnostic process measures are indeed substantially related to diagnostic quality without being the same. Incorporating measures of the diagnostic process into assessment is also warranted because several developments may increase the importance of the diagnostic process in the future. Collaborative diagnosing in teams (Kiesewetter, Fischer, & Fischer, 2017) and the use of digital decision support systems (Castaneda et al., 2015), are two examples of recent developments that rely firmly on a correct and well-documented diagnostic process.

Another variable of interest is professional knowledge. As pointed out before, different frameworks for diagnostic competences, such as illness script theory (Schmidt et al., 1990), highlight the importance of knowledge for diagnosing. Moreover, the strong contribution of knowledge to diagnosing has been demonstrated in empirical studies (Schmidt & Rikers, 2007). While professional knowledge is not a process variable by itself, it is theorized to be employed in diagnostic activities (Heitzmann et al., 2019). Consequently, it was an open question to what extent professional knowledge and diagnostic activities each make a unique contribution to diagnostic quality. Article 2 (Fink et al., submitted) revealed that both professional knowledge and diagnostic activities each predict a medium amount of variance in a comprehensive diagnostic score. The result that professional knowledge explains variance in diagnostic quality concurs with many other studies that stress the importance of knowledge in diagnosing (Schmidt & Rikers, 2007). The result that diagnostic activities also make a unique contribution to diagnostic quality has important implications for frameworks for diagnostic competences. Several frameworks for diagnostic competences argue, based on the broad notion of competences by Blömeke et al. (2015), that diagnostic competences consist of knowledge, the diagnostic process, and diagnostic quality (Heitzmann et al., 2019; Herppich et al., 2018; Loibl et al., 2020). This conceptualization entails that the diagnostic process is more than a mere embodiment of knowledge and makes a unique contribution to diagnostic quality. Article 2 (Fink et al., submitted) is among the first studies to provide evidence for this assumption, specifically in terms of the conceptualization of diagnostic activities and professional knowledge put forward by Heitzmann et al. (2019). Because this research also highlights that professional knowledge plays a vital role in diagnosing in the context of simulations, formative and summative assessments with simulations should capture this variable in the future. In this case, formative and summative assessment could then, for instance, pinpoint whether performance deficits in simulations are caused by a lack of knowledge or deficits in knowledge application and inform participants about these issues.

Next, I will discuss the relationship between perceived authenticity and diagnostic accuracy. The medical education literature assumes that perceived authenticity is only minimally related to diagnostic quality (Norman et al., 2012). This is opposed to meta-analytic findings from various domains showing that greater simulation authenticity is associated with improved learning (Chernikova, Heitzmann, Fink et al., 2020; Chernikova, Heitzmann, Stadler et al., 2020). Article 1 (Fink, Reitmeier et al., 2021) reported very small and non-significant correlations between perceived authenticity variables and diagnostic accuracy. These findings are in line with results that raising the perceived authenticity above a certain threshold is not associated with considerable performance gains (Norman et al., 2012). A plausible explanation for these results might be the operationalization of authenticity in the present study. In Article 1 (Fink, Reitmeier et al., 2021), authenticity was operationalized according to the concept of perceived authenticity (Schubert et al., 2001). In the meta-analysis by Chernikova, Heitzmann, Stadler et al. (2020), simulation authenticity was determined not by participants' judgments but by coding aspects similar to the concept of functional fidelity (Hamstra et al., 2014; Maran & Glavin, 2003; see Section 1.3). Consequently, differences in the operationalization of authenticity might explain the discrepant reported relationships with diagnostic accuracy. This point also highlights that future research should examine the relationships between diagnostic accuracy and functional fidelity more closely. Moreover, based on my analyses and the review by Norman et al. (2012), I can advise practitioners to not spend a large share of their budget on increasing perceived authenticity in simulation-based assessment beyond a necessary, beneficial threshold.

Regarding cognitive load, several studies in educational psychology and medical education have shown that extraneous and intrinsic cognitive load are negatively related to performance in problem-solving and diagnosing (Sweller et al., 2019; Young et al., 2014). In Article 1 (Fink, Reitmeier et al., 2021), intrinsic and extraneous cognitive load were both negatively associated with diagnostic accuracy. These findings align with the reported results from educational psychology and medical education, but also highlight that cognitive load can be detrimental in assessment settings. Consequently, instructors and instructional designers should monitor and control cognitive load in formative and summative assessments. In formative assessment, the instructor could be informed of participants' current cognitive load and then adapt their instructional strategies accordingly. In summative assessment, cognitive load should be taken into account in instructional design and be measured during pilot evaluations. In this way, optimal testing conditions can be created for assessees.

In a nutshell, the diagnostic process, as operationalized by diagnostic activities, makes a unique and important contribution to diagnostic quality and is more than an embodiment of professional knowledge. This finding adds to the support that diagnostic competences, as stated by multiple frameworks, encompass knowledge, the diagnostic process, and diagnostic quality (Heitzmann et al., 2019; Herppich et al., 2018; Loibl et al., 2020). Concerning authenticity, practitioners should not spend a large share of their budget on increasing perceived authenticity above a necessary, beneficial level. Moreover, cognitive load should be monitored and controlled in formative and summative assessments.

5.3 Implications for Facilitating Diagnostic Competences with Simulations

This dissertation's third central research question was to what extent the simulation-based learning of diagnostic competences can be facilitated with scaffolding. A recent meta-analysis showed that simulation-based learning has a large effect on acquiring complex skills (Chernikova, Heitzmann, Stadler et al., 2020). Moreover, this effect could be increased when particular combinations and types of scaffolding well-matched to learners' prerequisites were added (Chernikova, Heitzmann, Stadler et al., 2020). While numerous types of scaffolding can be used in simulation-based learning (see Section 1.5), this dissertation focused on reflection phases. More specifically, it investigated the effectiveness of reflection phases in simulation-based learning, the relationship between prior knowledge and learning to diagnose accurately through virtual patients and reflection phases, and the diagnostic process during reflection phases. These research questions and their implications will be discussed before providing a conclusion on the central research question.

