

---

# Learning Collaborative Reasoning

Foundations of Adaptive Reflection Support in Agent-  
Based Simulations

---



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

am Munich Center of the Learning Sciences

der Ludwig-Maximilians-Universität

München

vorgelegt von

*Constanze Catharina Richters*

aus Bad Soden am Taunus

München, 2024

First Supervisor / Erstgutachter: Prof. Dr. Frank Fischer, LMU München

Second Supervisor / Zweitgutachter: Prof. Dr. Martin Fischer, LMU Klinikum

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Matthias Stadler, LMU Klinikum

Prof. Dr. Michael Sailer, Universität Augsburg

Datum der mündlichen Prüfung: 05.06.2024

---

## Acknowledgments

Deep appreciation extends to a number of people who have accompanied me on my journey to writing this dissertation and pursuing my PhD. First and foremost, I would like to thank my primary supervisor, Frank Fischer. Frank, you have been an exceptionally fantastic supervisor, guide, and mentor, who has mastered his craft perfectly, always has a listening ear, and possesses an amazing sense of the development of his doctoral students. Your sharing of knowledge and feedback as well as your great support over the years have been instrumental in my professional and personal growth. Likewise, I would like to extend a heartfelt thank you to my second supervisor, Martin Fischer, whose critical perspective, combined with determined support and trust in my progress, has strongly paved my way. Also, thank you, Martin, for reminding me regularly of the importance of our research for medicine. The same gratitude goes to Ralf Schmidmaier, with whom I have greatly enjoyed working over the past years and whose valuable feedback has greatly improved my research. I would also like to thank Vitaliy Popov, who warmly welcomed me to the University of Michigan during my research visit in the final stages of my dissertation. Thank you, Vitaliy, for the enriching, challenging, and enjoyable working sessions that made my research stay a pure pleasure. In this course, I would also like to thank Stephanie Teasley, who thoughtfully assumed the role of a mentor during an extended lunch meeting.

As my day-to-day supervisor and postdoc buddy, I would like to express my deepest gratitude to Matthias Stadler. From the beginning, it felt like we were navigating the same waters, and I found comfort in knowing that I could approach you with any question or concern, whether it was grappling with the persistent R error message, seeking advice on article submission, or navigating networking opportunities at conferences. Your remarkable ability for simplifying complex issues, along with your stoic calmness and patience, has deeply resonated with me and will continue to leave a lasting impression.

Moreover, I would like to thank the entire DFG research unit COSIMA, within which this dissertation was developed. Among the many researchers associated with COSIMA, special thanks go to a few members who accompanied me in particular on this journey. First of all, I would like to thank my doctoral predecessor, Anika Radkowsch, for her tremendous support, especially in the early stages of my dissertation project. I am also grateful to Laura Brandl, who was part of the TP6 team first as a student assistant and later as a doctoral student. Thank you for the wonderful collaboration and your loyalty and reliability even in difficult times. Many thanks also go to Katharina Bach and Andreas Wildner for their invaluable support as student assistants. Thank you, Kathi, for the stimulating conversations

we had about research topics other than mine, such as those related to issues of feminism or educational equity. I would also like to thank my peers Amadeus Pickal and Caroline Corves for all the entertaining, supportive, and sometimes serious conversations, on both a professional and personal level. Special thanks also go to Olga Chernikova and Nicole Heitzmann for their great support in keeping us in touch across the projects. I am also grateful to Michael Sailer, who not only provided me with critical and invaluable feedback on my dissertation, but also engaged in countless entertaining and humorous conversations with me. Whether at conference dinners in Bern, Essen, or Thessaloniki, or at the Christmas parties of our department, his company has added an enjoyable dimension to my academic journey. I would also like to thank the other members of our department, particularly my peers Meral Roeben and Zoya Kozlova for always supporting me emotionally and for sharing their enlightening perspectives on our academic world. Additionally, I thank Alexander Kacina and Simone Steiger for their untiring assistance in various technical and administrative matters, as well as for the enjoyable moments we shared over lunch.

Last but not least, I would like to thank my parents, my brother, and my friends, who have been there for me more than ever over the past few years and have never lost hope in me in times of despair. In particular, I would like to thank my longtime best friends Natalija Simeunović, Juliana Bonitz, Julia Schwengers, Claudia Heiland, Annabel Wolf, Sophia Sauter, & Guillemette Dupont. Thank you for everything—you are one of the most important people in my life. And to my dad: I hope you're finally happy that I make use of my middle name—well, not in the published articles, but at least in my dissertation.



---

## Extended Summary

As a complex skillset, collaborative diagnostic reasoning is crucial in various professional contexts. Professionals (e.g., physicians or teachers) engage in collaborative diagnostic activities, which include individual activities—such as generating and evaluating evidence and hypotheses and drawing conclusions—and collaborative activities—such as eliciting and sharing evidence and hypotheses. High-quality diagnostic outcomes such as accurate diagnoses with well-supported, evidence-based justifications require collaborating professionals to apply different types of knowledge such as content knowledge and collaboration knowledge. Recently, simulation-based learning and scaffolding have been found to be effective instructional means for developing complex skills such as collaborative diagnostic reasoning in higher education. However, a major challenge that educational and psychological researchers have emphasized in light of recent technological advances is how to support learners on the basis of their individual needs. Understanding how learner characteristics such as prerequisites, behavior, or performance are related to their needs for support is critical for effectively adapting instructional support. Various coarse and fine-grained approaches can be used to provide foundations for adaptation. Researchers have frequently used conventional product data, such as prior knowledge data, to investigate the effects of scaffolding for learners with different prior knowledge levels. A newer direction involves analyzing computer-system-generated process data, which can help researchers understand problem-solving processes and their relationships with task outcomes. With help of machine learning, process data may facilitate finer adjustments in real time.

Addressing both approaches, the present PhD dissertation aims to lay foundations for adaptive instructional support for learning collaborative diagnostic reasoning. Previous studies have demonstrated that agent-based simulations, which enable a highly standardized training of collaborative processes, effectively enhance collaborative diagnostic reasoning when combined with collaboration scripts that additionally facilitate collaborative processes. The research in this dissertation builds on and extends previous research by proposing reflection guidance, which encourages learners to reflect on their own activities and performance, as a new effective type of scaffolding in collaborative diagnostic reasoning.

The dissertation comprises three studies conducted in the same agent-based medical simulation where participants in the role of internists diagnosed diseases for several patient cases while collaborating with an agent-based expert radiologist to gather further evidence for the cases. Experimental Studies 1 and 2 investigated conditions under which various types of scaffolding—notably reflection guidance—enhanced the learning of collaborative diagnostic

---

reasoning. The effectiveness of different forms of reflection guidance, tailored to different collaborative diagnostic activities and providing different levels of structure, was examined on the basis of a priori hypotheses. Study 3 used machine learning to analyze collaborative diagnostic reasoning processes and their relationships to the diagnostic outcome.

Study 1 examined the effects of reflection guidance addressing individual activities and collaboration scripts as a function of learners' prior content and collaboration knowledge on collaborative diagnostic reasoning. Collaborative diagnostic reasoning was operationalized as performance in evidence and hypothesis sharing (collaborative activities) and diagnostic accuracy and justification (diagnostic outcomes). Furthermore, Study 1 explored how reflection and collaboration affected the accuracy of suspected diagnoses throughout the reasoning process. Medical students were given questions to help them individually reflect on their initial suspected diagnoses, scripts while collaborating with the radiologist, both, or no support. Results showed that reflection improved hypothesis sharing for learners with high levels of content knowledge, whereas collaboration scripts improved evidence sharing for learners with low levels of content knowledge, suggesting that reflecting on individual activities activates prior content knowledge and prepares learners for collaboration if they have sufficient prior knowledge. Whereas neither collaboration scripts nor reflection guidance improved diagnostic outcomes, collaboration alone improved learners' diagnostic accuracy regardless of their prior knowledge level. These findings may be explained by the integration of external knowledge into the diagnostic process through collaboration with the agent.

Study 2 examined the effects of reflection guidance addressing collaborative activities on collaborative diagnostic reasoning, using the same operationalization as Study 1 and considering learners' prior collaboration knowledge. Medical students received either low-structured (no detailed questions) or high-structured (detailed questions) guidance to help them individually reflect on their collaborative activities or no support at all. Results revealed that reflection guidance was beneficial for learners with low levels of collaboration knowledge. Low-structured guidance improved evidence sharing, diagnostic accuracy, and diagnostic justification, indicating that reflecting on collaborative activities holds promise for not only activating but also restructuring prior knowledge. High-structured guidance improved only diagnostic justification, indicating that different levels of structure in reflection are differentially beneficial for different subskills because different underlying knowledge bases result in different subskill levels. Both low- and high-structured guidance were unhelpful or even detrimental for learners with high collaboration knowledge, suggesting that these learners may require a broader reflection prompt.

---

Study 3 investigated whether and how quickly diagnostic accuracy (diagnostic outcomes) could be predicted from collaborative diagnostic activities using machine learning. Log files of medical students and physicians working in the agent-based simulation were coded as collaborative diagnostic activities, including evidence generation, evidence elicitation, evidence sharing, hypothesis sharing, and drawing conclusions. Bigrams depicting the time spent on and switches between activities were used to train classification algorithms to predict the diagnostician's final diagnosis as either correct or incorrect. Results indicated that diagnostic success was more reliably predicted than failure and before case completion, suggesting that the behavior of unsuccessful diagnosticians underlies diverse cognitive misbehavior, whereas successful diagnosticians exhibit less behavioral variation. Successful diagnosticians spent more time on individual activities, indicating they have an appropriate initial cognitive case representation, whereas unsuccessful diagnosticians spent more time on collaborative activities and switched between individual and collaborative activities.

The dissertation provides theoretical and practical implications for adaptive instructional support for learning collaborative diagnostic reasoning in agent-based simulations. First, guidance on how to reflect on collaborative activities seems particularly promising for learning different subskills of collaborative diagnostic reasoning. A lower degree of structure is thereby likely to promote learning more than a higher degree of structure, regardless of learners' prior knowledge levels. Considering learners' levels in specific subskills beyond prior knowledge seems promising for designing effective reflection support. Nonetheless, the diverse results on the effectiveness of reflection for learners with different levels of prior knowledge in Studies 1 and 2 also highlight the difficulty of comparing and generalizing reflection effects, as well as the difficulty of quantifying the complexity of reflection processes. A complex interplay between factors, such as the content of reflection (e.g., diagnostic decision-making vs. collaboration), learners' prior knowledge and skill level, and the level of structure provided influences the effectiveness of reflection. Future research could continue to strive to objectively scale different levels of structure in reflection support to allow reliable comparisons of effects in the future. Second, the dissertation highlights the importance of theory-based process data to identify subtle differences in collaborative diagnostic reasoning processes between successful and unsuccessful diagnosticians. These findings offer reliable indications of learners' areas of struggle or proficiency in diagnostic cases, allowing for more fine-grained and dynamic instructional support. Such support could enhance the overall effectiveness of simulation-based learning for complex skills such as collaborative diagnostic reasoning in the future.

## Zusammenfassung

Kollaboratives diagnostisches Denken ist eine komplexe Fähigkeit, die in verschiedenen beruflichen Kontexten von großer Bedeutung ist. Während des kollaborativen Diagnostizierens sind Fachkräfte, wie zum Beispiel Mediziner:innen oder Lehrer:innen, an kollaborativen diagnostischen Aktivitäten beteiligt, die individuelle Aktivitäten wie das Generieren und Evaluieren von Evidenzen und Hypothesen und das Ziehen von Schlussfolgerungen sowie kollaborative Aktivitäten wie das Elizitieren und Teilen von Evidenzen und Hypothesen umfassen. Qualitativ hochwertige diagnostische Ergebnisse, genauer gesagt akkurate Diagnosen mit fundierten, evidenzbasierten Begründungen, erfordern von kollaborierenden Fachkräften die Anwendung verschiedener Arten von Wissen, wie zum Beispiel inhaltsbezogenes Wissen und Kollaborationswissen. Aktuelle pädagogisch-psychologische Forschung hat gezeigt, dass simulationsbasiertes Lernen und Scaffolding wirksame instruktionale Unterstützungsmethoden für die Entwicklung komplexer Fähigkeiten wie kollaboratives diagnostisches Denken in der Hochschulbildung sind. Eine große Herausforderung, die angesichts der jüngsten technologischen Fortschritte zunehmend in der Forschung diskutiert wird, ist jedoch, wie Lernende entsprechend ihren individuellen Bedürfnissen angemessen unterstützt werden können. Das Verständnis, wie Lernmerkmale wie Lernvoraussetzungen, Lernverhalten oder Leistung mit dem Unterstützungsbedarf zusammenhängen, ist entscheidend für eine effektive adaptive Unterstützung. Verschiedene grobkörnige und feinkörnige Ansätze können verwendet werden, um Grundlagen für die Adaption zu schaffen. Bisher wurden häufig konventionelle Produktdaten, wie beispielsweise Vorwissensdaten, verwendet, um die Auswirkungen von Scaffolding bei Lernenden mit unterschiedlichem Vorwissen zu untersuchen. Eine neuere Richtung ist die Analyse von Prozessdaten, die von Computersystemen generiert werden und dazu beitragen können, Problemlösungsprozesse und ihre Beziehung zu Aufgabenergebnissen zu verstehen. Mit Methoden wie dem maschinellen Lernen werden Prozessdaten vielversprechend für eine feinere instruktionale Anpassung in Echtzeit.

Ziel der vorliegenden Dissertation ist es, Grundlagen für eine adaptive Unterstützung beim Erlernen des kollaborativen diagnostischen Denkens zu schaffen. Frühere Studien haben gezeigt, dass agentenbasierte Simulationen, die ein hoch standardisiertes Training kollaborativer Prozesse ermöglichen, die Fähigkeit zum kollaborativen diagnostischen Denken effektiv verbessern, insbesondere wenn sie mit Kollaborationsskripts kombiniert werden, die kollaborative Prozesse zusätzlich erleichtern. Die Forschung in dieser Dissertation baut nicht nur auf diesen Ergebnissen auf, sondern erweitert sie, indem sie

Reflexionsunterstützung, die Lernende dazu anregt, über ihre eigenen Aktivitäten und Leistungen nachzudenken, als eine neue und effektive Form des Scaffolding zur Förderung des kollaborativen diagnostischen Denkens vorschlägt.

Die Dissertation umfasst drei Studien, die in derselben agentenbasierten medizinischen Simulation durchgeführt wurden. Die Teilnehmer:innen diagnostizierten in der Rolle von Internist:innen Erkrankungen bei verschiedenen Patient:innenfällen und kollaborierten dabei mit einer agentenbasierten Radiologin, um weitere Evidenz für die Fälle zu generieren. Die erste und zweite experimentelle Studie untersuchten die Bedingungen, unter denen verschiedene Arten von Scaffolding, insbesondere Reflexionsunterstützung, das Erlernen kollaborativen diagnostischen Denkens verbessern. Basierend auf a-priori-Hypothesen wurde die Wirksamkeit verschiedener Reflexionsinstruktionen untersucht, die unterschiedliche kollaborative diagnostische Aktivitäten adressieren und unterschiedliche Grade an Strukturierung bieten. In der dritten Studie wurde maschinelles Lernen eingesetzt, um die Prozesse des kollaborativen diagnostischen Denkens und ihre Beziehung zum diagnostischen Ergebnis zu analysieren.

Die erste Studie untersuchte die Auswirkungen von Reflexionsanleitungen (Reflexion individueller Aktivitäten) und Kollaborationsskripts auf das kollaborative diagnostische Denken unter Berücksichtigung des Inhalts- und Kollaborationswissens der Lernenden. Das kollaborative diagnostische Denken wurde durch die Leistung im Teilen von Evidenzen und Hypothesen (kollaborative Aktivitäten) und die diagnostische Akkuratheit und Begründung (diagnostische Ergebnisse) operationalisiert. Zusätzlich wurde der Einfluss von Reflexion und Kollaboration auf die Akkuratheit von Verdachtsdiagnosen während des Diagnoseprozesses untersucht. Während der Bearbeitung der Simulation erhielten Medizinstudierende Fragen zur individuellen Reflexion ihrer anfänglichen Verdachtsdiagnosen, Kollaborationsskripts während der Zusammenarbeit mit der Radiologin, beides oder keine Unterstützung. Die Ergebnisse zeigten, dass die Reflexion das Teilen von Hypothesen bei Lernenden mit hohem inhaltlichen Vorwissen verbesserte, während Kollaborationsskripts das Teilen von Evidenzen bei Lernenden mit niedrigem inhaltlichen Vorwissen verbesserten. Dies deutet darauf hin, dass die Reflexion das inhaltliche Wissen aktiviert und die Lernenden auf die Kollaboration vorbereitet, sofern sie über ausreichendes Vorwissen verfügen. Während weder Kollaborationsskripts noch die Reflexionsanleitung die diagnostischen Ergebnisse verbesserten, verbesserte Kollaboration allein die diagnostische Akkuratheit der Lernenden unabhängig von ihrem Vorwissen. Diese Ergebnisse können durch die Integration von

externem Wissen in den Diagnoseprozess durch die Kollaboration mit der Radiologin erklärt werden.

Die zweite Studie untersuchte die Auswirkungen von Reflexionsanleitungen (Reflexion über kollaborative Aktivitäten) unter Berücksichtigung des Kollaborationswissens der Lernenden auf das kollaborative diagnostische Denken. Die Studie verwendete die gleiche Operationalisierung des kollaborativen diagnostischen Denkens wie die vorangegangene Studie. Die Medizinstudierenden erhielten entweder eine wenig strukturierte Anleitung, das heißt keine detaillierten Fragen, eine stark strukturierte Anleitung, das heißt detaillierte Fragen zur individuellen Reflexion ihrer kollaborativen Aktivitäten, oder gar keine Unterstützung. Die Reflexion zeigte positive Effekte für Studierende mit geringem Kollaborationswissen. Die wenig strukturierte Anleitung verbesserte das Teilen von Evidenzen, die diagnostische Akkuratheit und die diagnostische Begründung, was darauf hindeutet, dass das Wissen durch die Reflexion nicht nur aktiviert, sondern auch umstrukturiert wurde. Die stark strukturierte Anleitung verbesserte nur die diagnostische Begründung, was darauf hindeutet, dass unterschiedliche Strukturierungsgrade für unterschiedliche Teilkompetenzen von unterschiedlichem Nutzen sein könnten, da den Teilkompetenzen unterschiedliche Wissensformen zugrunde liegen, die potenziell zu unterschiedlichen Kompetenzniveaus führen. Sowohl die wenig als auch die stark strukturierte Anleitung waren für Lernende mit hohem Kollaborationswissen nicht hilfreich oder sogar lernhinderlich, was darauf hindeutet, dass diese Lernenden möglicherweise eine noch weniger detaillierte Aufforderung zur Reflexion benötigen.

Die dritte Studie untersuchte, ob und wie schnell die diagnostische Akkuratheit (diagnostisches Ergebnis) auf Basis von kollaborativen diagnostischen Aktivitäten durch maschinelles Lernen vorhergesagt werden kann. Logfiles von Medizinstudierenden und Internist:innen, die in der agentenbasierten Simulation arbeiteten, wurden als kollaborative diagnostische Aktivitäten kodiert, einschließlich des Generierens und Elitzierens von Evidenzen, des Teilens von Evidenzen und Hypothesen und des Ziehens von Schlussfolgerungen. Bigramme, die die für die Aktivitäten aufgewendete Zeit und den Wechsel zwischen den Aktivitäten repräsentieren, wurden zum Training von Klassifikationsalgorithmen verwendet, um die endgültige Diagnose als richtig oder falsch vorherzusagen. Die Ergebnisse zeigten, dass eine korrekte Diagnose zuverlässiger vorhergesagt werden konnte als eine inkorrekte Diagnose und vor dem Abschluss des Falles, was darauf hindeutet, dass das Verhalten von erfolglosen Diagnostiker:innen auf verschiedenen kognitiven Fehlern beruht, während erfolgreiche Diagnostiker:innen weniger

Verhaltensvariationen aufweisen. Erfolgreiche Diagnostiker:innen verbrachten mehr Zeit mit individuellen Aktivitäten, was darauf hindeutet, dass sie eine angemessene anfängliche kognitive Repräsentation des Falles haben, während erfolglose Diagnostiker:innen mehr Zeit mit kollaborativen Aktivitäten verbrachten und zwischen individuellen und kollaborativen Aktivitäten wechselten.