Regarding the effectiveness of reflection phases, findings in the literature are mixed. A literature review in medical education focused mainly on learning from text-based cases and reported largely positive effects (Mamede & Schmidt, 2017). Moreover, a meta-analysis on problem-based learning uncovered a positive effect of including reflection phases on the acquisition of diagnostic competences (Chernikova, Heitzmann, Fink et al., 2020). The studies included in this meta-analysis primarily used text-based cases in the domains of medical education and teacher education. However, the meta-analysis by Chernikova, Heitzmann, Stadler et al. (2020) found no additional benefit of including reflection phases on the acquisition of complex skills in simulation-based learning in various domains. In Article 3 (Fink, Heitzmann et al., 2021), reflection phases did not have an added benefit for learning to diagnose accurately through virtual patients. This finding contradicts the positive effects of reflection phases reported for learning from text-based cases (Chernikova, Heitzmann, Fink et al., 2020; Mamede & Schmidt, 2017) but is in line with the results on acquiring complex skills in simulation-based learning (Chernikova, Heitzmann, Stadler et al., 2020). One explanation for this finding could be that the large effect of simulation-based learning may have diluted the effectiveness of including this type of scaffolding. However, other types of scaffolding and certain combinations of scaffolding have displayed added benefits in simulation-based learning (Chernikova, Heitzmann, Stadler et al., 2020). Another explanation could be that reflection phases are not as effective in simulation-based learning as in text-based learning. On the one hand, the case format deviates between these two types of learning (Huwendiek et al., 2009; Kiesewetter et al., 2020). In simulation-based learning, serial cue cases are mostly used, in which learners dynamically construct a case and gather information in a step-by-step manner. In text-based learning, whole cases are typically used, in which learners remember and interpret information that is provided in full. On the other hand, according to the cognitive theory of multimedia learning, information processing differs between simulation-based learning and text-based learning (Mayer & Moreno, 1998; see also Article 3, Fink, Heitzmann et al., 2021). In line with this theory, reflection phases might be highly beneficial in text-based learning because they support learners in selecting and organizing content from the verbal channel. However, reflection phases could be less beneficial in simulation-based learning, in which integrating information from the visual and verbal channels is of prime importance for learning. If the finding that reflection phases have no added benefit for learning to diagnose accurately in simulations replicates, other scaffolding types should be used to support learners. In this specific context, adaptive scaffolding, combinations of different types of scaffolding, and support that takes into account the learner's expertise could be particularly effective (Chernikova, Heitzmann, Fink et al., 2020; Chernikova, Heitzmann, Stadler et al., 2020; Plass & Pawar, 2020b).

Next, I will discuss the relationship between prior knowledge and improvement in diagnostic accuracy in learning from reflection phases. There is meta-analytic evidence that reflection phases are more beneficial for acquiring diagnostic competences among learners with high prior knowledge than learners with low prior knowledge (Chernikova, Heitzmann, Fink et al., 2020; Chernikova, Heitzmann, Stadler et al., 2020). In these meta-analyses, prior knowledge was operationalized as the level of educational training and content familiarity. A study by Mamede et al. (2010) showed that only physicians in specialist training, but not medical students, benefitted from reflective thought in solving complex problems. According to the authors, this might indicate that only the physicians and not medical students had the necessary expertise to improve through reflective thought (Mamede et al., 2010). Article 3 (Fink, Heitzmann et al., 2021) reported a low and not statistically significant correlation between prior strategic and conceptual knowledge and the improvement in diagnostic accuracy in the reflection phase conditions. This finding does not support the association between prior knowledge and acquiring diagnostic competences through reflection phases indicated by meta-analytic findings (Chernikova, Heitzmann, Fink et al., 2020; Chernikova, Heitzmann, Stadler et al., 2020). One explanation for this finding could be that the participants in the underlying study had, from an expertise development perspective, low to medium expertise and thus did not benefit from reflective thought, similar to the medical students in Mamede et al. (2010). Consequently, Article 3 (Fink, Heitzmann et al., 2021) and Mamede et al. (2010) together indicate that reflection phases' effectiveness could, perhaps, depend more on large expertise differences than on prior knowledge differences.

There is scant literature on the diagnostic process during simulation-based learning and reflection phases. One study indicated that participants' hypotheses improved at a second measurement point after reflection phases compared to a first measurement point before reflection phases (Mamede et al., 2020). Also, it seems plausible that hypotheses improve over the course of simulation-based learning as participants gather more and more diagnostic cues and reason about the case. Article 3 (Fink, Heitzmann et al., 2021) provides empirical evidence for both of these assumptions. Participants substantially improved their hypotheses during simulation-based learning. Significantly, but to a lesser extent, participants also improved their hypotheses during reflection phases. The latter finding highlights that reflection phases improved diagnosing in the learning phase. This effect, however, did not transfer to improved diagnostic accuracy from pre- to post-test.

Other important implications for facilitating diagnostic competences with scaffolding can be gained from this dissertation's analyses of the relationship between diagnostic process variables and diagnostic quality. These analyses stem from an assessment context without scaffolding (Article 1 and Article 2) but may also be applicable in a facilitation context that includes scaffolding. Article 2 (Fink et al., submitted) showed that diagnostic activities uniquely contribute to diagnostic quality and should not be considered a mere embodiment of professional knowledge. Article 1 (Fink, Reitmeier et al., 2021) complemented these findings by highlighting that evidence generation correlates positively with diagnostic accuracy in virtual patients. These findings indicate, together with other literature (e.g., Kramer et al., 2021), that diagnostic activities are associated with diagnostic quality. Scaffolding could make use of this association between diagnostic activities and diagnostic quality for learning purposes. I believe two different ways of implementing this are possible. First, scaffolding can be used to support learners in carrying out diagnostic activities. Later, the support may be removed by fading, and learners may be able to conduct the diagnostic activities without support. Second, practitioners can develop explicit quality criteria for diagnostic activities. The diagnostic process may then be taught systematically throughout higher education by using these quality criteria for instruction and support. Both of these ways of scaffolding based on the diagnostic process could, in turn, enhance the acquisition of diagnostic competences in medical education.

The effectiveness of simulation-based learning and scaffolding for facilitating diagnostic competences have been repeatedly demonstrated and should not be disputed. Nevertheless, meta-analytical results (Chernikova, Heitzmann, Stadler et al., 2020) and the results from Article 3 (Fink, Heitzmann et al., 2021) indicate that reflection phases do not have an added benefit in simulation-based learning. It is an open question why reflection phases seem to be more effective in other contexts apart from simulations, such as problem-based learning from text-based cases (Chernikova, Heitzmann, Fink et al., 2020; Mamede & Schmidt, 2017). I provide two explanations for this effect: differences in the case format and information processing between text-based learning and simulation-based learning. If the finding that reflection phases have no added benefit for the simulation-based learning of diagnostic competences replicates, other types of scaffolding should be provided in this context.

5.4 Limitations and Directions for Future Research

The next section first discusses some limitations of this dissertation and then provides directions for future research in this field.

One limitation of the empirical studies conducted for this dissertation concerns case specificity. Case-specificity means that an individual's performance in a sample of cases typically does not generalize well to other contexts (Elstein, Shulman, & Sprafka, 1978). This problem is inherent in all assessment methods that aim to achieve high face validity by providing contextual cues and also affects simulations (Swanson & Roberts, 2016; van der Vleuten & Schuwirth, 2019). Due to the use of a relatively small number of cases with realistic durations in the conducted studies, the analyses may suffer from case specificity to some extent. The key feature approach (Page & Bordage, 1995) showed that case specificity can be reduced by using a larger number of short text vignettes that contain critical features of a case. Like the key feature approach, future research could use shorter simulations with a narrower focus on the critical features of each case. Then, a larger number of cases could be used and case specificity could be mitigated.

Other limitations concern the measures used and the underlying conceptualization of the diagnostic process in this dissertation. The measures used to capture the diagnostic process focused mainly on the three diagnostic activities of hypothesis generation, evidence generation, and evidence evaluation from the framework by Heitzmann et al. (2019). Other diagnostic activities contained within this framework, such as communicating the process/results (see Table 1), were not captured in the conducted studies. Moreover, different patterns and sequences of diagnostic activities were not investigated in the conducted studies. This is important, since recent literature indicates that patterns and sequences can also explain variance in problem-solving (Stadler, Fischer, & Greiff, 2019). Finally, other conceptualizations of the diagnostic process from medical education, such as the illness script theory (Schmidt et al., 1990) and the dual-process theory (Croskerry, 2009; Eva, 2004), were not investigated in this dissertation. However, the main advantage of conceptualizing diagnosing as problem-solving in line with Heitzmann et al. (2019) is that the diagnostic process and associated diagnostic activities can be compared across domains. Thus, the results of this dissertation can contribute to future interdisciplinary research on diagnostic processes.

A third limitation concerns the methodology used to assess the diagnostic process. Instead of the behavior- and prompt-based approach employed in this dissertation, think-aloud, eye-tracking, and concept mapping can also be used to capture diagnostic processes (Hege, Kononowicz, Kiesewetter, & Foster-Johnson, 2018; Kok & Jarodzka, 2017; Pinnock, Young, Spence, Henning, & Hazell, 2015). Eye-tracking is minimally intrusive for the participant. However, this method requires theory-driven, purposeful, and prospective adjustments to instructional design in simulations in order to gain insights into specific diagnostic processes. Think-aloud necessitates sophisticated qualitative analyses and may induce the beneficial learning process of self-explanation (VanLehn, Jones, & Chi, 1992). Concept mapping and the behavior- and prompt-based approach used in this dissertation can be integrated well into simulations without needing to extensively adjust the instructional design in advance. However, these two methods may trigger learning through elaboration (Pressley, McDaniel, Turnure, Wood, & Ahmad, 1987). Nevertheless, even though the method used in this dissertation may sometimes have induced elaboration processes, it seems particularly suitable for capturing diagnostic processes in larger quantitative studies applying realistic simulations.

I now provide directions for future research, touching upon research on case characteristics, adaptivity of instruction and assessment, and the possible transfer of findings from assessment to facilitation settings.

I believe that more research is warranted on the characteristics of the cases used in simulation-based education. A licensed physician developed the cases for the conducted studies, which I validated by conducting an expert workshop and a pilot study (see Section 1.6). Nevertheless, the licensed physician found it difficult to systematically develop cases with a specific level of complexity and difficulty. This is because there is little theory on case characteristics such as case typicality and complexity in medical education (Braun, Lenzer, Fischer, & Schmidmaier, 2019; Custers, Boshuizen, & Schmidt, 1996; Papa & Elieson, 1993). Compared to the elaborate theories from psychology and cognitive science dealing with the properties of items in intelligence tests (e.g., Carpenter, Just, & Shell, 1990; Meo, Roberts, &

Marucci, 2007), this type of research is still in its infancy in medical education. Therefore, research on case characteristics can lead to important advances. For simulation-based assessment, research on this topic can improve case development and provide insight into the relationships between case characteristics and the diagnostic process. For simulation-based learning, this type of research may uncover to what extent case characteristics themselves evoke learning and how case characteristics interact with scaffolding.

Another interesting avenue for research on simulations concerns the adaptivity of instruction and assessment. Adaptive instruction means that teaching and support are tailored to individual learners' characteristics and performance in order to facilitate learning (Plass & Pawar, 2020a). Research on adaptive instruction explores which degree of adaptivity is optimal and what learner characteristics and performance indicators best facilitate learning. This type of instruction and support could be particularly effective in contexts such as Article 3 (Fink, Heitzmann et al., 2021), in which regular scaffolding did not have an added benefit for acquiring diagnostic competences. Adaptive assessment, on the other hand, refers to a test that selects test items from an item pool based on continuous scoring of the assessee's performance (Weiss & Kingsbury, 1984). Promising research could be conducted on topics such as the optimal number of simulations for reliable and valid assessment and the suitability of different algorithms for estimating competence levels. As this type of research progresses and incorporates theoretical advances on case characteristics, issues such as case specificity that might also pertain to the empirical studies conducted can be further reduced.

The final direction for future research to be discussed here is the question of to what extent findings from assessment settings can be transferred to learning settings and vice versa. Article 1 (Fink, Reitmeier et al., 2021) and Article 2 (Fink et al., submitted) focused on simulation-based assessment, while Article 3 (Fink, Heitzmann et al., 2021) explored simulation-based learning. Only in the simulation-based learning setting did the simulations include additional scaffolding and expert solutions. It is known that scaffolding can affect diagnostic process variables such as cognitive load (Hmelo-Silver, Duncan, & Chinn, 2006). Moreover, watching expert solutions could improve the diagnostic process via vicarious learning and conveying feedback (Bandura, 1971; Hattie & Timperley, 2007). In light of these points, it is an open question to what extent the reported findings are invariant across assessment and learning settings. For instance, it is not known whether the finding that diagnostic activities explain a medium proportion of the variance in diagnostic quality in assessment settings (Article 2, Fink et al., submitted) is also valid in learning settings with scaffolding and expert solutions. I believe that research integrating data from assessment and learning settings could answer this and similar research questions with techniques such as meta-analytic structural equation modeling (Cheung, 2015).

5.5 Conclusions

I conducted two empirical studies and wrote three articles to enhance our understanding of the assessment and facilitation of diagnostic competences with simulations. In both studies, diagnostic competences were assessed in the context of medicine. My research contributes the following main points to the existing literature. First, the use of different assessment methods can evoke differences in diagnostic accuracy. As reported in Article 1 (Fink, Reitmeier et al., 2021), diagnostic accuracy was higher in standardized patients than in virtual patients, and participants would have achieved a better grade in standardized patients than in virtual patients. This finding highlights, particularly for summative assessment, that simulation modalities should not be substituted one-for-one for each other but be adapted to the specific context. Second, as pointed out by Article 2 (Fink et al., submitted), the diagnostic process can be operationalized well with diagnostic activities and is related to diagnostic quality. In fact, diagnostic activities explain medium proportions of variance in diagnostic quality, and this unique contribution of the diagnostic process adds to professional knowledge's contribution to diagnostic quality. Third, as shown in Article 3 (Fink, Heitzmann et al., 2021), reflection phases did not have an added benefit for learning to diagnose accurately in virtual patients. This finding can perhaps be explained by the case format and information processing used in simulation-based learning. Together, the three articles add to the literature on assessing and facilitating diagnostic competences with simulations. I believe that these lines of research will further grow in popularity and prove fruitful for development and changes in education.