Die in dieser Dissertation vorgestellten Forschungsergebnisse liefern theoretische und praktische Implikationen für die adaptive Unterstützung des Lernens kollaborativen diagnostischen Denkens in agentenbasierten Simulationen. Erstens scheint die Anleitung zur Reflexion über kollaborative Aktivitäten besonders vielversprechend für das Erlernen verschiedener Teilfähigkeiten des kollaborativen diagnostischen Denkens zu sein. Ein geringes Maß an Struktur ist wahrscheinlich lernförderlicher als ein hohes Maß an Struktur, unabhängig vom Vorwissen der Lernenden. Die Berücksichtigung des Kompetenzniveaus der Lernenden in spezifischen Teilkompetenzen über das Vorwissen hinaus erscheint vielversprechend für die Gestaltung einer effektiven Reflexionsunterstützung. Die unterschiedlichen Ergebnisse zur Wirksamkeit von Reflexion bei Lernenden mit unterschiedlichem Vorwissen in den beiden Studien verdeutlichen jedoch auch die Schwierigkeit, Reflexionseffekte zu vergleichen und zu verallgemeinern sowie die Komplexität von Reflexionsprozessen zu quantifizieren. Ein komplexes Zusammenspiel von Faktoren wie dem Inhalt der Reflexion (z. B. diagnostische Entscheidungsfindung vs. Kollaboration), dem Vorwissen und dem Kompetenzniveau der Lernenden sowie dem Grad der Strukturierung beeinflusst die Wirksamkeit der Reflexion. Zukünftige Forschung könnte sich weiter mit der objektiven Skalierung verschiedener Strukturierungsgrade in der Reflexionsunterstützung beschäftigen, um in Zukunft zuverlässige Vergleiche der Effekte zu ermöglichen. Zweitens unterstreicht die Arbeit die Bedeutung theoriebasierter Prozessdaten zur Identifizierung subtiler Unterschiede in kollaborativen diagnostischen Prozessen zwischen erfolgreichen und erfolglosen Diagnostiker:innen. Diese Ergebnisse liefern verlässliche Hinweise auf Aktivitäten in diagnostischen Fällen, bei denen Lernende Schwierigkeiten oder Fähigkeiten haben, was eine feinere und dynamischere Anpassung der instruktionalen Unterstützung ermöglicht. Dies könnte in Zukunft die Gesamteffektivität simulationsbasierten Lernens für komplexe Fähigkeiten wie kollaboratives diagnostisches Denken verbessern.

## Table of Contents

Acknowledgments .....	i
Extended Summary .....	iii
Zusammenfassung .....	vi
Table of Contents .....	x
<b>1 General Introduction .....</b>	<b>1</b>
1.1 Aim and Structure of the Dissertation .....	2
1.2 Collaborative Diagnostic Reasoning as a Complex Skill Set .....	5
1.2.1 Diagnostic Reasoning .....	6
1.2.2 Collaborative Diagnostic Reasoning .....	9
1.3 Instructional Support for Learning Collaborative Diagnostic Reasoning .....	12
1.3.1 Simulation-Based Learning Environments .....	13
1.3.2 Scaffolding .....	15
1.3.3 Adaptivity in Simulation-Based Learning .....	24
1.4 Cumulative Dissertation .....	28
1.4.1 Agent-Based Simulation .....	29
1.4.2 Outline of Study 1 .....	30
1.4.3 Outline of Study 2 .....	32
1.4.4 Outline of Study 3 .....	33
<b>2 Study 1: Fostering Collaboration in Simulations: How Advanced Learners Benefit from Collaboration Scripts and Reflection .....</b>	<b>35</b>
<b>3 Study 2: Reflection on Collaborative Action: Fostering Collaborative Diagnostic Reasoning in an Agent-Based Medical Simulation .....</b>	<b>47</b>
<b>4 Study 3: Who is on the Right Track? Behavior-Based Prediction of Diagnostic Success in a Collaborative Diagnostic Reasoning Simulation .....</b>	<b>52</b>
<b>5 General Discussion .....</b>	<b>77</b>
5.1 Summary and Interpretation of Central Results .....	78
5.2 Theoretical Implications for Fostering Collaborative Diagnostic Reasoning Through Reflection in Agent-Based Simulations .....	83
5.3 Theoretical Implications for Adaptive Simulation-Based Learning of Collaborative Diagnostic Reasoning Using Process Data .....	88
5.4 Limitations .....	92
5.5 Transferability to Other Fields and Contexts .....	95
5.6 Practical Implications .....	96
5.7 Directions for Future Research .....	98
<b>6 Conclusion .....</b>	<b>101</b>
<b>7 References .....</b>	<b>105</b>



---

<b>8</b>	<b>Appendices.....</b>	<b>130</b>
	Appendix A: Case Material.....	132
	Appendix B: Coding Schemes .....	138
	Coding Manual for Diagnostic Outcomes.....	138
	Metrics and Sample Solutions for Collaborative Diagnostic Activities .....	139
	Appendix C: Knowledge Tests .....	144
	Content Knowledge.....	144
	Collaboration Knowledge .....	146
	Appendix D: Reflection Guidance .....	148
	Reflection on Individual Activities (Study 1) .....	148
	Reflection on Collaborative Activities (Study 2).....	148
	Appendix E: Collaboration Script (Study 1).....	152
	Appendix F: Additional Graphs for the Inferential Statistics in Study 2.....	155
	<b>Statement of Scientific Integrity .....</b>	<b>158</b>

# 1 General Introduction

*Constanze Catharina Richters*

## 1.1 Aim and Structure of the Dissertation

To successfully navigate the complexities of the 21<sup>st</sup> century, individuals require a range of complex skills simultaneously (Dede, 2009; Partnership for 21st Century Skills, 2009). These skills include, among others, problem solving (Fiore et al., 2017; Graesser et al., 2018; Liu et al., 2015; Roschelle & Teasley, 1995; Rummel & Spada, 2005), (scientific) reasoning (F. Fischer et al., 2014), reflection (Saleh, 2019), and collaboration (Griffin & Care, 2015; Van Laar et al., 2017). The interconnectedness of these skills is particularly evident in professional practice, where practitioners need not only to be able to engage in and reflect on cognitive activities—such as problem identification, asking questions, evaluating and generating evidence and hypotheses, and drawing conclusions—but also to be able to collaborate effectively with diverse others in this process. Collaboration can thereby take many forms, including collaboration between professionals from different disciplines or established areas of expertise, such as teachers from different subjects or physicians from different specialties. Physician collaboration is particularly important when global challenges are considered, such as the COVID-19 pandemic, where finding an appropriate solution is likely to be nearly impossible without the collaboration of experts from different areas. When participants make constructive and substantive contributions, collaboration offers several advantages over individual practice, including the integration of the different knowledge sources, skills, and perspectives, leading to better learning opportunities and problem-solving outcomes (Chi & Wylie, 2014; Graesser et al., 2018; Kirschner et al., 2018). However, collaboration is also inherently complex and challenging, requiring individuals to engage with different perspectives, negotiate conflicting ideas, and effectively coordinate their efforts (Hesse et al., 2015; Liu et al., 2015). Thus, educational and psychological research has increasingly focused on promoting the development of these skills in higher education in various contexts.

The present dissertation focuses on a form of collaboration in which the aforementioned skills of scientific reasoning, problem solving, and collaboration are highly interconnected and which often occurs in professional practice: *collaborative diagnostic reasoning* (e.g., Abele, 2018; Radkowitz et al., 2022). To name just a few examples, teachers from different subject areas (e.g., physics and biology) collaboratively diagnose students' skill levels (Pickal et al., 2022); physicians from different subspecialties (e.g., gynecology and oncology) collaboratively diagnose endometrial cancer (Emons et al., 2018); and mechatronics experts with different roles and tasks collaboratively diagnose faults in automotive systems (Abele, 2018). Collaborative diagnostic reasoning can be defined as a

coordinated process of diagnosing a malfunction in a system with at least one other diagnostician, involving several key actions, such as *generating*, *evaluating*, *sharing*, *eliciting*, and *negotiating evidence and hypotheses* (Radkowsch et al., 2022). The ultimate goal of diagnostic reasoning is to reduce diagnostic uncertainty to thereby facilitate the achievement of an accurate diagnosis in order to take appropriate action (Heitzmann et al., 2019). Previous research has shown that students and even practitioners struggle with sharing processes while engaging in collaborative diagnostics (Tschan et al., 2009).

In recent years, *simulation-based learning* has become a widely used instructional approach for fostering the development of complex skills such as collaborative diagnostic reasoning. Previous empirical research has provided robust meta-analytic evidence that simulations are appropriate for facilitating the learning of complex skills critical to the development of professional expertise (Chernikova, Heitzmann, Stadler, et al., 2020). Simulations reduce the complexity of real-world requirements (Gegenfurtner et al., 2014) while providing opportunities for the repetitive, deliberate practice of targeted subskills (Ericsson, 2004). *Agent-based simulations*, in which humans collaborate with computer agents, are particularly useful for standardizing specific collaborative processes (e.g., information sharing) and allowing learners to practice these processes in a targeted manner (Radkowsch, F. Fischer, et al., 2020). Furthermore, additional support, such as *scaffolding* (Belland et al., 2017; Wood et al., 1976), which is support provided to learners by more knowledgeable humans or computer systems to help them complete tasks or solve problems that would otherwise be challenging, can thereby increase the effectiveness of simulation-based learning (Chernikova, Heitzmann, Stadler, et al., 2020).

However, as learners differ in their learning prerequisites, such as prior knowledge or cognitive abilities, a current topic of much debate in learning and instructional research is the extent to which instructional support such as simulations or scaffolding should be adapted to learners' needs in order to increase its effectiveness (Belland et al., 2017; Chernikova, Heitzmann, Stadler, et al., 2020; F. Fischer et al., 2022; Plass & Pawar, 2020). Recent technological developments, such as AI-based generative systems (e.g., Chat GPT), have further intensified these ongoing discussions. Researchers have argued that *adaptive instructional support* is particularly promising for self-regulated learning (Azevedo & Hadwin, 2005; Munshi et al., 2023; Pea, 2004; Plass & Pawar, 2020). Major challenges in the appropriate implementation of adaptive instructional support involve how to decide *what* to adapt and *how* to adapt it (Plass & Pawar, 2020). To answer these questions, instructional designers and educators need a solid understanding of which learner characteristics (e.g., prior

knowledge, learner behavior, or learner performance) are associated with which needs for support (Plass & Pawar, 2020; Tetzlaff et al., 2021, 2023). For instance, the effectiveness of scaffolding has been found to be influenced by learners' prior knowledge, with different levels of guidance benefiting learners with different skill levels (Chernikova, Heitzmann, Stadler, et al., 2020; Kalyuga, 2007; Simonsmeier et al., 2021; Snow, 1978, 1991). Such robust empirical evidence seems to serve as a solid basis for adapting instructional support. Furthermore, besides product data (e.g., prior knowledge), process data that are collected during the learning process may offer insights into subtle variations in learners' problem-solving approaches (Goldhammer et al., 2017; Greiff et al., 2016; Stadler et al., 2020, 2023; Tetzlaff et al., 2021). Analyzed using advanced techniques such as *machine learning* (Desmarais & Baker, 2012; Gašević et al., 2016), these data may allow for a more precise tailoring of instructional support to individual learners' needs.

To date, not much research (Pickal, Engelmann, Chinn, Girwidz, et al., 2023; Radkowsch et al., 2021) has been conducted on instructional support aimed at fostering collaborative diagnostic reasoning, let alone research addressing adaptive instructional support. Therefore, this dissertation was aimed at establishing foundations of adaptive instructional support for learning collaborative diagnostic reasoning using medical education as an exemplar. Medicine was selected as the investigative context due to the critical importance of collaborative diagnostic reasoning in high-stakes professions (Epstein & Hundert, 2002). A comprehensive understanding of diseases and optimal treatments often requires expertise from various medical specialties (Shafran et al., 2017).

The research in this dissertation combines different instructional support approaches: simulation-based learning and two types of scaffolding, namely, *reflection guidance* (Mamede & Schmidt, 2017) and *collaboration scripts* (e.g., Vogel et al., 2017). The particular focus is thereby on reflection guidance. *Reflection* refers to the process of learning from experience by returning to past events or activities and re-evaluating them in light of new knowledge (Kolb, 1984). Reflection skills are considered highly important in professional practice for promoting autonomy and self-regulation, which are central to professional growth (Cressey & Boud, 2006; Nguyen et al., 2014). Guiding learners in reflection involves providing assistance and resources to help them engage in thoughtful introspection and critical analysis of their activities, leading to enhanced learning (Coulson & Harvey, 2013). In this light, the dissertation suggests new adaptive approaches for fostering reflection in collaborative diagnostic reasoning.

The dissertation is divided into three main parts: The *first part* consists of a general theoretical introduction in which collaborative diagnostic reasoning and relevant aspects of adaptive instructional support in agent-based simulations for learning collaborative diagnostic reasoning are elaborated (see Section 1). Relevant research gaps are derived from previous research, especially on reflection guidance, and the aims of the dissertation are presented. The *second part* contains the complete manuscripts of three studies that used the same agent-based simulation developed and validated by Radkowsch et al. (2020). Study 1 investigated how the effectiveness of scaffolding, particularly reflection guidance, may vary on the basis of learners' prior knowledge (see Section 2). Building on Study 1, Study 2 delved more deeply into reflection guidance by investigating additional circumstances under which its effectiveness depends on learners' prior knowledge (see Section 3). Study 3 examined the relationships between engagement in collaborative diagnostic reasoning processes and diagnostic outcomes as well as the early prediction of diagnostic outcomes from the engagement in collaborative diagnostic reasoning processes to inform future adaptive instructional support (see Section 4). Whereas Studies 1 and 2 employed traditional experimental designs and used regression analyses to examine interaction effects, Study 3 used machine learning algorithms to predict outcomes based on process data. Through these different approaches, the studies contribute to various conceptual and methodological foundations for adaptive instructional support for learning collaborative diagnostic reasoning. In the *third part*, the main findings of the studies are summarized and their individual and joint implications are discussed with regard to the aims and research questions of the dissertation, educational practice, and transferability to other fields and contexts while considering the limitations of the studies as well as further research (see Sections 5). Finally, an overall conclusion is drawn for the dissertation (see Section 6).

## 1.2 Collaborative Diagnostic Reasoning as a Complex Skill Set

A closer look at the processes involved in collaborative diagnostic reasoning is necessary to understand it as the complex skill set that it is. The conceptualization of collaborative diagnostic reasoning in this dissertation is based on the collaborative diagnostic reasoning (CDR) model, which was recently developed by Radkowsch et al. (2022). According to the *CDR model*, collaborative diagnostic reasoning processes are defined by two types of interacting activities, namely, *individual activities*, which refer to cognitive processes related to complex problem solving (diagnostic reasoning), and *collaborative activities*, which refer to the interactions among diagnosticians (collaboration). The timing and nature of

engagement in these processes are thought to be largely determined by different types of knowledge (i.e., *content* and *collaboration knowledge*) and *cognitive* and *social skills* (Radkowsch et al., 2022). Whereas *professional knowledge* (e.g., medical knowledge; Charlin et al., 2007) and *cognitive skills* (e.g., intelligence; Stadler et al., 2015) are assumed to primarily influence individual processes, *collaboration knowledge* (T. Engelmann & Hesse, 2011; F. Fischer et al., 2013) and *social skills* (Hesse et al., 2015; Liu et al., 2015) are assumed to primarily influence collaborative processes. In this chapter, collaborative diagnostic reasoning and its factors of influence are described in more detail, whereas the related concepts of (*collaborative*) *problem solving* (e.g., Graesser et al., 2018; Liu et al., 2015; Roschelle & Teasley, 1995; Rummel & Spada, 2005), (*scientific*) *reasoning* (e.g., F. Fischer et al., 2014; Kahneman & Frederick, 2001; Norman et al., 2017; Pelaccia et al., 2011), and further research on collaboration (e.g., F. Fischer et al., 2013) are considered, in line with Radkowsch et al. (2022), who drew on and integrated a number of these research findings into their conceptualization. To this end, diagnostic reasoning is described first (see Section 1.2.1), before it is extended to collaboration (see Section 1.2.2).

### **1.2.1 Diagnostic Reasoning**

As an umbrella term for diagnostic reasoning (Radkowsch et al., 2022), *complex problem solving* describes the transition of a system from a current state to a target state that cannot be achieved through the application of routine tasks (Jonassen, 2000). When applied to diagnostic problems, this process involves the transition from identifying a problem, such as an abnormality in a body system (e.g., a bacterial infection causing symptoms), to successfully solving the problem by, for example, eliminating the abnormality (e.g., treating a bacterial infection with antibiotics to cure the patient). *Diagnosing* is related to the aspect of problem solving that is concerned with the identification of the problem (Abele, 2018). It refers broadly to the “goal-oriented collection and interpretation of case- or problem-specific information to reduce uncertainty” (Heitzmann et al., 2019, p. 4). In professional practice, the central aim of diagnosing is to achieve an accurate diagnosis, referred to as *diagnostic accuracy* (Chinn et al., 2011; Monteiro et al., 2015; Simmons, 2010). The diagnosis serves as a decision point that enables actions such as an optimal treatment plan for the patient (Daniel et al., 2019; Eva et al., 2007). Besides making an accurate diagnosis, adequately supporting that diagnosis with evidence (e.g., key clinical findings), referred to as *diagnostic justification* (Yudkowsky et al., 2015), and communicating these justifications to third parties (e.g., other collaborating professionals, students, or patients) facilitates the traceability of decisions.

Previous research has described general reasoning processes that are relevant in the context of diagnosing. For example, Klahr and Dunbar (1988) proposed the *Scientific Discovery as Dual Search (SDDS) model*, in which reasoning moves between two hypothetical problem spaces: a hypothesis space containing potential hypotheses and an experiment space for testing the hypotheses. The SDDS model emphasizes the interdependencies between different evidence- and hypothesis-based cognitive processes, such as specifying hypotheses and testing them against evidence. F. Fischer et al. (2014) further specified the processes between the spaces in the context of domain-independent knowledge generation by proposing eight *epistemic activities*. Based on selected epistemic activities, an exemplary reasoning process could look as follows: *generating evidence* to support or reject a claim; *evaluating evidence* to assess how well a particular piece of evidence supports a claim or theory; *generating hypotheses*, which refers to formulating possible answers and deriving them from plausible models, theoretical frameworks, or empirical evidence; and *drawing conclusions*, which involves weighting different pieces of evidence in accordance with the method of generation and the rules and criteria of the discipline (F. Fischer et al., 2014). For instance, when encountering a patient, physicians (e.g., internal specialists) generate differential diagnoses (hypotheses) on the basis of findings and symptoms (evidence). They weigh these hypotheses with newly gathered evidence, such as history-taking or laboratory tests, until they settle on a final suspected diagnosis (drawing conclusions). Whereas these epistemic activities are part of the diagnostic process, they are not necessarily performed in a specific order or sequence.