I hope that the exciting discoveries made in this dissertation will be critically examined, questioned, and taken up by the literature. I would like to close this dissertation with a personal remark and an annotated reference to the quote with which I began this text. After working on this dissertation for several years, I am now even more convinced that simulations should retain a crucial role in the assessment and facilitation of diagnostic competences in medical education. Moreover, I anticipate that the popularity of simulations will grow in other fields of education as well. I fullheartedly agree with the quote that learning is "a continuous, life-long process resulting from acting in situations” (Brown et al., 1989, p. 33) and thus needs to be assessed with respect to relevant tasks and in realistic settings, preferably through appropriate methods such as simulations.

References

Bandura, A. (1971). Vicarious and self-reinforcement processes. In R. Glaser (Ed.), *The nature of reinforcement: A symposium of the Learning Research and Development Center, University of Pittsburgh* (pp. 228–278). New York, NY: Academic Press.

Barrows, H. S. (1996). Problem-based learning in medicine and beyond: A brief overview. *New Directions for Teaching and Learning*, *1996*(68), 3–12. https://doi.org/10.1002/tl.37219966804

Barrows, H. S., & Abrahamson, S. (1964). The programmed patient: A technique for appraising student performance in clinical neurology. *Academic Medicine*, *39*(8), 802–805.

Beauchamp, C. (2015). Reflection in teacher education: Issues emerging from a review of current literature. *Reflective Practice*, *16*(1), 123–141. https://doi.org/10.1080/14623943.2014.982525

Belland, B. R. (2014). Scaffolding: Definition, current debates, and future directions. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed., Vol. 26, pp. 505–518). Dordrecht, Netherlands: Springer. https://doi.org/10.1007/978-1-4614-3185-5_39

Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2017). Synthesizing results from empirical research on computer-based scaffolding in STEM education: A meta-analysis. *Review of Educational Research*, *87*(2), 309–344. https://doi.org/10.3102/0034654316670999

Berliner, D., Schneider, N., Welte, T., & Bauersachs, J. (2016). The differential diagnosis of dyspnea. *Deutsches Ärzteblatt International*, *113*(49), 834–845. https://doi.org/10.3238/arztebl.2016.0834

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(Suppl. 5A), S2-S23. https://doi.org/10.1016/j.amjmed.2008.01.001

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194f

Boshuizen, H. P. A., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science*, *16*(2), 153–184. https://doi.org/10.1016/0364-0213(92)90022-M

Boulet, J. R. (2008). Summative assessment in medicine: The promise of simulation for high-stakes evaluation. *Academic Emergency Medicine*, *15*(11), 1017–1024. https://doi.org/10.1111/j.1553-2712.2008.00228.x

Boulet, J. R., & Durning, S. J. (2019). What we measure … and what we should measure in medical education. *Medical Education*, *53*(1), 86–94. https://doi.org/10.1111/medu.13652

Braun, L. T., Zottmann, J. M., Adolf, C., Lottspeich, C., Then, C., Wirth, S., . . . Schmidmaier, R. (2017). Representation scaffolds improve diagnostic efficiency in medical students. *Medical Education*, *51*(11), 1118–1126. https://doi.org/10.1111/medu.13355

Braun, L. T., Lenzer, B., Fischer, M. R., & Schmidmaier, R. (2019). Complexity of clinical cases in simulated learning environments: Proposal for a scoring system. *GMS Journal for Medical Education*, *36*(6), Doc80. https://doi.org/10.3205/zma001288

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42. https://doi.org/10.3102/0013189X018001032

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404

Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., . . . Suh, K. S. (2015). Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of Clinical Bioinformatics*, *5*, 4. https://doi.org/10.1186/s13336-015-0019-3

Charlin, B., Boshuizen, H. P. A., Custers, E., & Feltovich, P. J. (2007). Scripts and clinical reasoning. *Medical Education*, *41*(12), 1178–1184. https://doi.org/10.1111/j.1365-2923.2007.02924.x

Charlin, B., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. (2000). The Script Concordance test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine*, *12*(4), 189–195. https://doi.org/10.1207/s15328015tlm1204_5

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., & Fischer, F. (2020). Facilitating diagnostic competences in higher education - A meta-analysis in medical and teacher education. *Educational Psychology Review*, *32*(1), 157–196. https://doi.org/10.1007/s10648-019-09492-2

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research*, *90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Cheung, M. W.-L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*, 1–7. https://doi.org/10.3389/fpsyg.2014.01521

Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1999). Computer-based case simulations from medicine: Assessing skills in patient management. In A. Tekian, C. H. McGuire, & W. C. McGaghie (Eds.), *Innovative simulations for assessing professional competence: From paper-and-pencil to virtual reality* (pp. 29–41). Chicago, IL: University of Illinois at Chicago.

Coderre, S., Wright, B., & McLaughlin, K. (2010). To think is good: Querying an initial hypothesis reduces diagnostic error in medical students. *Academic Medicine*, *85*(7), 1125–1129. https://doi.org/10.1097/ACM.0b013e3181e1b229

Cohen, D. S., Colliver, J. A., Marcy, M. S., Fried, E. D., & Swartz, M. H. (1996). Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. *Academic Medicine*, *71*(Suppl. 1), S87-S89. https://doi.org/10.1097/00001888-199601000-00052

Cook, D. A., Brydges, R., Hamstra, S. J., Zendejas, B., Szostek, J. H., Wang, A. T., . . . Hatala, R. (2012). Comparative effectiveness of technology-enhanced simulation versus other instructional methods: A systematic review and meta-analysis. *Simulation in Healthcare*, *7*(5), 308–320. https://doi.org/10.1097/SIH.0b013e3182614f95

Cook, D. A., Erwin, P. J., & Triola, M. M. (2010). Computerized virtual patients in health professions education: A systematic review and meta-analysis. *Academic Medicine*, *85*(10), 1589–1602. https://doi.org/10.1097/ACM.0b013e3181edfe13

Cook, D. A., Hamstra, S. J., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., . . . Hatala, R. (2013). Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. *Medical Teacher*, *35*(1), e867-e898. https://doi.org/10.3109/0142159X.2012.714886

COSIMA research unit (2021). COSIMA framework model. Retrieved from https://www.for2385.uni-muenchen.de/aktuelles/rahmenmodellccby/cosima-frame-model_eng_phase1.pdf (accessed 2021/04/26).