In addition to such approaches that are applied to describe different cognitive activities in the reasoning process, *dual-process theories* are used to describe different modes of reasoning, distinguishing between *nonanalytical* (intuitive, automated, experiential, rapid) and *analytical* (nonintuitive, deliberate, rational, slow) reasoning (Kahneman & Frederick, 2001; Norman et al., 2017; Pelaccia et al., 2011). *Nonanalytical reasoning* relies on recognizing patterns on the basis of readily available information, especially visual cues (Norman et al., 2007). Diagnosticians who use this system process only part of the information, make holistic judgments, and provide approximate responses (Norman et al., 2007). *Analytical reasoning* involves actively gathering information, applying learned rules, and demanding cognitive effort (Kahneman, 2003). Both reasoning approaches contribute to diagnostic errors, but the role of cognitive biases, especially premature case closure in nonanalytic processes, is unclear (Norman et al., 2017). In medicine, nonanalytical reasoning, also referred to as *System 1* (Kahneman & Frederick, 2001), prevails among experts in



diagnosing easy, typical cases (Charlin et al., 2007). Analytical reasoning, also referred to as *System 2* (Kahneman & Frederick, 2001), complements System 1 (Tay et al., 2016) by enhancing decision-making and action-taking (Quirk, 2006) and helping users overcome misleading information (Eva et al., 2007). Both the epistemic activities and the various modes of reasoning suggest that diagnosing is a complex, iterative, and nonlinear process in which there is no single correct path to a solution (Charlin et al., 2012).

Educational and psychological researchers have emphasized that diagnosticians' diagnostic processes differ because of different factors (Kahneman et al., 1982; Kahneman & Frederick, 2001; Pelaccia et al., 2011). First and foremost, differences (e.g., in diagnostic speed and success) depend on the quantity, structure, and organization of the diagnostician's *professional knowledge* (Boshuizen & Schmidt, 1992) and *expertise* (Goldhammer et al., 2014; Sherbino et al., 2012). The nomenclature for professional knowledge, also known as *domain knowledge* (Hetmanek et al., 2018) or *content knowledge* (Förtsch et al., 2018), is not uniform. A commonly used classification emphasizes the difference between *conceptual* and *strategic knowledge*, where *conceptual knowledge* refers to facts or “what” information and *strategic knowledge* refers to “how” information (Förtsch et al., 2018; Schmidmaier et al., 2013). In this dissertation, the term content knowledge is used to refer to conceptual and strategic knowledge to clearly distinguish it from collaboration knowledge (see Section 1.2.2).

In medicine, conceptual knowledge refers to the pathophysiological relationships underlying a disease, also known as *biomedical knowledge* (Boshuizen & Schmidt, 1992; Woods, 2007). Strategic knowledge refers to clinical knowledge about problem solving (Schmidmaier et al., 2013). The development of medical expertise is often described by the process of knowledge encapsulation (Feltovich & Barrows, 1984). Through the repeated application of complex biomedical and clinical knowledge, this knowledge becomes more and more encapsulated into simplified but efficient models, so-called *illness scripts* (Feltovich & Barrows, 1984). Illness scripts serve as cognitive representations of diseases, encompassing typical symptoms and findings derived from these encapsulated biomedical and clinical knowledge structures (Schmidt & Rikers, 2007). Enriched illness scripts comprise three components (Custers, 2015): fault (pathophysiological processes), enabling conditions (patients' characteristics and contextual factors), and consequences (signs and symptoms). Experienced physicians rely on this encapsulated but less consciously retrievable (unless necessary) knowledge of symptoms over isolated signs and pathophysiological knowledge, ultimately enhancing diagnostic efficiency (Boshuizen & Schmidt, 1992; Rikers et al., 2000).

Thus, encountering numerous clinical cases over time leads to nonanalytical reasoning (System 1), known as pattern recognition (Bowen, 2006), enabling rapid and accurate diagnoses (Charlin et al., 2007) with few diagnostic errors (Graber, 2009). During the initial patient encounter, activated illness scripts guide information gathering and its alignment with the scripts (Mamede, 2020). However, the choice between pattern recognition (System 1) and analytical reasoning processes (System 2) depends on factors such as the interaction of situation complexity and individual knowledge and skills, experience, and self-confidence (Tay et al., 2016). For instance, as experts usually use nonanalytical reasoning (System 1) to handle easy, typical cases, they are more likely to use analytical reasoning (System 2) to handle more complex and atypical cases because these more complex cases require conscious access to knowledge, as pattern recognition is ineffective when case characteristics are unknown (Charlin et al., 2007).

### **1.2.2 Collaborative Diagnostic Reasoning**

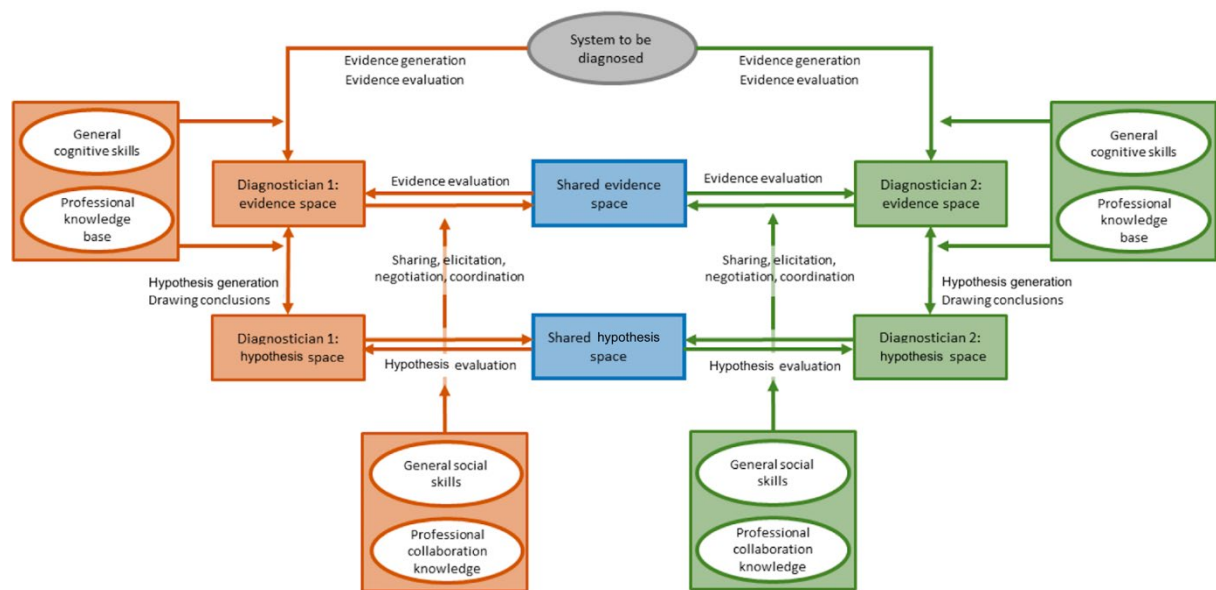
Collaboration is widespread in professional practice and important for improving diagnostic outcomes (e.g., Bosch & Mansell, 2015; Hautz et al., 2015). In medicine, it is particularly common for physicians from different subspecialties to collaboratively diagnose patients. Such collaboration is also referred to as *interdisciplinary collaboration* (Houldin et al., 2004; Mansilla et al., 2000). More precisely, physicians are not always able to get the full picture of the patient's condition through their own history-taking. Instead, they often depend on evidence generated by physicians from other subspecialties to reduce diagnostic uncertainty. Following the example from above, if an internal specialist suspects pneumonia, a patient's chest X-ray must be performed and reviewed by a radiologist to determine whether the findings are in fact consistent with pneumonia. Internal specialists rely on radiologists to provide additional information (evidence) to help them reduce diagnostic uncertainty to identify the problem that the patient is experiencing.

Consistent with previous research that viewed diagnostic reasoning as a problem-solving process, recent studies have expanded this concept to include collaborative diagnostic reasoning, conceptualizing it as a *collaborative problem-solving* process (e.g., Abele et al., 2018; Radkowsch et al., 2022). Collaborative problem solving refers to the process of working with another person or computer agent (e.g., a more specialized colleague) with the goal of finding the best solution to a problem (OECD, 2017). To describe collaborative diagnostic reasoning, Radkowsch et al. (2022) recently introduced the CDR model (see Figure 1). According to the CDR model, collaborative diagnostic reasoning refers to an

individual skill set that can be assessed at the individual level. The CDR model characterizes collaborative diagnostic reasoning as a coordinated process involving at least two diagnosticians with distinct knowledge backgrounds. Building on previous research such as the SDDS model (Klahr & Dunbar, 1988), the CDR model incorporates epistemic activities (i.e., evaluating and generating evidence and hypotheses; F. Fischer et al., 2014) and extends them by including collaborative activities from previous research on collaborative problem solving (i.e., *eliciting, sharing, negotiating, and coordinating*; Chi, 2009; Hesse et al., 2015; Liu et al., 2015; Stasser & Titus, 1985; Sun et al., 2020; Tschan et al., 2009) and by taking into account research on how knowledge and its distribution among collaborators affect collaborative processes and performance (Cannon-Bowers et al., 1993; T. Engelmann & Hesse, 2010; Stasser & Titus, 1985; Tschan et al., 2009).

**Figure 1**

*Collaborative Diagnostic Reasoning (CDR) Model*



*Note.* Boxes represent storage for outcomes of individual and collaborative processes. Ovals represent individual prerequisites for diagnostic and collaborative activities. Vertical lines represent individual diagnostic activities. Horizontal lines represent collaborative diagnostic activities. The figure was adopted from Radkowitz et al. (2022).

More precisely, the CDR model distinguishes between *collaborative diagnostic activities*, namely, *eliciting, sharing, negotiating, and coordinating evidence and hypotheses*, and *diagnostic activities*, which refer to the epistemic activities, namely, generating and evaluating evidence and hypotheses and drawing conclusions (Radkowitz et al., 2022). As

diagnostic activities precede and drive collaborative diagnostic activities in the collaborative diagnostic reasoning process (e.g., diagnosticians need to generate evidence in order to share it in the next step), in this dissertation, the term collaborative diagnostic activities refers to all activities involved in collaborative diagnostic reasoning, and *individual* and *collaborative activities* are distinguished within collaborative diagnostic activities. Throughout the diagnostic reasoning process, collaborative activities help diagnosticians construct and maintain a shared understanding of the problem (Roschelle & Teasley, 1995), and their quality is assumed to be crucial for the success of the collaboration (Radkowsch et al., 2022). Existing studies have shown that students often lack collaborative skills (e.g., Hall & Buzwell, 2013; O'Neill et al., 2013; Pauli et al., 2008), and practitioners struggle to pool (i.e., elicit and share) relevant information, as it has been observed that the content knowledge that some team members have and others do not is less likely to be shared than common knowledge (e.g., Davies et al., 2018; Tschan et al., 2009). For instance, radiologists are often not informed about previous surgeries or other conditions by the treating clinician, thus making it difficult for radiologists to interpret radiological results appropriately (Brady et al., 2012).

To engage effectively and successfully in such collaborative activities, diagnosticians need to have *collaboration knowledge*, also called *meta-knowledge* (T. Engelmann & Hesse, 2010), which refers to information about the collaborators' roles, knowledge backgrounds (i.e., how information is distributed among collaborators; Wegner, 1987), and tasks (Cannon-Bowers et al., 1993). The application of collaboration knowledge enables diagnosticians to anticipate, evaluate, and adapt to the knowledge, roles, and tasks of their counterparts (Cannon-Bowers et al., 1993). In the exemplary collaboration between radiology and internal medicine, the internist needs to know what information about the patient is relevant for the radiologist to perform the test. If the internist orders a contrast X-ray, the radiologist needs to know the patient's potential risk factors (e.g., chronic kidney disease, pregnancy) in order to assess whether the benefits of the test outweigh the risks to the patient. Thus, the amount and organization of collaboration knowledge presumably influences whether relevant information is shared and elicited (Fiore et al., 2010) and is therefore crucial for successful collaboration.

The *script theory of guidance* proposed by F. Fischer et al. (2013) provides a suitable theoretical explanation for how the amount and organization of collaboration knowledge determines the success of collaboration. F. Fischer et al. (2013) proposed that collaboration knowledge is organized into *internal collaboration scripts* that guide a person's understanding and actions during collaboration. Using the theater metaphor introduced by

Schank (1999), F. Fischer et al. (2013) defined these scripts as highly flexible configurations of knowledge components (*play, scene, role, and scriptlet*) that are related in a hierarchical way, with the play level forming the highest configuration of a collaboration script. As a person gains experience with a collaborative situation, script components develop at increasingly higher levels.

In addition to the general potential that collaboration has for improving diagnostic outcomes by reducing diagnostic uncertainty (Radkowsch et al., 2022), collaborative diagnostic reasoning is particularly relevant and advantageous over individual diagnostic reasoning in situations where the diagnostic problem cannot be solved by an individual (Graesser et al., 2018; Radkowsch, F. Fischer, et al., 2020). The importance of a collaborator (e.g., a radiologist) in the diagnostic process may be diminished if the diagnostician (e.g., the internist) has much prior knowledge (e.g., advanced illness scripts). In this case, the evidence that can be obtained from the radiologist may be less critical to the diagnostic outcome than when the diagnostician has less knowledge.

Overall, collaborative diagnostic reasoning within interdisciplinary teams is crucial for enhancing diagnostic outcomes, yet it poses significant challenges. It involves complex skills that are needed for individual activities (e.g., generating and evaluating evidence and hypotheses, drawing conclusions) and collaborative activities (e.g., eliciting and sharing evidence and hypotheses) that influence diagnostic outcomes (e.g., diagnostic accuracy and justification). The quality of collaboration and diagnostic reasoning outcomes depends on factors such as expertise, content and collaboration knowledge, and case characteristics. Furthermore, the contribution that collaboration makes to diagnostic outcomes may depend on the diagnostician's content knowledge. Study 1 in this dissertation explored whether the contribution that collaboration makes toward diagnostic outcomes depends on the content knowledge of the diagnostician (see Study 1 in Section 1.4, Research Question 3).

Based on the CDR model, it seems beneficial that successful training in collaborative diagnostic reasoning addresses both individual and collaborative activities.

### **1.3 Instructional Support for Learning Collaborative Diagnostic Reasoning**

Given the complexity of collaborative diagnostic reasoning, the challenges observed in practice in performing certain subskills (e.g., Davies et al., 2018; Tschan et al., 2009), and the importance of collaborative diagnostic reasoning in professional practice (Hautz et al., 2015), it seems necessary to help future diagnosticians (e.g., physicians and teachers) learn collaborative diagnostic reasoning. Before diving into simulation-based learning and

scaffolding as instructional support approaches for collaborative diagnostic reasoning, it is important to mention that fostering collaboration skills has gained particular attention in the context of *collaborative learning* (CL) and *computer-supported collaborative learning* (CSCL; e.g., F. Fischer et al., 2013; Vogel et al., 2017). Learners in CSCL are thought to be cognitively engaged at higher levels by interacting with each other, which enhances learning, provided they collaborate constructively (Chi & Wylie, 2014) and do not become cognitively overloaded (Kirschner et al., 2018). Although collaborative problem solving is not directly equivalent to collaborative learning, as the goal of collaboration is primarily different, successful problem solving is closely related to learning (Dillenbourg et al., 1996). More precisely, there is recent evidence that learning is an important component for the successful completion of a complex problem-solving task (Herrmann et al., 2023). Moreover, the joint consideration of both collaborative problem solving and collaborative learning has been increasingly emphasized in recent years (Tsang et al., 2019). On the basis of previous work (e.g., Radkowsch et al., 2021), it can therefore be assumed that learning theories and findings from CL and CSCL, such as *collaborative cognitive load theory* (Kirschner et al., 2018) and *collaborative inhibition* (Hood et al., 2023), can generally be applied to learning collaborative diagnostic reasoning.

In this dissertation, drawing on instructional support approaches previously used to foster collaborative diagnostic reasoning, namely, simulation-based learning (Pickal et al., 2022; Radkowsch, F. Fischer, et al., 2020) and collaboration scripts (Pickal, Engelmann, Chinn, Girwidz, et al., 2023; Radkowsch et al., 2021), reflection guidance (Mamede & Schmidt, 2017) is introduced as a new promising scaffolding approach. In the following, first, a brief conceptual introduction to simulation-based learning is given (see Section 1.3.1), followed by a more detailed section on scaffolding, including collaboration scripts and reflection guidance (see Section 1.3.2). Finally, taking into account the previous sections, aspects of adaptivity in simulation-based learning that are important for fostering collaborative diagnostic reasoning in simulations are discussed (see Section 1.3.3).

### **1.3.1 *Simulation-Based Learning Environments***

For effective knowledge restructuring and developing complex skills, such as collaborative diagnostic reasoning (Kolodner, 1992), early exposure to authentic situations is crucial (Boshuizen et al., 2020; Eva, 2005). In high-stakes fields such as medicine, simulations are particularly well suited to create such authentic situations, while offering reduced real-world complexity and risk (Gegenfurtner et al., 2014). Simulation-based learning

environments are extensively utilized in education, including scenarios such as flight simulators in pilot training (Landriscina, 2012) and patient simulations in medicine (Al-Kadi & Donnon, 2013). Referred to as *approximations of practice* (Grossman et al., 2009), simulation-based learning environments provide authentic representations of real-world scenarios in which learners (e.g., medical students) can engage with critical aspects of tasks and apply their knowledge to realistic cases in a standardized setting (Grossman et al., 2009; Siebeck et al., 2011). Allowing medical students to apply their knowledge to realistic patient cases facilitates knowledge reorganization, encapsulation, and the development of illness scripts (Feltovich & Barrows, 1984; Schmidt & Rikers, 2007). Simulation-based learning primarily targets a few subskills for repetitive, deliberate practice that is critical to developing professional expertise (Ericsson, 2004). In high-stakes professions such as medicine (Ziv et al., 2003), simulated patients offer unique advantages, such as access to rare scenarios (e.g., disruptive patient behaviors or uncommon diseases), time-outs and the exploration of failure (e.g., *productive failure*; Kapur, 2008), and systematic debriefing (Gegenfurtner et al., 2014; Grossman et al., 2009).

A more recent approach to fostering collaborative problem solving is *agent-based simulation*, in which one or more learners collaborate with a human or computer agent to solve a problem (Graesser et al., 2018; OECD, 2017). Unlike human-to-human simulations (e.g., role-playing; Gardner & Ahmed, 2014; Pickal et al., 2022; Zottmann et al., 2018), human-to-agent collaboration offers distinct advantages. By using a computer agent, certain characteristics can be held constant (e.g., the computer agent's prior knowledge), thus providing highly standardized training that gives learners the opportunity to repeatedly practice specific, exceptionally difficult subskills.

Radkowsch, F. Fischer, et al. (2020) introduced an authentic simulation-based learning environment with a standardized agent-based expert radiologist, validating its effectiveness for measuring and fostering individual skills associated with collaborative diagnostic reasoning. Acting as internists, participants diagnose diseases in a series of fictitious patient cases, actively interacting with the agent-based radiologist by repeatedly requesting and justifying radiological examinations. The simulation allows learners to deliberately and repeatedly practice the collaborative activities that have been identified as challenging for diagnosticians (i.e., information eliciting and sharing; Tschan et al., 2009).

In human-to-human collaboration, a focus on such repetitive training would create ethical and economic limitations and could potentially undermine the motivation of the human collaborator. Agent-based collaboration might also have drawbacks, such as less

authenticity, or it might decrease participant motivation because they have to interact with the computer agent. However, a wider range of interaction possibilities can actually be programmed in a human-to-agent environment (Rosen, 2015). Moreover, recent empirical studies have found no significant differences between agents and human collaborators in terms of students' overall performance in collaborative problem solving (Herborn et al., 2020; Rosen, 2015) and even higher levels of shared understanding, progress monitoring, and feedback in human-to-agent interaction (Rosen, 2015).