Crawford, M. P. (1966). Dimensions of simulation. *American Psychologist*, *21*(8), 788–796. https://doi.org/10.1037/h0023974

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine*, *84*(8), 1022–1028. https://doi.org/10.1097/ACM.0b013e3181ace703

Custers, E. J. F. M., Boshuizen, H. P. A., & Schmidt, H. G. (1996). The influence of medical expertise, case typicality, and illness script component on case processing and disease probability estimates. *Memory & Cognition*, *24*(3), 384–399. https://doi.org/10.3758/bf03213301

Daniel, M., Rencic, J., Durning, S. J., Holmboe, E., Santen, S. A., Lang, V., . . . Gruppen, L. D. (2019). Clinical reasoning assessment methods: A scoping review and practical guidance. *Academic Medicine*, *94*(6), 902–912. https://doi.org/10.1097/ACM.0000000000002618

Dankbaar, M. E. W., Alsma, J., Jansen, E. E. H., van Merrienboer, J. J. G., van Saase, J. L. C. M., & Schuit, S. C. E. (2016). An experimental study on the effects of a simulation game on students' clinical cognitive skills and motivation. *Advances in Health Sciences Education*, *21*(3), 505–521. https://doi.org/10.1007/s10459-015-9641-x

Davidsson, P., & Verhagen, H. (2013). Types of simulation. In B. Edmonds & R. Meyer (Eds.), *Simulating social complexity: A handbook* (pp. 23–36). Berlin, Germany: Springer.

de Jong, T. (1991). Learning and instruction with computer simulations. *Education and Computing*, *6*(3-4), 217–229. https://doi.org/10.1016/0167-9287(91)80002-F

Dillon, G. F., Boulet, J. R., Hawkins, R. E., & Swanson, D. B. (2004). Simulations in the United States Medical Licensing Examination^{TM} (USMLE^{TM}). *Quality and Safety in Health Care*, *13*(Suppl. 1), i41-i45. https://doi.org/10.1136/qshc.2004.010025

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction*, *13*(5), 533–568. https://doi.org/10.1016/S0959-4752(02)00025-7

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Edelstein, R. A., Reid, H. M., Usatine, R., & Wilkes, M. S. (2000). A comparative study of measures to evaluate medical students' performances. *Academic Medicine*, *75*(8), 825–833. https://doi.org/10.1097/00001888-200008000-00016

Ellaway, R., Candler, C., Greene, P., & Smothers, V. (2006). MedBiquitous Virtual Patients (MVP): An architectural model for MedBiquitous Virtual Patients. Retrieved from http://groups.medbiq.org/medbiq/display/VPWG/MedBiquitous+Virtual+Patient+Architecture (accessed 2021/04/26).

Elstein, A. S. (2009). Thinking about diagnostic thinking: A 30-year perspective. *Advances in Health Sciences Education*, *14*(Suppl. 1), 7–18. https://doi.org/10.1007/s10459-009-9184-0

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: The study of clinical performance. *Medical Education*, *41*(12), 1124–1130. https://doi.org/10.1111/j.1365-2923.2007.02946.x

Eva, K. W. (2004). What every teacher needs to know about clinical reasoning. *Medical Education*, *39*(1), 98–106. https://doi.org/10.1111/j.1365-2929.2004.01972.x

Eva, K. W., Hatala, R. M., LeBlanc, V. R., & Brooks, L. R. (2007). Teaching from the clinical reasoning literature: Combined reasoning strategies help novice diagnosticians overcome misleading information. *Medical Education*, *41*(12), 1152–1158. https://doi.org/10.1111/j.1365-2923.2007.02923.x

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Fink, M. C., Heitzmann, N., Reitmeier, V., Siebeck, M., Fischer, F., & Fischer, M. R. (submitted). Diagnosing virtual patients with partial prior knowledge: The interplay between knowledge and diagnostic activities. Manuscript submitted for publication to Advances in Health Sciences Education**,** June 17, 2022.

Fink, M. C., Heitzmann, N., Siebeck, M., Fischer, F., & Fischer, M. R. (2021). Learning to diagnose accurately through virtual patients: Do reflection phases have an added benefit? *BMC Medical Education*, *21*, 523. https://doi.org/10.1186/s12909-021-02937-9

Fink, M. C., Radkowitsch, A., Bauer, E., Sailer, M., Kiesewetter, J., Schmidmaier, R., . . . Fischer, M. R. (2021). Simulation research and design: A dual-level framework for multi-project research programs. *Educational Technology Research and Development*, *69*, 809–841. https://doi.org/10.1007/s11423-020-09876-0

Fink, M. C., Reitmeier, V., Siebeck, M., Fischer, F., & Fischer, M. R. (2022). Live and video simulations of medical history taking: Theoretical background, design, development and validation of a learning environment. In F. Fischer & A. Opitz (Eds.), *Learning to diagnose with simulations: Examples from teacher education and medical education* (109-122). Cham, Switzerland: Springer.

Fink, M. C., Reitmeier, V., Stadler, M., Siebeck, M., Fischer, F., & Fischer, M. R. (2021). Assessment of diagnostic competences with standardized patients versus virtual patients: Experimental study in the context of history taking. *Journal of Medical Internet Research*, *23*(3), e21196. https://doi.org/10.2196/21196

Frank, J. R., & Danoff, D. (2007). The CanMEDS initiative: Implementing an outcomes-based framework of physician competencies. *Medical Teacher*, *29*(7), 642–647. https://doi.org/10.1080/01421590701746983

Friedman, C. P., France, C. L., & Drossman, D. A. (1991). A randomized comparison of alternative formats for clinical simulations. *Medical Decision Making*, *11*(4), 265–272. https://doi.org/10.1177/0272989X9101100404

Gaba, D. M. (2004). The future vision of simulation in health care. *Quality and Safety in Health Care*, *13*(Suppl. 1), i2-i10. https://doi.org/10.1136/qshc.2004.009878

Gegenfurtner, A., Quesada-Pallarès, C., & Knogler, M. (2014). Digital simulation-based training: A meta-analysis. *British Journal of Educational Technology*, *45*(6), 1097–1114. https://doi.org/10.1111/bjet.12188

Ghanem, C., Kollar, I., Fischer, F., Lawson, T. R., & Pankofer, S. (2016). How do social work novices and experts solve professional problems? A micro-analysis of epistemic activities and the use of evidence. *European Journal of Social Work*, *21*(1), 3–19. https://doi.org/10.1080/13691457.2016.1255931

Graber, M. L., Tompkins, D., & Holland, J. J. (2009). Resources medical students use to derive a differential diagnosis. *Medical Teacher*, *31*(6), 522–527. https://doi.org/10.1080/01421590802167436

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, *111*(9), 2055–2100.

Groves, M., O'Rourke, P., & Alexander, H. (2003). Clinical reasoning: The relative contribution of identification, interpretation and hypothesis errors to misdiagnosis. *Medical Teacher*, *25*(6), 621–625. https://doi.org/10.1080/0142159031000160568 8

Guagnano, M. T., Merlitti, D., Manigrasso, M. R., Pace-Palitti, V., & Sensi, S. (2002). New medical licensing examination using computer-based case simulations and standardized patients. *Academic Medicine*, *77*(1), 87–90. https://doi.org/10.1097/00001888-200201000-00020

Haji, F. A., Cheung, J. J. H., Woods, N., Regehr, G., de Ribaupierre, S., & Dubrowski, A. (2016). Thrive or overload? The effect of task complexity on novices' simulation-based learning. *Medical Education*, *50*(9), 955–968. https://doi.org/10.1111/medu.13086

Hallinger, P., & Wang, R. (2020). The evolution of simulation-based learning across the disciplines, 1965–2018: A science map of the literature. *Simulation & Gaming*, *51*(1), 9–32. https://doi.org/10.1177/1046878119888246

Hamstra, S. J., Brydges, R., Hatala, R., Zendejas, B., & Cook, D. A. (2014). Reconsidering fidelity in simulation-based training. *Academic Medicine*, *89*(3), 387–392. https://doi.org/10.1097/ACM.0000000000000130

Harden, R. M., Stevenson, M., Wilson Downie, W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, *1*(1), 447–451. https://doi.org/10.1136/bmj.1.5955.447

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hawkins, R., MacKrell Gaglione, M., LaDuca, T., Leung, C., Sample, L., Gliva-McConvey, G., . . . Ciccone, A. (2004). Assessment of patient management skills and clinical skills of practising doctors using computer-based case simulations and standardised patients. *Medical Education*, *38*(9), 958–968. https://doi.org/10.1111/j.1365-2929.2004.01907.x

Hege, I., Kononowicz, A. A., Kiesewetter, J., & Foster-Johnson, L. (2018). Uncovering the relation between clinical reasoning and diagnostic accuracy - An analysis of learner's clinical reasoning processes in virtual patients. *PloS One*, *13*(10), e0204900. https://doi.org/10.1371/journal.pone.0204900

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., . . . Fischer, F. (2019). Facilitating diagnostic competences in simulations in higher education: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research*, *7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Heitzmann, N., Fischer, M. R., & Fischer, F. (2017). Towards more systematic and better theorised research on simulations. *Medical Education*, *51*(2), 129–131. https://doi.org/10.1111/medu.13239

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., . . . Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model.

Teaching and Teacher Education, *76*(4), 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2006). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, *42*(2), 99–107. https://doi.org/10.1080/00461520701263368

Hofer, M. (2016). *Presence und Involvement* (1st ed.). Baden-Baden, Germany: Nomos. https://doi.org/10.5771/9783845263540

Huang, G., Reynolds, R., & Candler, C. (2007). Virtual patient simulation at US and Canadian medical schools. *Academic Medicine*, *82*(5), 446–451. https://doi.org/10.1097/ACM.0b013e31803e8a0a

Huwendiek, S., de Leng, B. A., Zary, N., Fischer, M. R., Ruiz, J. G., & Ellaway, R. (2009). Towards a typology of virtual patients. *Medical Teacher*, *31*(8), 743–748. https://doi.org/10.1080/01421590903124708

Issenberg, S. B., McGaghie, W. C., Petrusa, E. R., Gordon, D. L., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher*, *27*(1), 10–28. https://doi.org/10.1080/01421590500046924

Jones, F., Passos-Neto, C. E., & Braghiroli, O. F. M. (2015). Simulation in medical education: Brief history and methodology. *Principles and Practice of Clinical Research*, *1*(2), 56–63. https://doi.org/10.21801/ppcrj.2015.12.8

Jünger, J. (2018). Kompetenzorientiert prüfen im Staatsexamen Medizin. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, *61*(2), 171–177. https://doi.org/10.1007/s00103-017-2668-9

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed.). New York, NY: Macmillan.

Kaplan, A. D., Cruit, J., Endsley, M., Beers, S. M., Sawyer, B. D., & Hancock, P. A. (2020). The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: A Meta-Analysis. *Human Factors*, *63*(4), 706-726. https://doi.org/10.1177/0018720820904229

Kaufman, D., & Ireland, A. (2016). Enhancing teacher education with simulations. *TechTrends*, *60*(3), 260–267. https://doi.org/10.1007/s11528-016-0049-0

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

Kelly, G. J. (2008). Inquiry, activity and epistemic practice. In R. Duschl & R. E. Grandy (Eds.), *Teaching scientific inquiry: Recommendations for research and implementation* (pp. 99–117). Rotterdam, Netherlands: Sense Publishers.

Khan, K. Z., Tolhurst-Cleaver, S., White, S., & Simpson, W. (2011). *Simulation in healthcare education. Building a simulation programme: A practical guide*. *AMEE guides: Vol. 2*. Dundee, Scotland: Association for Medical Education in Europe.

Kiesewetter, J., Fischer, F., & Fischer, M. R. (2017). Collaborative clinical reasoning - A systematic review of empirical studies. *The Journal of Continuing Education in the Health Professions*, *37*(2), 123–128. https://doi.org/10.1097/CEH.0000000000000158

Kiesewetter, J., Sailer, M., Jung, V. M., Schönberger, R., Bauer, E., Zottmann, J. M., . . . Fischer, M. R. (2020). Learning clinical reasoning: How virtual patient case format and prior knowledge interact. *BMC Medical Education*, *20*(1), 73. https://doi.org/10.1186/s12909-020-1987-y

Kim, S., Phillips, W. R., Pinsky, L., Brock, D., Phillips, K., & Keary, J. (2006). A conceptual framework for developing teaching cases: A review and synthesis of the literature across disciplines. *Medical Education*, *40*(9), 867–876. https://doi.org/10.1111/j.1365-2929.2006.02544.x

Klahr, D., & Simon, H. A. (2001). What have psychologists (and others) discovered about the process of scientific discovery? *Current Directions in Psychological Science*, *10*(3), 75–79. https://doi.org/10.1111/1467-8721.00119

Knox, J. D. E. (1975). *The modified essay question: ASME medical education booklet No. 5*. Dundee, Scotland: Association for the Study of Medical Education.

Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education*, *51*(1), 114–122. https://doi.org/10.1111/medu.13066

Kolodner, J. L. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review*, *6*(1), 3–34. https://doi.org/10.1007/BF00155578

Kramer, M., Förtsch, C., Seidel, T., & Neuhaus, B. J. (2021). Comparing two constructs for describing and analyzing teachers' diagnostic processes. *Studies in Educational Evaluation*, *68*(7), 100973. https://doi.org/10.1016/j.stueduc.2020.100973

LeBlanc, V. R., Brooks, L. R., & Norman, G. R. (2002). Believing is seeing: The influence of a diagnostic hypothesis on the interpretation of clinical features. *Academic Medicine*, *77*(Suppl. 10), S67-S69. https://doi.org/10.1097/00001888-200210001-00022

LeBlanc, V. R., Norman, G. R., & Brooks, L. R. (2001). Effect of a diagnostic suggestion on diagnostic accuracy and identification of clinical features. *Academic Medicine*, *76*(Suppl. 10), S18- S20.