Overall, simulation-based learning has gained recognition as an effective approach to skill development and problem solving across various fields and contexts (e.g., Chernikova, Heitzmann, Stadler, et al., 2020; Cook et al., 2013; Gegenfurtner et al., 2014). It enables learners to solve complex problems in controlled settings. However, for early-stage learners, unsupported problem solving can overwhelm working memory (Belland et al., 2017; Kirschner et al., 2006; Renkl, 2014). In response, scaffolding provides valuable support throughout the learning process (Chernikova, Heitzmann, Fink, et al., 2020; Hmelo-Silver et al., 2007).

### 1.3.2 Scaffolding

Whereas, tasks that promote learning are ideally slightly more challenging than those learners can easily solve by themselves (Roosevelt, 2008), such challenging tasks can cognitively overwhelm early-stage learners who lack sufficient *prior knowledge* (Renkl, 2014). Prior knowledge refers to the information stored in a learner's long-term memory at the start of learning (Simonsmeier et al., 2021). *Cognitive load* broadly refers to the amount of mental effort a task requires (Sweller et al., 2011). *Intrinsic load*, which is determined by the inherent complexity of the learning task (structure and interactivity) as a result of the learner's level of prior knowledge, is essential for learning. Thus, a lack of prior knowledge can lead to cognitive overload. In addition, *extraneous load*, which is additional mental effort that results from the way information is presented or the instructional design, can hinder learning by overloading cognitive resources. To counteract cognitive overload for early-stage learners, scaffolding, the temporary support and guidance of learners—historically given by more experienced individuals, such as educators or peers—has emerged to assist learners with complex problem-solving tasks that would normally be beyond learners' independent abilities without support (Tabak & Kyza, 2018; Wood et al., 1976). Scaffolding thereby facilitates essential learning progress, an assumption rooted in Vygotsky's (1978) idea of the *zone of proximal development*. Supporting learners in their learning processes through scaffolding



promotes the development of both domain knowledge and higher order thinking skills by bridging the gap between the current skill level and the desired level (Quintana et al., 2004; Wood et al., 1976). Effective scaffolding aligns instructional support with task-specific cognitive processes, promoting a deeper understanding and keeping extraneous cognitive load as low as possible (Renkl, 2014). The ultimate goal is for learners to internalize external guidance so they can develop more self-regulated problem-solving skills (Wood et al., 1976).

Nowadays, scaffolding usually refers to instructional support that is implemented in computer-based learning environments (e.g., Belland et al., 2017; Pea, 2004; Tabak & Kyza, 2018). As such, scaffolding can be defined as the support offered while a learner works on a task; the support involves a temporary transfer of control over the learning process from the learner to a teacher or learning environment (Tabak & Kyza, 2018). To date, instructional researchers have investigated a wide range of scaffolding approaches designed to address different learning processes and enhance learning at different levels, including *worked examples* to foster cognitive processes (Paas & Van Gog, 2006; Renkl, 2014), *external collaboration scripts* to foster sociocognitive processes (F. Fischer et al., 2013; Radkowitz et al., 2021; Vogel et al., 2017), or *reflection phases* to foster (meta-)cognitive processes (e.g., Ibiapina et al., 2014; Mamede & Schmidt, 2017). Meta-analytic evidence indicates that computer-based scaffolding is generally beneficial for learning (Belland et al., 2017) and that scaffolding enhances diagnostic reasoning skills (Chernikova, Heitzmann, Fink, et al., 2020).

The interplay between prior knowledge and scaffolding has long been explored in *Aptitude-Treatment-Interaction* research (Snow, 1978, 1991). Despite several limitations in this strand of research, as it has generally yielded small effects and has rarely offered concrete guidance for instructional purposes (Driscoll, 1987; Tetzlaff et al., 2021), such research has consistently found that learners with high ability or prior knowledge require less support, whereas those with lower ability benefit significantly from more support (Jiang et al., 2018; Kalyuga, 2007). Similarly, a recent study found that learners with lower reasoning abilities benefited from more external guidance, whereas learners with higher reasoning abilities benefited from less external guidance or more self-guidance (Ziegler et al., 2021). A recently published meta-analysis (Chernikova, Heitzmann, Fink, et al., 2020) supported this finding in the context of diagnostic reasoning by showing that scaffolding types that provide substantial guidance are more effective for less advanced learners, whereas scaffolding types that emphasize higher levels of self-regulation are more effective for advanced learners. However, some types of scaffolding, such as classic prompts (i.e., hints for diagnostic problem solving as opposed to reflection prompts), have also been shown to have similar effects for learners

with low and high levels of prior knowledge. The meta-analysis by Simonsmeier et al. (2021) further complemented these findings by highlighting the stronger positive correlation between prior knowledge and learning outcomes for scaffolding with higher cognitive demands (*Matthew Effect*; Walberg & Tsai, 1983) than for scaffolding with lower demands (*Expertise-Reversal Effect*; Kalyuga et al., 2003). These results suggest that the designing of effective scaffolding is more about the variation in cognitive and self-regulatory demands than about the choice of scaffold.

In sum, over the last few decades, studies have demonstrated that prior knowledge moderates the effectiveness of scaffolding in various learning contexts (e.g., Kalyuga, 2007; Simonsmeier et al., 2021; Snow, 1978, 1991; Ziegler et al., 2021). Thus, collaboration scripts and reflection guidance seem particularly promising for helping to improve learners' collaborative diagnostic reasoning when the scaffolds are aligned with learners' prior knowledge.

### 1.3.2.1 External Collaboration Scripts

External collaboration scripts are used to structure and enhance collaborative processes in collaborative learning by guiding collaborative activities (F. Fischer et al., 2013; Vogel et al., 2017). The use of external scripts can help reduce the cognitive resources that learners need for role engagement and collaboration during collaborative activities (Nokes-Malach et al., 2015) by allowing these resources to be redirected to cognitive processes, such as knowledge restructuring or problem solving (Vogel et al., 2017). In this dissertation, the term collaboration scripts refers to external collaboration scripts. Collaboration scripts are believed to help learners construct critical components of functional scripts, particularly internal collaboration scripts (F. Fischer et al., 2013).

However, empirical evidence of the effectiveness of collaboration scripts remains mixed, as some studies have demonstrated effects on collaborative learning and collaboration quality (e.g., Dillenbourg & Hong, 2008; Noroozi et al., 2013; Rummel & Spada, 2005), whereas others have reported no effects (e.g., Rummel et al., 2009; Strauß et al., 2023). Nevertheless, meta-analytic studies have consistently shown robust medium-sized effects of scripts on collaboration skills (Radkowsch, Vogel, et al., 2020; Vogel et al., 2017) and have provided initial insights into the conditions under which scripts may prove effective, such as at a more structured level (scriptlet level) or in combination with content-specific support. In addition, scripts have been implemented to adaptively improve collaborative diagnostic reasoning skills (Radkowsch et al., 2021).

In an agent-based simulation, Radkowitz et al. (2021) examined collaboration scripts adapted to learners' performance in collaborative diagnostic activities and compared their effectiveness with static scripts. Whereas both scripts facilitated collaborative diagnostic activities with no observable differences in performance, only learners in the adaptive condition showed successful knowledge transfer to a new case, indicating that adaptivity supported the internalization of collaborative diagnostic reasoning. These findings are consistent with a recent study conducted by Strauß et al. (2023), who found no effect of static collaboration scripts on knowledge transfer. Collectively, these findings support the notion that, besides fading (Belland et al., 2017; Stegmann et al., 2011), other forms of adaptation to learners' internal collaboration scripts are also essential for knowledge transfer and the development of implicit knowledge, reinforcing the current advocacy for adaptive collaboration scripts in the literature (Kollar et al., 2018). For instance, learners with a high level of prior knowledge who have well-developed internal collaboration scripts may derive less benefit from external collaboration scripts (F. Fischer et al., 2013).

Whereas Radkowitz et al. (2021) aimed to measure learners' internal scripts, it is possible that learners made errors for other reasons and that focusing solely on errors might not adequately capture internal scripts. Moreover, given that diagnosticians must handle both individual and collaborative cognitive demands, resulting in a considerable cognitive load (Kirschner et al., 2018), supporting learners individually in improving individual and collaborative activities for collaborative diagnostic reasoning seems essential and promising. Addressing the double load, Vogel et al. (2017) showed descriptively that additional content support (e.g., reflection guidance) can increase the effectiveness of collaboration scripts by prestructuring the learning material. However, such synergistic effects of scaffolding (Tabak, 2004) on the learning of collaborative diagnostic reasoning have not been investigated so far.

Overall, collaboration scripts that are aligned with learners' prior collaboration knowledge seem promising for facilitating collaborative diagnostic reasoning. As a first step, such an approach requires insights into the effects of collaboration scripts as a function of collaboration knowledge. Moreover, as supplemental content support, reflection guidance could serve as an appropriate preparatory scaffold for the subsequent collaboration script and could enhance its effectiveness. Study 1 in this dissertation examined whether collaboration scripts could foster the learning of collaborative diagnostic reasoning as a function of collaboration knowledge and whether additional reflection support could increase the effectiveness of collaboration scripts (see Study 1 in Section 1.4, Research Questions 1 and 2).

### 1.3.2.2 Reflection Guidance

Reflection is a concept that is deeply ingrained in people's daily lives. Besides physical phenomena, such as mirrored images, the term reflection is used to refer to the complex process through which individuals deliberately engage with their personal experiences, fostering a deeper understanding of the self and enabling a shift in perspective (Boud, 2001). Reflection skills are essential in professional practice, as practitioners need to be able to work in an autonomous and self-regulated manner. As a gateway to *experiential learning* (Kolb, 1984), reflection plays a central role in both personal (Schön, 1983; Shulman, 1986) and professional development (Gustafsson & Fagerberg, 2004; Körkkö et al., 2016; Korthagen & Vasalos, 2005). In terms of professional development, reflection has the potential to promote cognitive growth by enhancing cognitive flexibility, problem-solving skills, and knowledge development (Boud, 2001; Boud et al., 2006; Lin et al., 1999; Moon, 1999). Reflection has intrigued researchers for over a century, and its roots can be traced back to the work of Dewey (1910), who defined reflection as “the active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusions to which it tends” (p. 9). This dynamic and multifaceted process involves a thoughtful examination of one's thoughts and actions with the aim of understanding, controlling, and adapting past experiences to guide future behaviors (Boud, 1985; Nguyen et al., 2014). Reflective thinking involves both cognitive and metacognitive processes, often referred to as “thinking about one's thinking” (Moon, 1999). However, reflection processes place high cognitive and motivational demands on learners and are highly unlikely to occur without instructional support, especially in novice learners who are less likely to spontaneously engage in reflection (Lin et al., 1999). In learning and instruction, various conceptual models (e.g., Boud, 1985; Hommel et al., 2023; Kolb, 1984; Moon, 1999; Schön, 1983) have been developed to describe reflection and related processes that form the basis for learner-centered approaches to fostering reflection in education and professional practice.

Reflection prompts and guidance that activate and structure reflection in the learning process (Coulson & Harvey, 2013; Van Den Boom et al., 2007) seem promising for effective learning through reflection (M. Ryan, 2013). In recent decades, a series of empirical studies testing assumptions about the nature of reflection and the effects of reflection prompts or guidance on learning and professional development have been conducted in different contexts.

For instance, Schoenfeld (1985) found that prompting reflective processes had positive effects on knowledge application, particularly in complex problem-solving tasks (e.g., Schoenfeld, 1985). Stark and Krause (2009) also found evidence of such effects for tasks that required declarative, procedural, or conditional knowledge. Lin and Lehman (1999) found that prompting students to think had positive effects on their reasoning processes and knowledge transfer. Nückles et al. (2009) found positive effects of prompts that focused on the use of cognitive and metacognitive learning strategies. Renner et al. (2016) found that detailed reflection instructions were more effective than more general instructions were. Recent results highlight that reflection prompts lead to deeper levels of reflection (Radović et al., 2021; Schellenbach-Zell et al., 2023).

In the context of diagnostic reasoning, reflection phases (Chernikova, Heitzmann, Fink, et al., 2020) with guided questions (Mamede & Schmidt, 2017) can encourage learners to think about the goals of a procedure, their performance, and the next useful steps (Heitzmann et al., 2019). In this way, learners can generate their own feedback (Nicol & Macfarlane-Dick, 2006) and options for action (Heitzmann et al., 2019).

Furthermore, in medical education, reflection guidance that structures reflection on diagnostic activities has been investigated with respect to its impact on individual diagnostic reasoning (Mamede & Schmidt, 2017) but has shown mixed effects. For instance, Ibiapina et al. (2014) found that medical students benefited more from reflecting on cases with additional content support or from studying examples of reflection than from reflection questions without additional content support. Mamede et al. (2014) found that students who were encouraged to reflect on their initial diagnosis by focusing on comparing signs, symptoms, and findings with activated illness scripts to identify errors achieved better diagnostic accuracy 1 week later for the same and similar diseases than those without structured reflection questions (Mamede et al., 2014). Current explanations for the learning mechanisms behind such effects have focused on the activation and reorganization of prior knowledge (Mamede & Schmidt, 2022) or the restructuring of cognitive representations of clinical cases (Mamede et al., 2014). The available literature suggests that approaches aimed at reorganizing knowledge to minimize diagnostic errors have consistently yielded small but positive benefits (Norman et al., 2017).

However, Braun et al. (2019) found no differences between conditions in which learners were given structured reflection, a cognitive representation scaffold, or feedback in terms of diagnostic accuracy and efficiency in immediate and delayed assessment. The authors concluded that the effectiveness of structured reflection likely depends on learner and

case characteristics (e.g., prior knowledge and case complexity) but that knowledge about how to adapt structured reflection to these circumstances is lacking. Furthermore, Fink et al. (2021) demonstrated that, unlike in problem-centered instruction (e.g., Ibiapina et al., 2014; Mamede et al., 2014), students did not achieve higher diagnostic accuracy when given structured reflection questions while diagnosing the diseases of virtual patient cases in a simulation-based learning environment, due to the use of serial cue cases. These cases, which are interactively constructed rather than presented all at once, may provide more space for implicit reflection, thus obviating the need for additional reflection instruction.

In sum, adapting the structure of reflection guidance to the reflection content, goals, timeframe, and prior knowledge of learners is a significant challenge. Specifically, the challenge lies in adapting the approach so that it will increase learner motivation, cognitive engagement, and processing without cognitively overwhelming learners, especially those in earlier stages of learning (Sweller, 2005).

Whereas reflection has long been largely conceptualized as an individual process, reflection in the collaborative context and its facilitation has received considerable attention in the last 2 decades (e.g., Cressey & Boud, 2006; Davis, 2000; Suthers, 2000). Building on older and newer perspectives on reflection to include the process (e.g., Korthagen & Vasalos, 2005), critical (e.g., Mezirow, 1990), and social (e.g., Prilla et al., 2013, 2020; Renner et al., 2016) perspectives, Hommel et al. (2023) recently proposed a comprehensive reflection model. The *process perspective* considers various triggers and influences on reflection, leading to specific outcomes. The *critical perspective* involves questioning the assumptions and content of problem solving and evaluating their consistency with the learner's knowledge, understanding, and beliefs in the current circumstances (*critical reflection*; Mezirow, 1998). This type of reflection is content-dependent and cognitively demanding (Hommel et al., 2023; Kmiecziak, 2020). The *social perspective* recognizes that reflection can take place in social contexts. Reflecting on diagnostic reasoning (e.g., Mamede et al., 2014) is aligned with the critical perspective, whereas reflecting in the context of collaborative diagnostic reasoning is aligned with both the critical and social perspectives.

Reflection in the social context has primarily been studied as *collaborative reflection* (Prilla et al., 2013), which includes joint reflective activities and is closely related to the notion of collaborative problem solving (Roschelle & Teasley, 1995). Collaborative reflection describes the process of learning together from shared experiences by articulating and sharing experiences, evaluating them together, and gaining valuable insights (e.g., Csanadi et al., 2021; Darmawansah et al., 2022; Fleck & Fitzpatrick, 2010; Foong et al., 2018; Kim et al.,

2011; Krogstie et al., 2013; Lin et al., 1999; Prilla et al., 2013, 2020; Radović et al., 2023; Strauß et al., 2023; Zhang et al., 2023). However, in the context of collaborative problem solving (e.g., collaborative diagnostic reasoning), research on the effects of reflection guidance on the quality of collaboration or the development of collaboration skills is scarce.

One study by Strauß et al. (2023) addressed this gap and found that reflection guidance did not have a stronger effect on the quality of collaboration compared with external collaboration scripts (e.g., F. Fischer et al., 2013). However, they found that reflection did have a stronger effect on the explicit knowledge of beneficial interactions during collaboration than collaboration scripts. The authors hypothesized that reflection triggers deeper levels of cognitive processing of information (Chi & Wylie, 2014) about successful collaboration than following a script, which was feared to limit learner autonomy (Wise & Schwarz, 2017). However, other studies, such as Radkowsch, Vogel, et al. (2020) and Radkowsch et al. (2021), have presented counterevidence to this criticism of collaboration scripts.

Although Strauß et al. (2023) examined collaborative reflective activities, these findings seem promising for individual reflection on collaborative activities as well. The potential of reflection guidance may even go beyond what Strauß et al. (2023) found in their study: namely, that reflection guidance, because it inherently structures collaboration to a lower extent but at the same time can be structured more flexibly than external collaboration scripts, may be similar (for learners with low levels of prior knowledge) or even more beneficial (for learners with high levels of prior knowledge) for enhancing the quality of collaboration and learning when aligned with learner's prior collaboration knowledge. In line with the meta-analytic evidence (Chernikova, Heitzmann, Fink, et al., 2020; Simonsmeier et al., 2021 see Section 1.2.2), reflection guidance adapted to the appropriate level of the learner's internal collaboration script may potentially offer a way to vary the structure of reflection guidance so that learners with different levels of prior knowledge can benefit.

In sum, when adapted to learners' internal collaboration scripts, individual reflection on collaborative activities may be a promising way to improve the quality of collaboration and diagnostic outcomes and to promote the learning of collaborative diagnostic reasoning by promoting learners' deep cognitive engagement without overwhelming them (Sweller, 2005).

Overall, the empirical evidence points to the potential that reflection guidance holds for the learning of collaborative diagnostic reasoning. In contrast to concerns that have been raised against collaboration scripts about how they might limit learner autonomy, which is a critical component of effective professional practice (cf. Radkowsch, Vogel, et al., 2020;

Wise & Schwarz, 2017), reflection guidance appears to be a more flexible and autonomy-supportive strategy (Nguyen et al., 2014; Strauß et al., 2023). However, there are also unanswered research questions about inconsistencies in the effects on individual diagnostic reasoning and a lack of research on the effects on collaborative diagnostic reasoning.

*First*, to understand the inconsistencies that have been reported, knowledge about the conditions (e.g., a certain level of prior knowledge on the diagnostic content) that optimize the effectiveness of reflection is necessary. The meta-analysis by Chernikova, Heitzmann, Fink, et al. (2020) suggested that learners with high levels of prior knowledge are particularly likely to benefit.

*Second*, there is a lack of empirical understanding of the suspected mechanisms behind reflection effects (i.e., knowledge reorganization; Mamede & Schmidt, 2022).

*Third*, there is a lack of research on guiding individual reflection on individual activities in collaborative diagnostic reasoning. As mentioned in Section 1.1.2, the knowledge and skills of each collaborator determine the quality of the exchange of relevant information and are therefore crucial for a common understanding of the diagnostic situation (Radkowsch et al., 2022). For example, for the internist to communicate critical information to the radiologist, the internist must first generate sufficient evidence (individual activities) to distinguish between relevant and irrelevant information for their partner (collaborative activities). Thus, reflecting on individual activities as a preparatory activity for collaboration seems promising overall (Vogel et al., 2017).