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education*, *91*, 103059. https://doi.org/10.1016/j.tate.2020.103059

Luctkar-Flude, M., Wilson-Keates, B., & Larocque, M. (2012). Evaluating high-fidelity human simulators and standardized patients in an undergraduate nursing health assessment course. *Nurse Education Today*, *32*(4), 448–452. https://doi.org/10.1016/j.nedt.2011.04.011

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, *60*(11), 225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007

Mamede, S., & Schmidt, H. G. (2017). Reflection in medical diagnosis: A literature review. *Health Professions Education*, *3*(1), 15–25. https://doi.org/10.1016/j.hpe.2017.01.003

Mamede, S., Hautz, W. E., Berendonk, C., Hautz, S. C., Sauter, T. C., Rotgans, J., . . . Schmidt, H. G. (2020). Think Twice: Effects on diagnostic accuracy of returning to the case to reflect upon the initial diagnosis. *Academic Medicine*, *95*(8), 1223–1229. https://doi.org/10.1097/ACM.0000000000003153

Mamede, S., Schmidt, H. G., & Penaforte, J. C. (2008). Effects of reflective practice on the accuracy of medical diagnoses. *Medical Education*, *42*(5), 468–475. https://doi.org/10.1111/j.1365-2923.2008.03030.x

Mamede, S., Schmidt, H. G., Rikers, R. M. J. P., Custers, E. J. F. M., Splinter, T. A. W., & van Saase, J. L. C. M. (2010). Conscious thought beats deliberation without attention in diagnostic decision-making: At least when you are an expert. *Psychological Research*, *74*(6), 586–592. https://doi.org/10.1007/s00426-010-0281-8

Mamede, S., van Gog, T., Moura, A. S., de Faria, R. M. D., Peixoto, J. M., Rikers, R. M. J. P., & Schmidt, H. G. (2012). Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Medical Education*, *46*(5), 464–472. https://doi.org/10.1111/j.1365-2923.2012.04217.x

Mamede, S., van Gog, T., Sampaio Moura, A. S., de Faria, R. M. D., Peixoto, J. M., & Schmidt, H. G. (2014). How can students' diagnostic competence benefit most from practice with clinical cases? The effects of structured reflection on future diagnosis of the same and novel diseases. *Academic Medicine*, *89*(1), 121–127. https://doi.org/10.1097/ACM.0000000000000076

Maran, N. J., & Glavin, R. J. (2003). Low- to high-fidelity simulation - A continuum of medical education? *Medical Education*, *37*(Suppl. 1), 22–28. https://doi.org/10.1046/j.1365-2923.37.s1.9.x

Mayer, R. E., & Moreno, R. (1998). A cognitive theory of multimedia learning: Implications for design principles. *Journal of Educational Psychology*, *91*(2), 358–368.

McCarthy, W. H., & Gonnella, J. S. (1967). The simulated Patient Management Problem: A technique for evaluating and teaching clinical competence. *British Journal of Medical Education*, *1*(5), 348–352. https://doi.org/10.1111/j.1365-2923.1967.tb01730.x

McGuire, C. H., & Babbott, D. (1967). Simulation technique in the measurement of problem-solving skills. *Journal of Educational Measurement*, *4*(1), 1–10.

Meller, G. (1997). A typology of simulators for medical education. *Journal of Digital Imaging*, *10*(Suppl. 3), 194–196.

Meo, M., Roberts, M. J., & Marucci, F. S. (2007). Element salience as a predictor of item difficulty for Raven's Progressive Matrices. *Intelligence*, *35*(4), 359–368. https://doi.org/10.1016/j.intell.2006.10.001

MFT (Medizinischer Fakultätentag der Bundesrepublik Deutschland e.V.) (2015). Nationaler Kompetenzbasierter Lernzielkatalog Medizin (NKLM). Retrieved from http://www.nklm.de/files/nklm_final_2015-07-03.pdf (accessed 2021/04/26).

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*(Suppl. 9), S63-S67. https://doi.org/10.1097/00001888-199009000-00045

Monteiro, S. M., & Norman, G. (2013). Diagnostic reasoning: Where we've been, where we're going. *Teaching and Learning in Medicine*, *25*(Suppl. 1), S26-S32. https://doi.org/10.1080/10401334.2013.842911

Morcke, A. M., Dornan, T., & Eika, B. (2013). Outcome (competency) based education: An exploration of its origins, theoretical basis, and empirical evidence. *Advances in Health Sciences Education*, *18*(4), 851–863. https://doi.org/10.1007/s10459-012-9405-9

Nendaz, M. R., & Bordage, G. (2002). Promoting diagnostic problem representation. *Medical Education*, *36*(8), 760–766. https://doi.org/10.1046/j.1365-2923.2002.01279.x

Nguyen, Q. D., Fernandez, N., Karsenti, T., & Charlin, B. (2014). What is reflection? A conceptual analysis of major definitions and a proposal of a five-component model. *Medical Education*, *48*(12), 1176–1189. https://doi.org/10.1111/medu.12583

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education*, *39*(4), 418–427. https://doi.org/10.1111/j.1365-2929.2005.02127.x

Norman, G., Dore, K., & Grierson, L. (2012). The minimal relationship between simulation fidelity and transfer of learning. *Medical Education*, *46*(7), 636–647. https://doi.org/10.1111/j.1365-2923.2012.04243.x

Page, G., & Bordage, G. (1995). The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. *Academic Medicine*, *70*(2), 104–110. https://doi.org/10.1097/00001888-199502000-00012

Papa, F. J., & Elieson, B. (1993). Diagnostic accuracy as a function of prototypicality. *Academic Medicine*, *68*(Suppl. 10), S58-S-60. https://doi.org/10.1097/00001888-199310000-00046

Patel, V. L., Evans, D. A., & Groen, G. J. (1989). Biomedical knowledge and clinical reasoning. In D. A. Evans & V. L. Patel (Eds.), *Cognitive science in medicine: Biomedical modeling* (53-112). Cambridge, MA: MIT Press.

Peterson, M. C., Holbrook, J. H., Von Hales, D., 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.

Pinnock, R., Young, L., Spence, F., Henning, M., & Hazell, W. (2015). Can think aloud be used to teach and assess clinical reasoning in graduate medical education? *Journal of Graduate Medical Education*, *7*(3), 334–337. https://doi.org/10.4300/JGME-D-14-00601.1

Plass, J. L., & Pawar, S. (2020a). Adaptivity and personalization in games for learning. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 263–282). Cambridge, MA: MIT Press.

Plass, J. L., & Pawar, S. (2020b). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943

Pressley, M., McDaniel, M. A., Turnure, J. E., Wood, E., & Ahmad, M. (1987). Generation and precision of elaboration: Effects on intentional and incidental learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(2), 291–300. https://doi.org/10.1037/0278-7393.13.2.291

Radkowitsch, A., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2020). Learning to diagnose collaboratively: Validating a simulation for medical students. *GMS Journal for Medical Education*, *37*(5), Doc51. https://doi.org/10.3205/zma001344

Renkl, A., Mandl, H., & Gruber, H. (1996). Inert knowledge: Analyses and remedies. *Educational Psychologist*, *31*(2), 115–121. https://doi.org/10.1207/s15326985ep3102_3

Rethans, J. J., Sturmans, F., Drop, R., & van der Vleuten, C. (1991). Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *British Journal of General Practice*, *41*, 97–99.

Saber Tehrani, A. S., Lee, H., Mathews, S. C., Shore, A., Makary, M. A., Pronovost, P. J., & Newman-Toker, D. E. (2013). 25-Year summary of US malpractice claims for diagnostic errors 1986-2010: An analysis from the National Practitioner Data Bank. *BMJ Quality & Safety*, *22*(8), 672–680. https://doi.org/10.1136/bmjqs-2012-001550

Salas, E., Rosen, M. A., Held, J. D., & Weissmuller, J. J. (2009). Performance measurement in simulation-based training: A review and best practices. *Simulation & Gaming*, *40*(3), 328–376. https://doi.org/10.1177/1046878108326734

Scalese, R. J., Obeso, V. T., & Issenberg, S. B. (2008). Simulation technology for skills training and competency assessment in medical education. *Journal of General Internal Medicine*, *23*(Suppl. 1), 46–49. https://doi.org/10.1007/s11606-007-0283-4

Schleicher, A., & Zoido, P. (2016). The policies that shaped PISA, and the policies that PISA shaped. In K. Mundy, A. Green, B. Lingard, & A. Verger (Eds.), *The handbook of global education policy* (Vol. 32, pp. 374–384). Chichester, England: Wiley. https://doi.org/10.1002/9781118468005.ch20

Schmidmaier, R., Eiber, S., Ebersbach, R., Schiller, M., Hege, I., Holzer, M., & Fischer, M. R. (2013). Learning the facts in medical school is not enough: Which factors predict successful application of procedural knowledge in a laboratory setting? *BMC Medical Education*, *13*(1), 28. https://doi.org/10.1186/1472-6920-13-28

Schmidt, H. G., Norman, G., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine*, *65*(10), 611–621. https://doi.org/10.1097/00001888-199010000-00001

Schmidt, H. G., & Rikers, R. M. J. P. (2007). How expertise develops in medicine: Knowledge encapsulation and illness script formation. *Medical Education*, *41*(12), 1133–1139. https://doi.org/10.1111/j.1365-2923.2007.02915.x

Schubert, T., Friedman, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Teleoperators & Virtual Environments*, *10*(3), 266–281. https://doi.org/10.1162/105474601300343603

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., . . . Mislevy, R. (2009). Epistemic Network Analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media*, *1*(2), 33–53. https://doi.org/10.1162/ijlm.2009.0013

Simpson, M. A. (1985). How to use role-play in medical teaching. *Medical Teacher*, *7*(1), 75–82. https://doi.org/10.3109/01421598509036794

Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology*, *10*, 777. https://doi.org/10.3389/fpsyg.2019.00777

Stark, R., Kopp, V., & Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: Two experimental studies in undergraduate medical education. *Learning and Instruction*, *21*(1), 22–33. https://doi.org/10.1016/j.learninstruc.2009.10.001

Stillman, P. L., Swanson, D. B., Smee, S., Stillman, A. E., Ebert, T. H., Emmel, V. S., . . . Willms, J. (1986). Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine*, *105*(5), 762–771. https://doi.org/10.7326/0003-4819-105-5-762

Stojan, J. N., Daniel, M., Morgan, H. K., Whitman, L., & Gruppen, L. D. (2017). A randomized cohort study of diagnostic and therapeutic thresholds in medical student clinical reasoning. *Academic Medicine*, *92*(Suppl. 11), S43-S47. https://doi.org/10.1097/ACM.0000000000001909

Swanson, D. B., Norman, G. R., & Linn, R. L. (1995). Performance-based assessment: Lessons from the health professions. *Educational Researcher*, *24*(5), 5–11. https://doi.org/10.3102/0013189X024005005

Swanson, D. B., & Roberts, T. E. (2016). Trends in national licensing examinations in medicine. *Medical Education*, *50*(1), 101–114. https://doi.org/10.1111/medu.12810

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/a:1022193728205

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tabak, I., & Kyza, E. A. (2018). Research on scaffolding in the learning sciences: A methodological perspective. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 191–200). New York, NY: Routledge. https://doi.org/10.4324/9781315617572-19

Taras, M. (2005). Assessment - summative and formative - some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466–478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Towle, A. (1998). Changes in health care and continuing medical education for the 21st century. *British Medical Journal*, *316*(7127), 301–304. https://doi.org/10.1136/bmj.316.7127.301

van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine*, *2*(2), 58–76. https://doi.org/10.1080/10401339009539432

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2019). Assessment in the context of problem-based learning. *Advances in Health Sciences Education*, *24*(5), 903–914. https://doi.org/10.1007/s10459-019-09909-1

van Thiel, J., Kraan, H. F., & van der Vleuten, C. P. M. (1991). Reliability and feasibility of measuring medical interviewing skills: The revised Maastricht History-Taking and Advice Checklist. *Medical Education*, *25*(3), 224–229. https://doi.org/10.1111/j.1365-2923.1991.tb00055.x

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, *2*(1), 1–59. https://doi.org/10.1207/s15327809jls0201_1

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Williams, R. G., Klamen, D. L., Markwell, S. J., Cianciolo, A. T., Colliver, J. A., & Verhulst, S. J. (2014). Variations in senior medical student diagnostic justification ability. *Academic Medicine*, *89*(5), 790–798. https://doi.org/10.1097/ACM.0000000000000215

Winters, B., Custer, J., Galvagno, S. M., Colantuoni, E., Kapoor, S. G., Lee, H., . . . Newman-Toker, D. (2012). Diagnostic errors in the intensive care unit: A systematic review of autopsy studies. *BMJ Quality & Safety*, *21*(11), 894–902. https://doi.org/10.1136/bmjqs-2012-000803

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

Wood, D. F. (2003). Problem based learning. *British Medical Journal*, *326*(7384), 328–330. https://doi.org/10.1136/bmj.326.7384.328

Woolliscroft, J. O., Calhoun, J. G., Billiu, G. A., Stross, J. K., MacDonald, M., & Templeton, B. (1989). House officer interviewing techniques. *Journal of General Internal Medicine*, *4*(2), 108–114. https://doi.org/10.1007/BF02602349

Wormald, B. W., Schoeman, S., Somasunderam, A., & Penn, M. (2009). Assessment drives learning: An unavoidable truth? *Anatomical Sciences Education*, *2*(5), 199–204. https://doi.org/10.1002/ase.102

Young, J. Q., van Merriënboer, J. J. G., Durning, S., & ten Cate, O. (2014). Cognitive load theory: Implications for medical education. *Medical Teacher*, *36*(5), 371–384. https://doi.org/10.3109/0142159X.2014.889290

Ziv, A. (2009). Simulators and simulation-based medical education. In J. A. Dent & R. M. Harden (Eds.), *A practical guide for medical teachers* (3rd ed., Vol. 2, pp. 217–222). Edinburgh, Scotland: Churchill Livingstone.