*Fourth*, there is a lack of empirical research on individual reflection on collaborative activities in collaborative diagnostic reasoning. Guidance for reflection on collaborative action seems promising for learning (Strauß et al., 2023) and even more promising when adapted to learners' prior collaboration knowledge (Chernikova, Heitzmann, Fink, et al., 2020) or internal collaboration scripts (F. Fischer et al., 2013). An open research question is whether the level of guidance in reflection can be varied (e.g., to a lower or higher degree of structure) to benefit learners with low and high levels of prior knowledge. Such empirical evidence could be used to adapt reflection guidance before learning to learners' prior knowledge. Study 1 in this dissertation addressed the first, second, and third points by examining the effects of individual reflection on individual activities in the context of collaborative diagnostic reasoning as a function of prior content knowledge and by further exploring the reflection process (see Study 1 in Section 1.4, Research Questions 1, 2, and 3). Study 2 addressed the fourth point by examining the effects of low- and high-structured guidance for individual reflection on collaborative activities in the context of collaborative



diagnostic reasoning as a function of prior collaboration knowledge (see Study 2 in Section 1.4).

### ***1.3.3 Adaptivity in Simulation-Based Learning***

*Adaptivity* refers to the ability of systems to dynamically adjust their behavior and responses to meet users' specific needs and preferences (G. Fischer, 2001; Peng et al., 2019). In the context of technology-enhanced learning, adaptivity is broadly referred to as the adjustment implemented by a computer to meet learners' individual needs (G. Fischer, 2001; Peng et al., 2019). In this sense, adaptivity can be defined as the "ability of a learning system to diagnose a range of learner variables, and to accommodate a learner's specific needs by making appropriate adjustments to the learner's experience with the goal of enhancing learning outcomes" (Plass & Pawar, 2020, p. 276). Adaptivity can be implemented in simulation-based learning environments in a variety of ways, including adaptive tasks or scaffolding (F. Fischer et al., 2022) as well as adaptive feedback (e.g., Sailer et al., 2023). Adaptive scaffolding refers to adjusting the level and type of support provided to the learner on the basis of their individual needs and abilities (Azevedo & Hadwin, 2005), such as reflection guidance with low or high degrees of structure, provided in accordance with learners' prior knowledge (see Section 1.3.2.2). It has been argued that the provision of adaptive scaffolding in simulations enables learners to progress more efficiently and effectively than with nonadaptive scaffolding (Plass & Pawar, 2020). However, a meta-analysis by Belland et al. (2017) comparing the effects of adaptive and nonadaptive scaffolding in STEM found no significant differences. The authors attributed these findings also to the lack of studies on specific adaptive scaffolding strategies included in the analysis and called for more research in this area. The lack of such studies may be at least partly due to a lack of knowledge about which scaffolding strategies are most effective for which learners.

The effective implementation of adaptivity requires robust evidence on so-called *learner variables* (Plass & Pawar, 2020), such as prior knowledge related to learning outcomes and needs for scaffolding, which is currently still lacking for many other learner variables in different contexts (Plass & Pawar, 2020; Tetzlaff et al., 2021). The relationship between prior knowledge and the effectiveness of scaffolding (e.g., Chernikova, Heitzmann, Fink, et al., 2020; Kalyuga, 2007; Simonsmeier et al., 2021) has recently been replicated in simulation-based learning studies that have emphasized the critical role of adapting scaffolding to learners' current knowledge when they learn with simulations (Chernikova, Heitzmann, Stadler, et al., 2020). Substantial external guidance, such as offered by worked

examples, benefits learners with low levels of prior knowledge, whereas lower external guidance, as in reflection phases, benefits learners with high levels of prior knowledge (Chernikova, Heitzmann, Stadler, et al., 2020). Thus, using prior knowledge as a basis for selecting differentially structured reflection guidance prior to simulation-based learning seems promising.

An optimal matching of scaffolding to learners' prior knowledge before learning can be classified as adaptivity at a so-called *macro level* (Plass & Pawar, 2020; Tetzlaff et al., 2021), which refers to adapting tasks or learner support between different simulations (F. Fischer et al., 2022). However, such macro-level adaptivity does not allow for dynamic adaptation to learners' evolving needs, which might be particularly promising for learning. For instance, *intelligent tutoring systems* that dynamically adapt to learners' prerequisites have shown robust effectiveness across learning contexts (Ma et al., 2014; Steenbergen-Hu & Cooper, 2014; VanLehn, 2011). Furthermore, a study by Sailer et al. (2023) recently showed that, compared with static feedback, dynamic feedback that adapts to learners' diagnostic explanations in real time between cases can improve the quality of diagnostic justification in simulation-based learning. Such an adaptation can be classified as adaptivity at a so-called *meso level* (Tetzlaff et al., 2021), which refers to adaptations that take place between different practice representations (e.g., cases) within a simulation (F. Fischer et al., 2022). However, meso-level adaptivity is based on the assumption of a linearly increasing learning process across cases (learners progress steadily and predictably from one case to the next). Such a progression does not necessarily correspond to the real learning process, as learners may encounter unexpected challenges or need to revisit earlier concepts. To address this issue, so-called *micro-level* adaptivity (Plass & Pawar, 2020; Tetzlaff et al., 2021) may be more promising. This adaptivity refers to adapting tasks or learner support within a case in a simulation (F. Fischer et al., 2022). An example of microadaptivity is the aforementioned study by Radkowitz et al. (2021), which examined the effects of adaptive collaboration scripts that were presented to the learner on the basis of their performance on each case that they worked on during the learning phase in the simulation. This kind of adaptivity could involve the real-time monitoring of changes in learners' behavior to inform the immediate adjustments of tasks, scaffolding, or feedback. Moreover, relying solely on product data, such as *learner products* (Gašević et al., 2015; e.g., the quality of diagnostic justification), when adapting to learners' evolving needs is limited because it cannot provide insights into the finer process differences that may be critical for understanding learning outcomes and identifying learners' needs (Goldhammer et al., 2017). Process data in the form of log files representing

learner behavior such as engagement in collaborative diagnostic activities hold promise for such micro-level adaptivity (e.g., Goldhammer et al., 2017; Greiff et al., 2016; Stadler et al., 2020; Tetzlaff et al., 2021).

In recent years, process data have gained prominence as an important source of scientific knowledge and a base for adaptive instructional support to research fields such as *educational data mining* and *learning analytics* (Gašević et al., 2015; Plass & Pawar, 2020; Tetzlaff et al., 2021). Process data in the form of log files can automatically be stored and analyzed by computer systems in real time by using complex methods, such as *machine learning* (e.g., Desmarais & Baker, 2012; Gašević et al., 2016). Previous analyses of process data from complex problem solving have been demonstrated to identify problem-solving approaches (Griffin & Care, 2015), common misconceptions during learning (Stadler et al., 2019), learning processes (Ifenthaler et al., 2012), and changing learning prerequisites (K. Engelmann & Bannert, 2021).

A major advantage of process analysis is that it can reveal differences between learners or activities in processes that may be relevant to learners' overall learning success and their needs for scaffolding but are not apparent in pure product data such as performance scores (e.g., Goldhammer et al., 2017). More precisely, two learners may have the same score on the outcome but may differ significantly in the processes that led to this outcome. For instance, Stadler et al. (2020) showed that process data can provide information about subtle differences between activities—such as time on task or number of clicks made—for learners with the same task outcome in complex problem solving. Other research from the field has suggested that such activities may be indicative of the task outcome (Cirigliano et al., 2020; Goldhammer et al., 2017). Using machine learning, these activities can be used to predict learners' performance before they complete a task (e.g., Ulitzsch et al., 2022). Such analyses are aimed at providing process-based scaffolding or at identifying learners who are at risk of failure (Gašević et al., 2016; Leitner et al., 2017) versus learners who are on the right track with the goal of removing scaffolding before it has a negative impact on learning (Kalyuga et al., 2003) in real time. In addition, if it is possible to identify learners who are on track and to obtain additional information about specific activities, then process-based scaffolding can also be provided on the basis of these activities because there may be learners who perform well in the end but who were not successful in every aspect of the process (Stadler et al., 2020).

However, interpreting certain activities from process data in the context of learning is far from straightforward. For instance, spending a longer time on an activity can in fact be either a sign of deeper information processing or a sign of excessive information generation or

cognitive overload (Goldhammer et al., 2017). In addition, the interpretations of process activities in the context of learning may depend on the specific characteristics of a learning environment. Therefore, activities within processes should be linked with learning theories and conventional product data (e.g., learners' performances) to obtain a comprehensive understanding of these activities and generalizable and replicable findings (e.g., Gašević et al., 2015). For example, Brandl et al. (2021) showed that engagement in collaborative activities in collaborative diagnostic reasoning processes reliably predicted the diagnostic outcome. They predicted the diagnostic outcome from machine learning based on process data in the form of log files linked to theoretically derived process activities (Brandl et al., 2021).

In sum, process analysis allows scaffolding, feedback, or tasks to be adapted to the learner on the basis of specific demonstrated and theory-based activities in the learning or problem-solving process. Collaborative diagnostic reasoning processes involving individual and collaborative activities appear to be promising predictors of the diagnostic outcome (performance), as they strongly build on theory (Brandl et al., 2021; F. Fischer et al., 2014; Heitzmann et al., 2019; Radkowitz et al., 2022). If machine learning can be used to make reliable predictions on the basis of collaborative diagnostic activities before the case is completed, it could be implemented in the simulation-based learning environment and thereby serve as the basis for real-time adjustments to support learners' performance in diagnostic reasoning. Therefore, Study 3 presented in this dissertation investigated to what extent and how early collaborative diagnostic activities can reliably predict diagnostic accuracy as the diagnostic outcome (see Study 3 in Section 1.4, Research Questions 1 and 2).

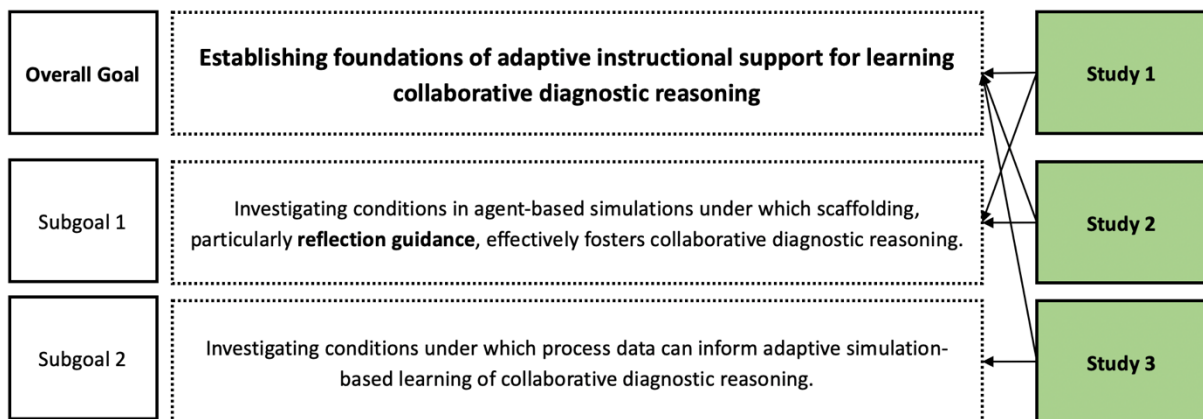
Overall, scaffolding such as collaboration scripts and reflection guidance adapted to learners' prior knowledge at the macro level appears to offer a promising approach for learning collaborative diagnostic reasoning with agent-based simulations. Furthermore, analyzing collaborative diagnostic reasoning processes with machine learning seems promising for deriving performance indicators to inform microadaptive simulation-based learning in the future.

## 1.4 Cumulative Dissertation

The overarching goal of this dissertation was to establish foundations of adaptive instructional support for learning collaborative diagnostic reasoning (see Figure 2). First, the conditions in agent-based simulations under which scaffolding, particularly reflection guidance, are effective for fostering collaborative diagnostic reasoning were investigated (Subgoal 1). Building on the research gaps outlined in the general introduction of this dissertation (see Section 1.3.2.2), the application of reflection guidance in the context of collaborative diagnostic reasoning has remained particularly unexplored and therefore formed the main scaffolding focus of the dissertation. Guidance for reflection is a potentially effective scaffold for enhancing not only collaborative diagnostic reasoning but also learner autonomy, which is particularly important for professional development (Nguyen et al., 2014). Second, the conditions under which process data can inform the adaptive simulation-based learning of collaborative diagnostic reasoning using machine learning were investigated (Subgoal 2). Whereas the first subgoal addressed macro-level adaptivity, the second subgoal addressed micro-level adaptivity.

**Figure 2**

*Aims of the Dissertation and Corresponding Studies*



To achieve these goals, three studies (i.e., two intervention studies and a process analysis study) that used different methodological approaches were conducted. The first intervention study (Study 1) examined the effects of guidance for reflection on individual activities and collaboration scripts as a function of learners' prior knowledge. The second intervention study (Study 2) examined the effects of low- and high-structured guidance for reflection on collaborative activities as a function of learners' prior knowledge. Thus, both intervention studies focused on the interaction effects of different types of reflection guidance and prior knowledge on different collaborative diagnostic reasoning outcomes, which form

the bases for macro-adaptive reflection guidance. The process analysis study (Study 3) examined collaborative diagnostic activities as predictors of diagnostic accuracy using machine learning. It examined the earliest point in the diagnostic process at which a reliable prediction could be used to inform microadaptive simulation-based learning in the future. Whereas this research contributes in general to the foundations of adaptive instructional support for learning collaborative diagnostic reasoning, it also contributes to the understanding of collaborative diagnostic reasoning and the mechanisms of individual reflection in collaborative diagnostic reasoning.

The studies were conducted in the agent-based simulation developed and validated by Radkowsch, F. Fischer, et al. (2020). The simulation models the collaboration between internal medicine and radiology to measure and facilitate collaborative diagnostic reasoning. It is briefly described below.

#### ***1.4.1 Agent-Based Simulation***

Collaborative diagnostic reasoning is central to many fields, such as automotive mechatronics, teaching, and medicine. In this dissertation, the medical context, more precisely the collaboration between internists and radiologists (Radkowsch et al., 2022), was chosen to investigate the foundations of adaptive scaffolding for fostering collaborative diagnostic reasoning and related skills. Previous research has shown that these skills are particularly important in medicine, yet medical students and practitioners lack them (Tschan et al., 2009). The proposed CDR model (Radkowsch et al., 2022) was used as the basis for the collaborative diagnostic processes. According to the authors, the generic process activities in this model have the potential to be applied outside of medicine, making it a universal framework. Therefore, medicine should be considered here only as an application context in which these activities are particularly visible. The simulation-based learning environment (Radkowsch, F. Fischer, et al., 2020) implemented on the CASUS learning platform ([www.instruct.eu](http://www.instruct.eu)) models a situation in a hospital emergency department in which two physicians from different specialties, an internist and a radiologist, collaboratively generate evidence for a patient case. Learners in the role of an internist interact with an agent-based expert radiologist to request radiological examinations (evidence elicitation) in order to obtain additional information about the patient (evidence generation), with the goal of reducing uncertainty about the final diagnosis. Participants work on several fictitious but realistic patient cases in which the presenting symptom is fever. The cases are structured as follows:

A participant begins by reviewing an electronic health record presented as a digital folder (evidence generation) containing patient admission details, emergency medical service protocol, medical history, and laboratory results. The participant then completes a radiological request form by selecting a test (e.g., MRI; evidence elicitation) and justifying their request with evidence (evidence sharing) and their suspected diagnosis (hypothesis sharing). On the basis of the relevance of the shared evidence and hypothesis, the agent-based radiologist decides whether to perform the test. If the request is rejected due to insufficient justification, the radiologist prompts the participant to revise the request form or proceed with the case. If the request is approved, detailed documentation of the requested radiological evidence is provided (evidence generation). Upon completing the case, the student submits their final suspected diagnosis (drawing conclusions), accompanied by individual and collaborative evidence rationales (diagnostic justification). Although there is no time limit on case completion, prompts suggest moving on after 15 min per case.

In each study, the simulation-based learning environment consisted of five such patient cases. The cases were typically arranged for the studies so that there was one pretest case, three learning cases, and one posttest case. The test cases differed from the learning cases in the number of examinations (request forms) that could be requested. Participants could request three examinations in test cases and 10 in learning cases. A full description of the case material can be found in Appendix A.

#### ***1.4.2 Outline of Study 1***

With respect to identifying conditions under which reflection guidance and collaboration scripts are effective for learning collaborative diagnostic reasoning, Study 1 examined the effects and mechanisms of guidance for reflection on individual activities and collaboration scripts as a function of learners' prior content and collaboration knowledge. Whereas reflection guidance in simulations (Chernikova, Heitzmann, Stadler, et al., 2020) is promising for fostering collaborative diagnostic reasoning (Chernikova, Heitzmann, Fink, et al., 2020), previous research has shown mixed results with respect to its effectiveness for individual diagnostic reasoning (cf. Fink et al., 2021; Ibiapina et al., 2014), leaving underlying mechanisms largely unexplored (Mamede et al., 2014; Mamede & Schmidt, 2022). In collaborative contexts, collaboration scripts have been shown to be effective for improving students' collaborative diagnostic reasoning (Radkowsch et al., 2021). However, there is a gap in understanding the potential combined effects of structured reflection and collaboration scripts, including possible synergistic benefits for learning collaborative

diagnostic reasoning (Tabak, 2004; Vogel et al., 2017). Additionally, it remains unclear how these types of scaffolding, individually and in combination, depend on learners' prior knowledge, which has been shown to influence the effectiveness of scaffolding for learning diagnostic skills (Chernikova, Heitzmann, Fink, et al., 2020). To address these gaps, the study used a 2x2-factorial design with  $N = 151$  advanced medical students from the fourth academic year and higher of a 6-year medical study program randomly distributed into four groups (reflection guidance, collaboration scripts, reflection and collaboration scripts, no scaffolding). This study adopted the structured reflection developed by Mamede et al. (2014). The reflection questions can be found in Appendix D, Subsection D1. The collaboration script was adopted from Radkowsch et al. (2021) and can be found in Appendix E. The study aimed to provide valuable insights into the learning of collaborative diagnostic reasoning and foundations for macro-adaptive scaffolding through reflection guidance and collaboration scripts. The following research questions and hypotheses were addressed:

RQ1: Can structured reflection and collaboration scripts in a medical simulation foster the learning of collaborative diagnostic reasoning by improving learners' performance in collaborative diagnostic activities (evidence sharing, hypothesis sharing) and learners' diagnostic outcomes (diagnostic accuracy, diagnostic justification)? We hypothesized that structured reflection (H1.1) and collaboration scripts (H1.2) would have positive individual effects and a synergistic (positive interaction) effect (H1.3) on collaborative diagnostic reasoning.

RQ2: What is the moderating effect of learners' prior knowledge with respect to the effects of structured reflection and collaboration scripts on collaborative diagnostic reasoning (evidence sharing, hypothesis sharing, diagnostic accuracy, diagnostic justification)? We hypothesized that learners with a high level of content knowledge would benefit more from structured reflection (H2.1), whereas learners with a low level of collaboration knowledge would benefit more from collaboration scripts (H2.2). We additionally hypothesized that the synergistic effect of structured reflection and collaboration scripts would depend on learners' levels of content knowledge (H2.3a) and collaboration knowledge (H2.3b).

In addition, other process-related exploratory research questions were investigated:

RQ3: To what extent does the accuracy of suspected diagnoses change during structured reflection as a function of prior content knowledge?

RQ4: To what extent does collaboration contribute to diagnostic accuracy as a function of prior content knowledge?



By investigating the extent to which reflection changes the initial suspected diagnosis (i.e., indicator of initial case representation) as a function of learners' prior content knowledge, RQ3 aimed to identify potential mechanisms behind reflection effects (i.e., knowledge activation or reorganization; Mamede et al., 2014; Mamede & Schmidt, 2022). By investigating the extent to which the contribution of collaboration to diagnostic outcomes depends on the content knowledge of the diagnostician, RQ4 aimed to investigate the extent to which the benefits of collaboration for diagnostic outcomes depend on learners' prior content knowledge. For example, when an internist has advanced illness scripts, collaboration may contribute less to diagnostic accuracy or may even be detrimental compared with when an internist has less advanced scripts (Kirschner et al., 2018).

### 1.4.3 *Outline of Study 2*

To focus more specifically on conditions under which reflection guidance is effective for adapting reflection at a macro level, Study 2 built on Study 1 but focused exclusively on reflection guidance, namely, on reflection that addressed collaborative activities. The meta-analysis by Chernikova, Heitzmann, Fink, et al. (2020) asked to what extent reflection can be more or less structured in order to be suitable for learners with low and high amounts of prior knowledge. Thus, Study 2 examined the effects of differentially structured reflection guidance as a function of learners' prior collaboration knowledge. Two variants of reflection guidance (low and high levels of structure) were developed on the basis of F. Fischer et al.'s (2013) *script theory of guidance*. More specifically, the actions in the collaboration with the radiologist were preassigned to the internal collaboration script components. Each collaborative diagnostic activity was defined as a *scene* and the subactivities that occurred within a scene as *scriptlets*. Learners in the reflection condition with a low level of structure received scene-level questions with information about which scene to reflect on, whereas learners in the reflection condition with a high level of structure received the same information about which scene to reflect on but with questions broken down to the scriptlet level. A detailed explanation of the reflection guidance with low and high levels of structure can be found in Appendix D, Subsection D2. It was assumed that learners with low levels of prior knowledge would benefit from the high structure, whereas learners with high levels of prior knowledge would benefit from the low structure. The study used a one-factorial design with  $N = 195$  advanced medical students between the third and sixth academic years of a 6-year medical study program randomly distributed to one of three groups (low-structured

reflection, high-structured reflection, no scaffolding). The following research question was addressed in this study:

RQ: Depending on prior knowledge, to what extent can low- and high-structured reflection offer support that stimulates learners to reflect on their collaborative activities and fosters the learning of collaborative diagnostic reasoning by improving learners' performance in collaborative diagnostic activities (evidence sharing, hypothesis sharing) and learners' diagnostic outcomes (diagnostic accuracy, diagnostic justification) in an agent-based medical simulation? We hypothesized that learners with a low level of collaboration knowledge would benefit from high-structured reflection, whereas for learners with a high level of collaboration knowledge, low-structured reflection would be sufficient.

The coding schemes for the performance in collaborative diagnostic activities and the quality of diagnostic outcomes for Studies 1 and 2 can be found in Appendix B. Additionally, the prior knowledge tests can be found in Appendix C.

#### **1.4.4 Outline of Study 3**

To address microadaptivity, Study 3 investigated learners' engagement in collaborative diagnostic activities on the basis of process data (log files) using machine learning to predict diagnostic accuracy (performance measure). Analyzing process data to infer learner behavior holds great promise for gaining insights into problem-solving approaches and needs for scaffolding in collaborative diagnostic reasoning (e.g., Stadler et al., 2020), yet there is a lack of empirical evidence on the relationship between learner behavior and scaffolding needs in collaborative diagnostic reasoning. The goals of the study were twofold. First, to provide a general and replicable approach for analyzing collaborative diagnostic reasoning processes, diagnostic accuracy was linked to broad behavioral indicators by analyzing the collaborative diagnostic activities displayed in the agent-based simulation. The goal was to investigate differences between successful and unsuccessful diagnostic reasoning processes and to determine the extent to which collaborative diagnostic activities could predict diagnostic accuracy. The second goal was to investigate how early diagnostic accuracy could be predicted from collaborative diagnostic activities on the basis of engagement in collaborative diagnostic activities to identify in an exploratory manner early starting points for effective ways to microadapt scaffolding. The following research questions were addressed:

- RQ1: To what extent can collaborative diagnostic activities predict diagnostic accuracy in a medical training simulation using machine learning classification models?
- RQ2: How early in the process of making a diagnosis can diagnostic accuracy be reliably predicted from collaborative diagnostic activities in a medical training simulation using machine learning classification models?

## 2 Study 1: Fostering Collaboration in Simulations: How Advanced Learners Benefit from Collaboration Scripts and Reflection

*Constanze Richters \* Matthias Stadler \* Anika Radkowitzsch \* Felix Behrmann \* Marc Weidenbusch \* Martin R. Fischer \* Ralf Schmidmaier \* Frank Fischer*

**Reference:** Richters, C., Stadler, M., Radkowitzsch, A., Behrmann, F., Weidenbusch, M., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2024). Fostering collaboration in simulations: How advanced learners benefit from collaboration scripts and reflection. *Learning and Instruction, 93*, 101912. <https://doi.org/10.1016/j.learninstruc.2024.101912>

© 2024 The Authors. This manuscript version is made available under the CC-BY-NC-ND 4.0 license. <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Parts of the manuscript have been previously published in the Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning - CSCL 2022. *International Society of the Learning Sciences*.

**Additional Reference:** Richters, C., Stadler, M., Radkowitzsch, A., Behrmann, F., Weidenbusch, M., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2022). Making the rich even richer? Interaction of structured reflection with prior knowledge in collaborative medical simulations. In A. Weinberger, W. Chen, D. Hernández-Leo, & B. Che (Eds.), *Proceedings of the 15<sup>th</sup> International Conference on Computer Supported Collaborative Learning – CSCL 2022* (pp. 155-162). International Society of the Learning Sciences. <https://repository.isls.org/handle/1/8270>

Learning and Instruction 93 (2024) 101912



Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: [www.elsevier.com/locate/learninstruc](http://www.elsevier.com/locate/learninstruc)

## Fostering collaboration in simulations: How advanced learners benefit from collaboration scripts and reflection

Constanze Richters<sup>a,\*</sup>, Matthias Stadler<sup>b</sup>, Anika Radkowitzsch<sup>c</sup>, Felix Behrmann<sup>b</sup>,  
Marc Weidenbusch<sup>b</sup>, Martin R. Fischer<sup>b</sup>, Ralf Schmidmaier<sup>d</sup>, Frank Fischer<sup>a</sup>

<sup>a</sup> Department of Psychology, LMU Munich, Munich, Germany

<sup>b</sup> Institute of Medical Education, LMU University Hospital, LMU Munich, Munich, Germany

<sup>c</sup> Department of Mathematics Education, IPN Leibniz Institute for Science and Mathematics Education, Kiel, Germany

<sup>d</sup> Department of Medicine IV, LMU University Hospital, LMU Munich, Munich, Germany

### ARTICLE INFO

**Keywords:**  
Simulation-based learning  
Reflection  
Collaboration (scripts)  
Medical education  
Collaborative diagnostic reasoning  
Prior knowledge

### STRUCTURED ABSTRACT

**Background:** Individual reflection and interdisciplinary collaboration can be critical for high-quality diagnostic outcomes. However, empirical findings on using instructional approaches to facilitate reflection and collaboration in collaborative diagnostic reasoning are inconclusive and limited. Previous studies on structured reflection and collaboration scripts have failed to consider learners' prior knowledge, but the benefits of different types of instructional support, which offer varying levels of external guidance, tend to differ across prior knowledge levels.

**Aims:** We aim to investigate individual and synergistic effects of structured reflection and collaboration scripts on collaborative diagnostic reasoning while considering knowledge in a simulation and to explore how individual reflection and collaborative engagement contribute to diagnostic outcomes.

**Sample:** Participants were 151 advanced medical students.

**Methods:** Participants received structured reflection, collaboration scripts, both, or no support while diagnosing fictitious patient cases with an agent-based radiologist.

**Results:** Structured reflection improved collaborative diagnostic reasoning performance for learners with extensive prior knowledge but impeded performance for learners with little prior knowledge. The opposite was found for collaboration scripts. Furthermore, learners with extensive prior knowledge benefited more from a combination of both kinds of support than learners with little prior knowledge. Whereas no main effect of instructional support on the diagnostic outcome was found, simply working with the collaborator had a positive effect.

**Conclusions:** Different types of instructional support in simulations are differentially effective for learners with little and extensive prior knowledge. Extensive knowledge is needed for effective learning through reflection. But for high-quality diagnostic outcomes in simulated collaborative settings, collaborative engagement is more important than individual reflection.

### 1. Introduction and theoretical background

In professional practice, such as medicine, individuals from different sub-specialties need to collaborate, for instance, in diagnosing potential causes of a problem, such as a patient's fever. Collaboration is essential for improving diagnostic outcomes and reducing diagnostic error rates (National Academies of Sciences, Engineering, and Medicine, 2015). Although collaborative diagnostic reasoning is common in various

professional practices and highly critical, both students and practitioners often lack the complex collaborative skills needed for success in this realm (Tschan et al., 2009). Thus, understanding collaborative diagnostic reasoning and related skills and fostering them in higher education, seems essential. Simulation-based learning offers a highly promising approach to foster collaborative diagnostic reasoning, allowing learners to apply their knowledge to realistic and practical scenarios (e.g., Siebeck et al., 2011)—a practice that is necessary for

\* Corresponding author. Leopoldstr. 13, 80802, München, Germany.  
E-mail address: [constanze.richters@psy.lmu.de](mailto:constanze.richters@psy.lmu.de) (C. Richters).

<https://doi.org/10.1016/j.learninstruc.2024.101912>

Received 17 August 2023; Received in revised form 16 February 2024; Accepted 24 March 2024  
0959-4752/© 2024 Elsevier Ltd. All rights reserved.



them to develop complex skills (Kolodner, 1992)—without the high stakes that exist in the real world (e.g., Gegenfurtner et al., 2014). When complemented by instructional support, such as interventions targeting reflection (Chernikova, Heitzmann, Stadler, et al. 2020), which have demonstrated broad effects on academic achievement (Zhai et al., 2023), simulations are particularly effective at facilitating the learning process.

However, in previous research, results on the effectiveness of reflection interventions on diagnostic reasoning have been mixed (e.g., Braun et al., 2019a; Fink et al., 2021; Mamede et al., 2014). Plus, the mechanisms underlying the effects remain largely unexplored (Mamede & Schmidt, 2022), and there is a lack of understanding of the benefits of reflection in collaborative reasoning contexts. In such collaborative contexts, collaboration scripts can help learners engage in collaborative activities (Vogel et al., 2017) and have been shown to be effective in improving students' collaborative diagnostic reasoning (Pickal et al., 2023; Radkowsch et al., 2021). While such previous research on structured reflection and collaboration scripts has separately examined the effects of these types of instructional support, there remains a gap in findings on potential combined effects, such as synergistic benefits (Tabak, 2004) for learning collaborative diagnostic reasoning. Collaboration scripts appear to be particularly conducive to learning when combined with content-specific instructional support such as structured reflection (Vogel et al., 2017). However, the extent to which the potential individual and combined effects of structured reflection and collaboration scripts depend on learners' prior knowledge is unclear. A previous meta-analysis showed that the effectiveness of instructional support for learning diagnostic skills varies for learners with low and high levels of prior knowledge, depending on the level of guidance provided (Chernikova, Heitzmann, Fink, et al. 2020). Hardly any systematic primary research has examined the effects of structured reflection and collaboration scripts on how well collaborative diagnostic reasoning can be learned with simulations, particularly in relation to prior knowledge.

Thus, this study aims to address this gap by providing valuable insights into the potential benefits of and the mechanisms underlying structured reflection and collaboration scripts in simulations that are designed to teach complex collaborative skills while taking into account learners' prior knowledge.

### 1.1. Facilitating individual diagnostic reasoning

*Individual diagnostic reasoning* as an epistemic process is based on the collection and interpretation of case-specific information to reduce uncertainty about the final diagnosis (Heitzmann et al., 2019). The epistemic activities involved in diagnostic reasoning—such as generating evidence (i.e., case-relevant information) and hypotheses, evaluating hypotheses against the background of evidence, and drawing conclusions (Fischer et al., 2014)—require knowledge and skills.

In medicine, physicians rely on *content knowledge*, which includes conceptual knowledge (i.e., biomedical knowledge about pathophysiological relationships; Boshuizen & Schmidt, 1992) and strategic knowledge that refers to the diagnostic problem-solving process. Content knowledge is necessary for applying conceptual knowledge (Stark et al., 2011), structured along so-called *illness scripts* (Feltovich & Barrows, 1984) to generate suspected and differential diagnoses on the basis of findings and symptoms. Illness scripts are cognitive representations of specific diseases with typical symptoms and findings that emerge from encapsulated structures of biomedical knowledge and are linked to clinical knowledge such as signs and symptoms (Schmidt and Rikers, 2007). As illness scripts and medical expertise increase, a learner can no longer easily access detailed biomedical knowledge, leading to a process known as pattern recognition (Bowen, 2006), which allows for quick and accurate diagnoses (Charlin et al., 2007). However, when diagnosticians (e.g., medical students) still have only a little medical expertise, such pattern recognition is prone to error. In addition, medical

students often struggle to connect evidence with hypotheses (Yudkovsky et al., 2015) and fail to adequately justify their diagnoses (Braun et al., 2019b).

One effective type of instructional support for improving medical diagnostic reasoning is structured (i.e., externally guided) reflection (Heitzmann et al., 2019; Mamede & Schmidt, 2017; Nguyen et al., 2014). Reflection involves analyzing and making judgments about what has happened and can help learners assess what they know, what they need to know, and how to fill the knowledge gap (Boud, 2001; Dewey, 1933). By encouraging medical students to reflect on initial suspected diagnoses, structured reflection potentially helps them identify flaws in their diagnostic reasoning (Mamede & Schmidt, 2017). More precisely, structured reflection encourages medical students to compare signs, symptoms, and findings with activated illness scripts, leading to potential improvements not only in diagnostic accuracy (i.e., the correctness of a final diagnosis) but also in understanding and explaining a final diagnosis, which holds promise for fostering diagnostic justification (Braun et al., 2019b; Mamede et al., 2014; Mamede & Schmidt, 2017).

Current explanations for the learning mechanisms behind the effects of structured reflection are related to the activation and reorganization of prior knowledge (Mamede & Schmidt, 2022) and the restructuring of cognitive case representations (Mamede et al., 2014). Thus, a certain level of prior knowledge may be a prerequisite for benefiting from structured reflection (Braun et al., 2019a; Chernikova, Heitzmann, Fink, et al. 2020; Mamede & Schmidt, 2022). Indicators of prior knowledge reorganization might include, for example, changes in the accuracy of current suspected diagnoses during reflection; such changes, in turn, are likely to depend on how much content knowledge learners have when they enter the diagnostic process. However, such learning mechanisms that are presumed to be associated with structured reflection have yet to be systematically measured and investigated.

### 1.2. Facilitating collaborative diagnostic reasoning with structured reflection and collaboration scripts

During *collaborative diagnostic reasoning*, diagnosticians share and elicit evidence and hypotheses (Radkowsch et al., 2020) to construct and maintain a shared conception of a problem (Liu et al., 2016; Roschelle & Teasley, 1995). Sharing and eliciting are collaborative activities that are necessary for complex problem solving (Liu et al., 2016). For instance, internists need to elicit new evidence not only by interviewing the patient themselves (i.e., patient interview) but also by examining materials they request from radiologists (e.g., an x-ray of the patient's thorax for suspected pneumonia). Thus, medical collaboration potentially leads to higher diagnostic accuracy than individual diagnostic reasoning and is often even essential for reducing diagnostic uncertainty (National Academies of Sciences, Engineering, and Medicine, 2015; Shafiran et al., 2017).

*Collaboration knowledge*, including meta-cognitive knowledge, which is information about the collaborators' knowledge, roles, and tasks, is critical for successful collaboration (Engelmann & Hesse, 2011). For example, an internist needs to know what patient-related information the radiologist requires to perform an examination (e.g., whether to look for a cause of inflammation or cancer; for examinations involving radiation exposure, whether a female patient is pregnant). This type of knowledge enables collaborators to anticipate and assess their counterpart's behavior and adapt to it accordingly (Fischer et al., 2013). Yet, physicians often struggle to share relevant information during collaboration, and this issue can negatively affect diagnostic accuracy (Tschan et al., 2009).

Collaborative diagnostic reasoning can be facilitated by methods such as external collaboration scripts, which can facilitate interaction during collaborative learning by initiating particular collaborative activities (Vogel et al., 2017). For instance, learners are provided with prompts to share specific or crucial information (Noroozi et al., 2013), which enhances the collaborative process and helps learners build

important functional script components internally (Fischer et al., 2013). Whereas collaboration scripts may be more beneficial for learners at earlier stages of collaboration skill development, such scripts may hinder learning for more experienced learners (i.e., *expertise reversal effect*; Fischer et al., 2013; Kalyuga, 2007). Thus, considering learners' prior collaboration knowledge as a basis for an adaptation of collaboration scripts seems promising because this type of knowledge may affect the extent to which a learner will benefit.

However, collaborative diagnostic reasoning involves diverse knowledge and skills and places high individual and collaborative diagnostic demands on learners, resulting in a considerable cognitive load (i.e., *collaborative cognitive load*; Kirschner et al., 2018). Due to this double load, it can be challenging to facilitate skill development through single scaffolds. Combining different types of scaffolds, so-called synergistic scaffolding (Tabak, 2004) seems promising for improving learners' individual and collaborative activities for collaborative diagnostic reasoning.

In contrast to different types of scaffolds that benefit the same aspect of a specific goal or need and are therefore potentially redundant, synergistic scaffolds refer to supports that interact to augment each other and therefore have the potential to provide more support than either scaffold can alone or even the sum of the scaffolds. The individual scaffolds in synergistic scaffolding address specific aspects of learners' needs and complement each other to support complex skills (Tabak, 2004).

A promising use of synergistic scaffolding for promoting collaborative diagnostic reasoning involves the combination of structured reflection and collaboration scripts. Content-specific scaffolds such as structured reflection seem promising for enhancing the effectiveness of external collaboration scripts on domain-specific knowledge (Vogel et al., 2017). Structured reflection that supports learners in reflecting on suspected diagnoses may help them prestructure the learning material and thereby better understand and implement the collaboration script and, thus, collaborate more effectively. However, the extent to which prior content and collaboration knowledge affects a potential synergistic effect of structured reflection and collaboration scripts is an open research question.

The degree to which diagnostic accuracy can be increased through collaboration is also likely to depend on collaborators' prior content knowledge. Collaborative diagnostic reasoning is particularly relevant and advantageous over individual reasoning in situations where diagnostic problems cannot be solved by an individual (Graesser et al., 2018). In the example from above, if the internist already has advanced illness scripts, the evidence that can potentially be elicited from the radiologist may be less critical for the internist's diagnostic accuracy than when the internist has less advanced illness scripts. Collaboration may even be detrimental to learning by generating extraneous load when the internist may be able to diagnose the case alone (Kirschner et al., 2018). It remains unclear to what extent the contribution of collaboration to diagnostic accuracy depends on the level of prior content knowledge learners have when they enter the diagnostic process.

## 2. Research questions

We derived the following research questions. RQ1: Can structured reflection and collaboration scripts in a medical simulation foster the learning of collaborative diagnostic reasoning by improving learners' performance in collaborative diagnostic activities (evidence sharing; hypothesis sharing) and learners' diagnostic outcomes (diagnostic accuracy; diagnostic justification)? We hypothesized that structured reflection (H1.1) and collaboration scripts (H1.2) would have positive individual effects and a synergistic (positive interaction) effect (H1.3) on collaborative diagnostic reasoning.

RQ2: What is the moderating effect of learners' prior knowledge with respect to the effects of structured reflection and collaboration scripts on collaborative diagnostic reasoning (evidence sharing; hypothesis

sharing; diagnostic accuracy; diagnostic justification)? We hypothesized that learners with a high level of content knowledge would benefit more from structured reflection (H2.1), whereas learners with a low level of collaboration knowledge would benefit more from collaboration scripts (H2.2). We additionally hypothesized that the synergistic effect of structured reflection and collaboration scripts would depend on learners' levels of prior content knowledge (H2.3a) and collaboration knowledge (H2.3b).

We also formulated two additional learning-process-related exploratory research questions. RQ3: To what extent does the accuracy of suspected diagnoses change during structured reflection as a function of prior content knowledge? RQ4: To what extent does collaboration contribute to diagnostic accuracy as a function of prior content knowledge?

## 3. Methods

### 3.1. Sample and design

We conducted an experiment with a 2x2 factorial design with structured reflection (levels: present, absent) and collaboration script (levels: present, absent) as between-subjects factors. We recruited advanced medical students in their 4th academic year and above from a 6-year medical study program and randomly assigned them to the four groups. These students had already attended courses in internal medicine and radiology and had experience with diagnosing real patient cases and collaborating with other physicians in medical clerkships.

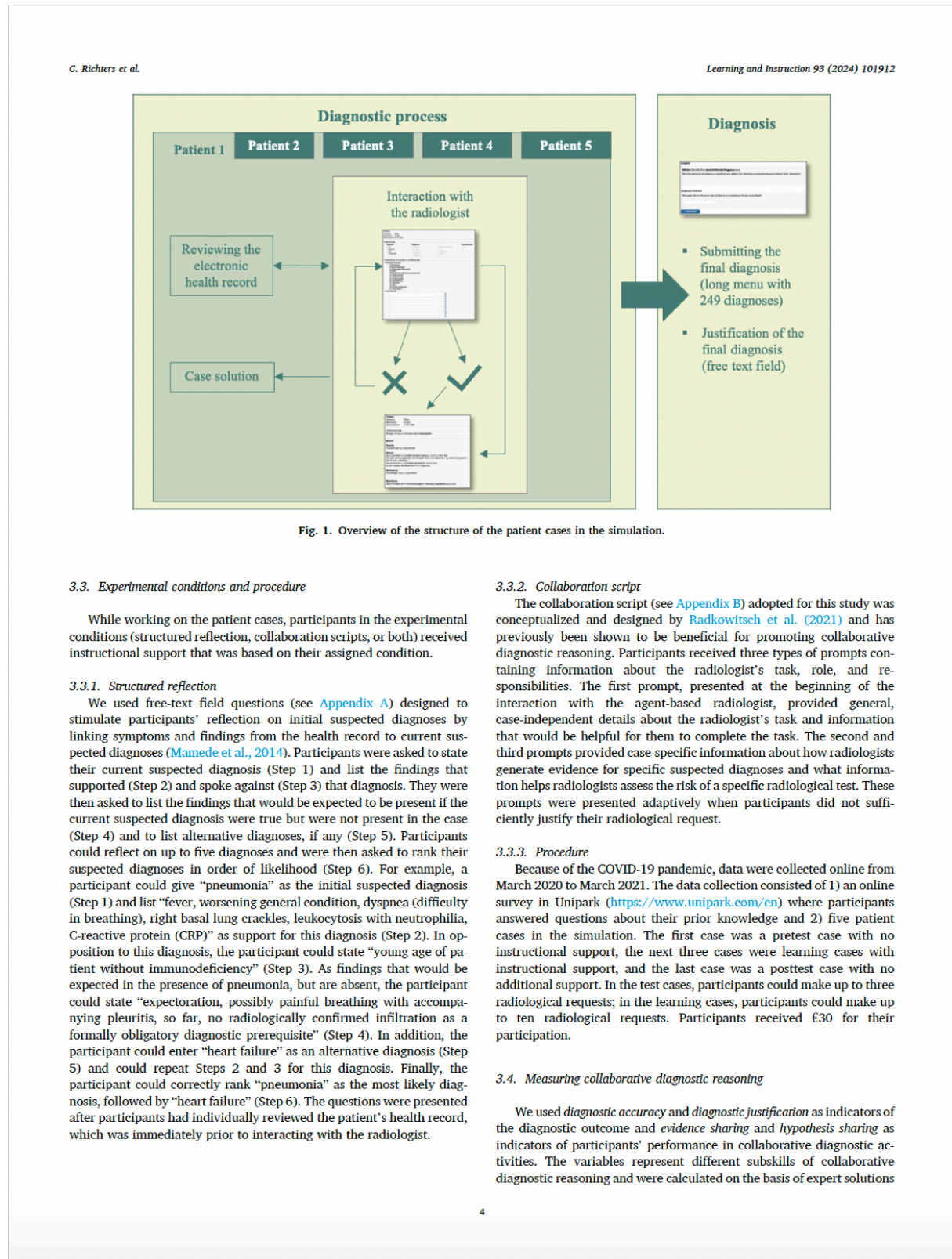
An a priori power analysis, presuming a medium effect size of  $f = 0.25$  with  $\alpha = 0.05$  and  $1 - \beta = 0.80$ , yielded a minimum number of participants of  $N = 128$ . Accordingly, a total of  $N = 151$  medical students ( $N_{\text{female}} = 109$ ) participated in the study and were used to address RQ1 and RQ2. The recruited medical students had been in medical school for an average of 5.33 years ( $SD = 0.09$ ) and were 25 years old ( $SD = 2.99$ ) on average. To address RQ3 and RQ4, we used a subsample ( $N = 79$ ) of the full data set, consisting only of participants from the structured reflection conditions.

### 3.2. The simulation and the learner task

The study was conducted in a simulated emergency department setting, embedded in a simulation-based learning environment. Participants collaborated with a computer agent (e.g., Graesser et al., 2018; Herborn et al., 2020) that played the role of a radiologist to diagnose five fictitious but realistic text-based patient cases with the leading symptom of fever (Fig. 1).

Participants were presented with an electronic health record containing information about the patient's admission, medical history, and laboratory results. They then filled out a request form for a radiological examination (e.g., chest x-ray) by providing the radiologist with information from the health record and suspected diagnoses from a long menu (249 different diagnoses). If the participants provided sufficient justification for their request, the radiologist performed the examination and shared their medical evaluation. The case was closed by selecting a final diagnosis from the long menu and justifying it in a free-text field. There was no time limit for completing the patient cases, but after participants had worked on a case for 15 min, they were prompted to present their solution to that case. On average, participants worked for 13.6 min on each patient case. The simulation was implemented in the CASUS learning platform (<https://www.instruct.eu/casus/virtual-patient-software>), which most medical students were already familiar with through their curriculum. A detailed description of the development, implementation, and validation of the simulation can be found in Radkowsch et al. (2020).







provided by a team of medical experts.

**Evidence sharing** was calculated as the ratio of the shared relevant evidence (patient information selected by participants and submitted to the radiologist) to all shareable relevant evidence. This indicator assesses learners' ability to identify what information a radiologist would need to perform the radiological examination and interpret the results. For each patient case, a score was calculated for each request and then a mean score was calculated for all requests, indicating learners' performance in evidence sharing. The range was from 0 points (indicating that no relevant evidence was shared) to 1 point (indicating all relevant evidence was shared). Across all cases, the internal consistency was  $\omega = 0.91$ .

**Hypothesis sharing** was calculated as the ratio of all shared relevant hypotheses (diagnoses selected by the learner from the long menu and submitted to the radiologist) to all shared hypotheses. This indicator assesses whether learners were able to state relevant hypotheses on the basis of case information and the extent to which irrelevant hypotheses were shared with the radiologist. We calculated how many of the shared diagnoses were also relevant for each case across all requests, resulting in a range of 0 points (indicating that no relevant hypotheses were shared) to 1 point (indicating that all shared hypotheses were relevant). Across all cases, the internal consistency was  $\omega = 0.66$ .

**Diagnostic accuracy** was assessed as the correctness of the final diagnosis. A *correct diagnosis with high specificity* (e.g., aspiration pneumonia) was assigned 1 point, a *correct diagnosis with low specificity* (e.g., pneumonia) was assigned 0.5 points, and an *incorrect diagnosis* (any diagnosis other than the correct one) was assigned 0 points. For RQ1 and RQ2, we used a binary indicator coded 0 (incorrect diagnosis) or 1 (correct diagnosis with high or low specificity). For the exploratory RQ4, we used the original coding to assess the diagnostic accuracy of the learner's final diagnosis in a learning case.

**Accuracy of suspected diagnoses** was assessed to address RQ3, also using the original coding of diagnostic accuracy to assess the accuracy of learners' suspected diagnoses, namely, the initial diagnosis (reflection step 1) and the post-reflection diagnosis before collaboration (reflection step 6) in the learning cases.

**Diagnostic justification** was assessed when participants submitted a correct final diagnosis. We calculated the proportion of relevant information mentioned in the free-text field out of all relevant information that would have fully justified the final diagnosis according to the medical experts. This indicator measures whether learners are able to correctly justify their diagnosis. For example, in the case of a patient with community-acquired pneumonia, this diagnosis would need to be properly justified with case information, such as "sudden onset of illness," "fever," "dyspnea (shortness of breath)," "fine crackles in the lower right lung on auscultation," "elevated inflammatory markers," and "chest X-ray/CT scan: infiltration or opacity or lobar pneumonia." We averaged the scores for each case. In the case of an incorrect final diagnosis, we assigned 0 points. We obtained a range from 0 points (indicating an insufficiently justified final diagnosis) to 1 point (indicating a sufficiently justified final diagnosis;  $\omega = 0.83$ ). Two independent raters achieved an overall interrater reliability of  $\kappa = 0.91$  for diagnostic accuracy and diagnostic justification in the first round, followed by three rounds of discussion with medical experts to achieve 100% agreement.

### 3.5. Measuring prior knowledge

**Prior content knowledge** was captured by measuring conceptual knowledge (Boshuizen & Schmidt, 1992) and strategic knowledge (Stark et al., 2011) from internal medicine and radiology. Conceptual knowledge was operationalized by 35 single-choice items that tested knowledge of pathophysiology and disease triggers (internal medicine) and knowledge of what can be detected by various radiological examinations and which radiological sign refers to which diagnosis (radiology). Participants were instructed to select one of five response options. An

example internal medicine item was "What are the most common pathogens of community-acquired pneumonia?" where the correct answer was "gram-positive bacteria".

Strategic knowledge was operationalized by 15 text-based cases using the key feature approach (Fischer et al., 2005). Key feature cases capture clinical knowledge and skills in multiple steps. Internal medicine knowledge was tested with seven cases, and radiology knowledge was tested with eight cases. After the patient was briefly introduced via a short patient vignette, internal medicine knowledge was tested with three questions, and radiology knowledge was tested with two questions. More precisely, internal medicine questions asked for the most likely diagnosis (e.g., "pertussis"), further examinations (e.g., "nasopharyngeal swab"), and the most important therapy (e.g., "symptomatic treatment"). Radiologic questions asked about imaging (e.g., "MRI with contrast") and potential risks (e.g., "renal failure is more likely"). Each question had eight possible answers, from which the learners were asked to select one. An example of a case description from internal medicine with sample solutions can be found in Appendix C.

For the final prior content knowledge score, we calculated mean scores across all conceptual and strategic knowledge questions, resulting in a range of 0 points to 1 point, indicating the learner's level of prior content knowledge ( $\omega = 0.81$ ).

**Prior collaboration knowledge** was measured with seven text-based patient cases with leading symptoms such as ascites, joint pain, impaired vigilance, B-symptoms (fever, night sweats and weight loss), back pain, dyspnea, and weakness, which required a radiological examination in the next step of the diagnostic work-up. Participants received a text-based introduction to the patient, including information about the patient's admission, signs and symptoms, and information about which examination needed to be performed to gather further evidence. Participants' task was to select the information that they would communicate to the radiologist performing the examination. Each case contained eleven items or pieces of information with or without radiological relevance. For example, a 28-year old patient was discovered unconscious with head injuries and left-sided abrasions. He exhibited disorientation, unresponsiveness, vomiting, and anisocoria, with vital signs indicating shallow breathing and low blood pressure. An emergency CT scan should be performed. The participant should correctly share "condition after fall from ladder," "impaired vigilance," "multiple episodes of vomiting," "contusion on the left forehead," and "anisocoria." The full case description of this example can be found in Appendix D. We assigned 1 point for each piece of information that was correctly rated. We calculated the proportion of correctly rated information out of all eleven pieces of information and then averaged across all cases, resulting in a range of 0 points to 1 point, indicating the learner's level of prior collaboration knowledge ( $\omega = 0.83$ ).

### 3.6. Statistical analyses

All analyses described below were conducted in R 4.0.3 (R Core Team, 2020). To address RQ1 and RQ2, we used multiple linear regression models to test the effects on evidence sharing, hypothesis sharing, and diagnostic justification and a multiple binomial logistic regression model to test the effects on diagnostic accuracy.

In each model, the posttest case score was used as the dependent variable, and the pretest case score as a covariate. Prior content and collaboration knowledge were included as moderators, and structured reflection and collaboration script were included as independent variables. In line with our hypotheses, we included two-way interaction terms (Structured Reflection x Collaboration Script; Content Knowledge x Structured Reflection; Collaboration Knowledge x Collaboration Script) and three-way interaction terms (Content Knowledge x Structured Reflection x Collaboration Script; Collaboration Knowledge x Structured Reflection x Collaboration Script) in each model.

All continuous variables (covariates and dependent variables) were z-standardized to simplify the interpretation of the results. Learners'

**Table 1**  
Descriptive statistics of collaborative diagnostic reasoning subskills by condition.

N	No Support		Structured Reflection		Collaboration Script		Both	
	M	SD	M	SD	M	SD	M	SD
<i>Pretest</i>								
Evidence Sharing	.43	.21	.43	.19	.40	.23	.40	.23
Hypothesis Sharing	.71	.35	.63	.35	.70	.36	.55	.39
Diagnostic Accuracy	.59	.30	.51	.26	.59	.28	.63	.32
Diagnostic Justification	.36	.24	.36	.28	.35	.24	.39	.27
<i>Posttest</i>								
Evidence Sharing	.41	.24	.44	.19	.43	.23	.47	.23
Hypothesis Sharing	.65	.35	.69	.33	.64	.34	.67	.36
Diagnostic Accuracy	.52	.40	.53	.33	.49	.33	.55	.36
Diagnostic Justification	.36	.24	.36	.28	.35	.24	.39	.27

Note. All variables are unstandardized and have a theoretical minimum of 0 and a maximum of 1. Caution should be exercised in interpreting the scores in absolute terms, as there is no normative sample to indicate which values reflect low, medium, or high performance. "Both" refers to the condition in which learners received both structured reflection and the collaboration script.

prior content and collaboration knowledge were included in each model as deviation values with respect to the predictor's levels of interest, namely, a low level of prior knowledge (*low prior knowledge*; one standard deviation below the sample mean and lower), an average level of prior knowledge (*average prior knowledge*; between one standard deviation below and one standard deviation above the sample mean), and a high level of prior knowledge (*high prior knowledge*; one standard deviation above the sample mean and higher). In line with our hypotheses, we focused our results on learners with low and high prior knowledge.

To investigate our exploratory research questions, we calculated rank correlations between the accuracy of the initial diagnosis and the post-reflection diagnosis (RQ3) and between the post-reflection diagnosis (before collaboration) and the final diagnosis (after collaboration) (RQ4) in the three learning cases in the reflection conditions. Moreover, we counted the absolute frequencies of the suspected diagnoses that were correct with high specificity, correct with low specificity, and incorrect at the three time points. We calculated the relative frequencies of learners who stuck to or deviated from their suspected diagnoses, including improvement and deterioration. Logistic regressions were

used to test the influence of prior content knowledge. For RQ3, the accuracy of the post-reflection diagnosis was used as the outcome, the initial diagnosis and prior content knowledge were used as predictors, and the interactions between the predictors were included as well. For RQ4, the final diagnosis was used as the outcome, the post-reflection diagnosis and prior content knowledge were used as predictors, and the interactions between the predictors were again included. For RQ4, a potential collaboration script effect was additionally controlled for.

**4. Results**

*4.1. Effects of structured reflection and collaboration scripts on collaborative diagnostic reasoning*

Table 1 presents the means and standard deviations of all numeric dependent variables, and Table 2 presents frequencies for the correct and incorrect diagnoses (the final indicator of diagnostic accuracy). Learners scored approximately one third to one half of the possible points across variables and conditions. An absolute interpretation of this

**Table 2**  
Frequencies of incorrect and correct diagnoses (final indicator of diagnostic accuracy) by condition.

	No Support		Structured Reflection		Collaboration Script		Both	
	correct	incorrect	correct	incorrect	correct	incorrect	correct	incorrect
<i>Pretest</i>								
Diagnostic Accuracy	9	23	6	35	10	28	11	20
<i>Posttest</i>								
Diagnostic Accuracy	10	21	9	29	7	28	9	20

Note: "Both" refers to the condition in which learners received both structured reflection and the collaboration script.

**Table 3**  
Summary of the effects on collaborative diagnostic activities and diagnostic outcomes on the posttest (one model for each outcome).

Predictor	Performance in collaborative diagnostic activities on the posttest				Diagnostic outcomes on the posttest			
	Evidence Sharing		Hypothesis Sharing		Diagnostic Accuracy		Diagnostic Justification	
	$\beta$	p	$\beta$	p	OR	p	$\beta$	p
H1.1 Structured Reflection	.09	.274	.05	.660	0.65	.465	.19	.080
H1.2 Collaboration Script	.09	.304	-.02	.884	0.34	.133	.02	.870
H1.3 Structured Reflection*Collaboration Script	.01	.928	-.01	.948	1.93	.546	-.09	.461
H2.1 Structured Reflection*Content Knowledge	.01	.961	.42	.007**	0.00	.235	.00	.996
H2.2 Collaboration Script*Collaboration Knowledge	.18	.183	-.03	.855	2.29	.306	.07	.710
H2.3a Structured Reflection*Collaboration Script*Content Knowledge	.29	.017*	-.21	.169	1.23e+09	.120	.01	.469
H2.3b Structured Reflection*Collaboration Script*Collaboration Knowledge	-.20	.129	-.04	.840	0.32	.261	-.11	.497

Note. All continuous variables are z-standardized. \*p < 0.05. \*\*p < 0.01.



C. Richters et al.

Learning and Instruction 93 (2024) 101912

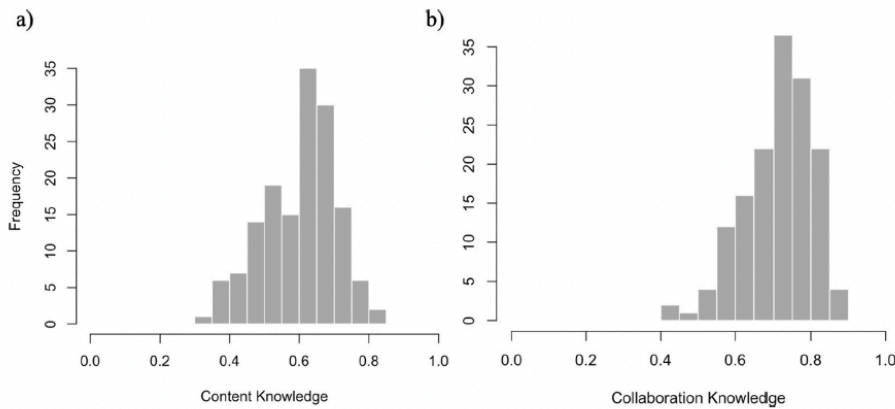


Fig. 2. Distributions of prior content and collaboration knowledge in the total sample.

Table 4  
Numbers and percentage of learners with low, medium, and high prior knowledge.

	Low	Medium	High
Content Knowledge	28 (19 %)	102 (66 %)	21 (14 %)
Collaboration Knowledge	28 (19 %)	98 (65 %)	25 (17 %)

Note. Low = One standard deviation below the mean; Medium = Mean; High = One standard deviation above the mean.

performance was not necessarily valid, as no norm sample was available. Radkowsch et al. (2021) previously used the same simulation and a similar metric for evidence sharing and found that learners scored higher. However, we substantially revised and improved this metric for the current study in collaboration with medical experts, resulting in increased internal consistency. Thus, we interpret the average moderate performance across learners as more indicative of the average difficulty of the items and the absence of ceiling effects. The descriptive results show that the learners in the different conditions hardly differed in the different subskills of collaborative diagnostic reasoning (Tables 1 and 2).

Table 3 summarizes all the hypothesized effects that were tested with the linear regression models (one per outcome).

Structured reflection and collaboration scripts did not have significant effects (separately or synergistically) on the performance of

collaborative diagnostic activities or on the diagnostic outcomes. Thus, the results did not support H1.1 through H1.3.

4.2. The moderating effect of prior knowledge with respect to the effects of structured reflection and collaboration scripts on collaborative diagnostic reasoning

As can be seen in Fig. 2, all learners scored in the middle to high range on prior content knowledge (Fig. 2a;  $M = 0.60, SD = 0.10$ ) and collaboration knowledge (Fig. 2b;  $M = 0.71, SD = 0.09$ ), reflecting the fact that the sample comprised advanced medical students. Table 4 shows the number of learners with low, medium, and high prior knowledge in the total sample.

Table 3 shows that, as a function of prior content knowledge, structured reflection did not have a significant effect on performance in evidence sharing, diagnostic accuracy, or diagnostic justification, but it had a moderate significant effect on hypothesis sharing ( $\beta = .42, p < .01$ ).

Fig. 3 shows that learners with low content knowledge performed significantly worse in hypothesis sharing when they reflected on their diagnostic process than when they did not ( $M_{\text{Difference}} = 3.57, SE = 1.69, p = .037$ ). Conversely, learners with high content knowledge performed significantly better when they reflected on their diagnostic process than when they did not ( $M_{\text{Difference}} = 3.73, SE = 1.68, p = .028$ ). Taken

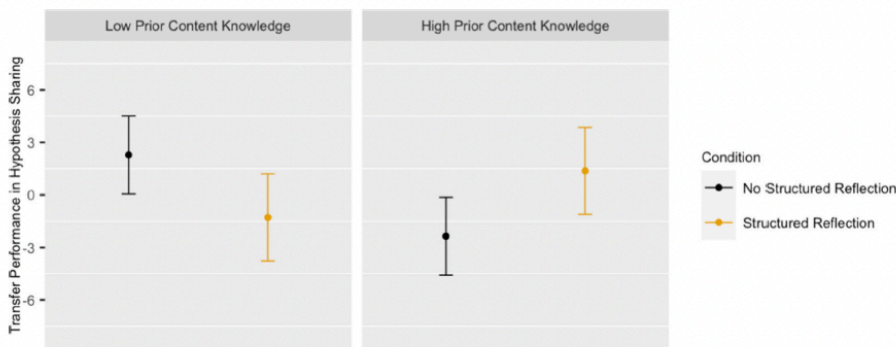


Fig. 3. Effects of structured reflection on the hypothesis sharing performance of learners with different levels of prior content knowledge. Note. Estimated means per group are represented by dots, accompanied by confidence intervals (represented by vertical lines).

C. Richters et al. Learning and Instruction 93 (2024) 101912

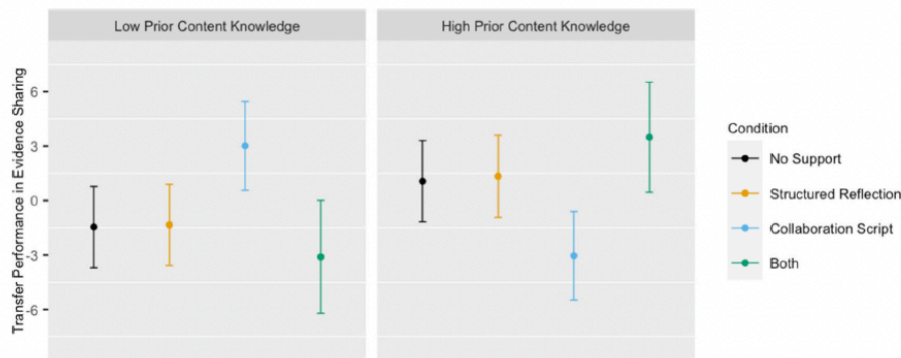


Fig. 4. Effects of structured reflection and collaboration scripts on performance in evidence sharing for learners with different levels of prior content knowledge. Note. Estimated means per group are represented by dots, accompanied by confidence intervals (represented by vertical lines). “Both” refers to the condition in which learners received both structured reflection and the collaboration script.

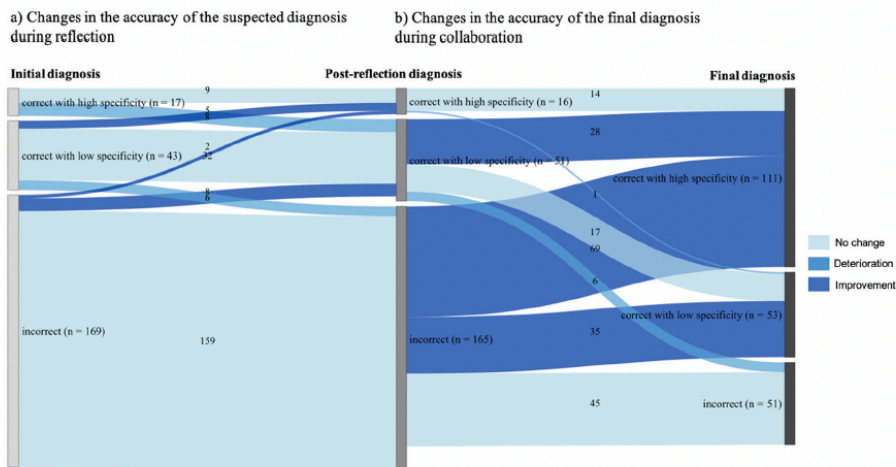


Fig. 5. Changes in the accuracy of the initial and final suspected diagnoses in the conditions with structured reflection. Note. Numbers represent frequencies. The thickness of the transitions corresponds to the frequency of the occurrence of the transition in the learning cases. Light blue = no change in the suspected diagnosis; medium blue = deterioration; dark blue = improvement.

together, these results partially supported H2.1.

The interaction of collaboration scripts and collaboration knowledge did not significantly affect performance in collaborative diagnostic activities or the diagnostic outcomes; thus, the data did not support H2.2.

Furthermore, structured reflection, collaboration scripts, and prior content knowledge had no significant effects on performance in hypothesis sharing, diagnostic accuracy, or diagnostic justification but had a small significant effect on evidence sharing ( $\beta = .29, p = .017$ ).

Fig. 4 shows that learners with low prior content knowledge performed significantly better in evidence sharing when they learned exclusively with collaboration scripts than when they only reflected on their diagnostic process ( $M_{\text{Difference}} = 4.35, SE = 1.67, p = .010$ ), additionally reflected on their diagnostic process ( $M_{\text{Difference}} = 6.11, SE = 2.00, p = .003$ ), or received no additional support ( $M_{\text{Difference}} = 4.46, SE = 1.67, p = .009$ ). There were no differences in evidence sharing when learners with low prior content knowledge learned without support, reflected on their diagnostic process, or learned with collaboration

scripts and additionally reflected on their diagnostic process.

By contrast, learners with high prior content knowledge had significantly weaker performances in evidence sharing when they received collaboration scripts alone compared with when they reflected on their diagnostic process ( $M_{\text{Difference}} = 4.37, SE = 1.68, p = .010$ ), reflected on their diagnostic process and additionally received collaboration scripts ( $M_{\text{Difference}} = 6.52, SE = 1.97, p = .001$ ), or received no additional support ( $M_{\text{Difference}} = 4.10, SE = 1.67, p = .016$ ). There were no differences when learners with high prior content knowledge learned without support, reflected on their diagnostic process, or learned with collaboration scripts and additionally reflected on their diagnostic process. Although the combination of structured reflection and collaboration scripts had no significant effects in the prior content knowledge groups compared with the respective control conditions, the combination was more conducive to learning for learners with high prior content knowledge than for those with low prior content knowledge ( $M_{\text{Difference}} = 6.59, SE = 2.20, p = .003$ ). Overall, these findings partially supported



C. Richters et al.

Learning and Instruction 93 (2024) 101912

#### H2.3a.

In addition, the three-way interaction between structured reflection, collaboration scripts, and collaboration knowledge did not significantly predict performance in collaborative diagnostic activities or the diagnostic outcomes, and thus, the data did not support H2.3b.

#### 4.3. Exploring the influence of structured reflection and collaboration scripts on the accuracy of suspected diagnoses and diagnostic accuracy as a function of prior content knowledge

Fig. 5a shows that for almost 90% of the learners, structured reflection did not bring about any change from the initial (mostly incorrect) diagnosis, and only a small number of learners improved. This finding was supported by a strong positive correlation between the accuracy of the initial diagnosis and the diagnosis after reflection ( $\rho = 0.84, p < .001$ ). Moreover, prior content knowledge had positive but smaller correlations with the accuracy of the initial diagnosis ( $\rho = 0.17, p = .009$ ) and the post-reflection diagnosis ( $\rho = 0.17, p = .010$ ). The logistic regression revealed that the strong correlation between the accuracy of the initial diagnosis and the post-reflection diagnosis was not influenced by prior content knowledge ( $b = -5.62, p = .491, OR = 0.00$ ), indicating that changes in the suspected diagnoses from structured reflection did not depend on prior content knowledge.

To address RQ3, regardless of learners' prior content knowledge, structured reflection did not change the accuracy of learners' suspected diagnoses.

Fig. 5b shows that nearly 60% of the learners improved after collaborating with the radiologist. Of those who began with a correct diagnosis with low specificity, 55% made it to the correct diagnosis with high specificity by the end of the diagnostic process. Of those with an incorrect diagnosis, just under 30% still maintained an incorrect diagnosis after collaborating with the radiologist. The remaining learners had at least a correct diagnosis with low specificity at the end of the diagnostic process. These results were mirrored by the significant, positive, but small correlation between the accuracy of the post-reflection diagnosis before collaboration and the final diagnosis ( $\rho = 0.28, p < .001$ ). There was also a small positive correlation between prior content knowledge and the accuracy of the final diagnosis after collaboration ( $\rho = 0.18, p = .010$ ). The logistic regression showed that the correlation between the accuracy of the post-reflection diagnosis and the accuracy of the final diagnosis was not influenced by prior content knowledge ( $b = -13.99, p = .142, OR = 0.00$ ), indicating that changes in the suspected diagnoses from collaboration did not depend on prior content knowledge.

To summarize the pattern of results that addressed RQ4, the accuracy of learners' diagnoses improved during collaboration regardless of their prior content knowledge.

## 5. Discussion

This study was not able to provide evidence of main effects of structured reflection or collaboration scripts or a synergistic effect of both scaffolds on advanced learners' collaborative diagnostic reasoning. As hypothesized, we found differential effects of structured reflection and collaboration scripts—when used separately or in combination—on the collaborative diagnostic reasoning of advanced learners with different levels of prior content knowledge. Because learners with different levels of prior knowledge benefit differently from structured reflection and collaboration scripts, it is not surprising that the combination of the two also does not generally have a positive effect on learning. Specifically, the two scaffolds did not address the same needs of learners and therefore did not complement each other so that learners could benefit more from them together than from either scaffold alone when the scaffolds were provided to learners with the appropriate level of prior knowledge (Tabak, 2004).

As hypothesized, structured reflection enhanced the collaborative

diagnostic activities of learners with high prior content knowledge. This interaction effect is consistent with recent meta-analytic findings (Chernikova, Heitzmann, Fink, et al. 2020; Chernikova, Heitzmann, Stadler, et al. 2020) and theoretical considerations in research (Mamede & Schmidt, 2022). More precisely, structured reflection improved the hypothesis sharing performance of learners with high prior knowledge but appeared to hinder the performance of learners with low prior knowledge. We assume that the repeated reflection in the learning phase helped high knowledge learners focus on relevant hypotheses in the case presented on the posttest, leading them to share fewer irrelevant hypotheses with their collaborator (the radiologist). Learners with low prior content knowledge, on the other hand, were more likely to deviate from the relevant hypotheses through repeated reflection.

It seems that learners with high prior knowledge are able to activate their knowledge as they go through the process of relating suspected diagnoses to information from the patient case at hand (Mamede & Schmidt, 2022) as they attempt to distinguish between relevant and irrelevant hypotheses. Such knowledge activation through structured reflection does not work for learners with low prior knowledge because their small knowledge base is less helpful for generating and evaluating alternative explanations. In addition, these learners may need to expend higher cognitive effort and may struggle to differentiate between relevant and irrelevant concepts or make connections between activated concepts (Wetzels et al., 2011). They may have shared a larger number of irrelevant hypotheses because they were repeatedly asked to think about different hypotheses in the structured reflection conditions.

Furthermore, collaboration scripts improved the evidence sharing performance of learners with low prior content knowledge, whereas such scripts were detrimental to the learning of learners with high prior content knowledge. Even though we expected that effectiveness would depend on collaboration knowledge, this result was largely consistent with our expectation and the meta-analytic findings (Chernikova, Heitzmann, Fink, et al. 2020). The difference in the effect of the interaction between the collaboration script and prior content knowledge and the main effect of the collaboration script found by Radkowsitch et al. (2021) may be due to the fact that Radkowsitch et al. investigated medical students who were in earlier semesters and had lower average content knowledge. The present results suggest that the collaboration scripts provided the optimal level of guidance for learners with low prior content knowledge without overloading working memory capacity, thus leading them to share more relevant evidence.

However, asking learners with low prior content knowledge to engage in structured reflection before they received the collaboration script may have cognitively overwhelmed these learners (Eckhardt et al., 2013). This situation may have left them without enough cognitive capacity to internally implement the elements of the external collaboration script, leading them to share less relevant evidence than when they received only the collaboration script. Although these learners usually generated relevant hypotheses (high performance in hypothesis sharing without structured reflection), the task of identifying case-relevant information to test these hypotheses during structured reflection does not yet seem feasible for them, as suggested by their low performance in evidence sharing with structured reflection.

For learners with high content knowledge, the collaboration scripts seemed to provide unnecessary guidance, possibly causing additional cognitive load that resulted in sharing less relevant evidence (Kalyuga, 2007). However, reflecting on the diagnostic process before receiving collaboration scripts compensated for this negative effect of the scripts for these learners, leading them to share a larger number of relevant pieces of evidence compared with when they received only collaboration scripts. Similar to note-taking, responding in writing during the structured reflection questions may help these learners activate concepts (illness scripts) and relate them to each other (identifying and organizing case-relevant evidence) without having to keep the individual pieces of evidence and the hypotheses active in working memory (Wetzels et al., 2011), thus potentially increasing cognitive capacity and



offsetting the negative effect of the collaboration scripts.

In sum, the interaction between structured reflection, collaboration scripts, and prior content knowledge found in this study does not indicate a synergistic effect of structured reflection and collaboration scripts (Tabak, 2004), from which learners benefit more or less depending on their prior knowledge. This finding is consistent with the lack of synergistic effect in RQ1. Rather, this interaction effect highlights the differential effects of the two forms of scaffolding for learners with different levels of prior knowledge.

Moreover, the effects of the instructional supports varied across the sharing activities. Structured reflection tended to improve hypothesis sharing, and collaboration scripts tended to improve evidence sharing. These differential effects can be seen as evidence of the validity of our different measures of collaborative diagnostic activities. Both instructional supports were designed to strongly promote these subskills.

Our exploratory analysis of the learning phase showed that structured reflection did not improve the accuracy of the initial suspected diagnosis, which can be considered a summative indicator of cognitive case representations (Charlin et al., 2012). According to recent explanations of the effects of structured reflection on diagnostic accuracy, reflection is expected to specifically improve early case representations by restructuring prior knowledge (Mamede & Schmidt, 2022). However, in this study, we found no evidence that structured reflection improved early case representations. This lack of evidence may be due to the characteristics of the cases or to the lack of developed illness scripts. However, even though reflection did not affect diagnostic accuracy, it did lead learners with high prior knowledge to share fewer irrelevant hypotheses. This finding suggests that structured reflection did not help learners actually specify hypotheses by reorganizing their knowledge, but rather, it helped them to consider a wider range of hypotheses by activating prior knowledge.

Furthermore, the interaction with the (simulated) radiologist helped learners improve the accuracy of their representation of the final case regardless of the collaboration script. This finding indicates that the opportunity to repeatedly consult with the collaboration partner as an external source of information (a feature of the simulation) appears to be more effective than additional support pertaining to reflection (Chernikova, Heitzmann, Stadler, et al. 2020), at least in the learning phase.

Overall, our findings contrast with previous theoretical assumptions that would have predicted the general effectiveness of structured reflection and collaboration scripts for learning individual diagnostic reasoning (Mamede & Schmidt, 2017) and collaborative diagnostic reasoning (e.g., Radkowsch et al., 2021), respectively. Structured reflection, which provides lower levels of guidance, is an effective type of support for learners with high prior knowledge, whereas collaboration scripts, which provide higher levels of guidance, are an effective way to support learners with low prior knowledge. Structured reflection can hinder learning for learners with low prior knowledge, whereas collaboration scripts can hinder learning for learners with high prior knowledge. Moreover, structured reflection compensates for the expertise reversal effect of collaboration scripts for learners with high prior content knowledge, whereas it cancels out the positive effect of such scripts for learners with low prior content knowledge. A combination of the two instructional supports leads to higher learning outcomes for learners with high prior knowledge than for learners with low prior knowledge (Eckhardt et al., 2013).

In the learning phase, learners used external information from their collaboration partner to confirm their final diagnoses, whereas internal resources used during reflection were insufficient for verification. In collaborative learning environments, learners sometimes increasingly rely on the collaboration partner as an external source of information after individual interactions with the learning material. Such a reliance has the potential to reduce their cognitive effort during their individual interactions with the learning material (i.e., *social loafing*; Karau & Williams, 1993). Another possible explanation for these findings is that participants' interactions with the collaboration partner itself may

trigger reflection because learners are aware that they need to think about the case information before engaging in collaboration; a phenomenon that could be called *collaboration apprehension*. The effects of structured reflection in learning environments designed to foster individual reasoning might not transfer easily to collaborative settings, where complex collaborative tasks introduce new demands but also provide valuable resources.

### 5.1. Limitations and further research

In this study, prior collaboration knowledge did not significantly interact with collaboration scripts. This lack of effect may be due to the performance-based adaptation of the script. Two of the three prompts were presented only when learners' requests were rejected by the radiologist, and rejection rates are likely to be negatively correlated with learners' collaboration knowledge. Thus, the collaboration script may have already been adapted to the learners' prior collaboration knowledge and therefore no longer interacted with it. However, the first prompt in the collaboration script was presented regardless of rejection, and factors other than collaboration knowledge can also contribute to rejection by the simulated radiologist. Future studies should examine the conditions under which learners receive negative feedback from their collaborators when sharing information.

Additionally, the test of collaboration knowledge used in this study focused exclusively on the critical aspect of information sharing in medical collaboration, but it did not cover other important aspects that may interact with the effectiveness of collaboration scripts (e.g., negotiation or regulation). However, information sharing is a challenge for aspiring physicians and practitioners, and we argue that it was therefore appropriate to prioritize it. To make valid statements about how collaboration knowledge interacts with other factors in collaborative settings, future research should also consider other knowledge facets and a broader range of collaborative activities.

Moreover, we interpreted the accuracy of suspected diagnoses at a given point in time during the diagnostic reasoning process as a summative indicator of case representations. However, the accuracy of suspected diagnoses does not fully capture the complexity and dynamics of how participants represent their cases, which may include other cognitive structures, such as the inclusion and exclusion of relevant and irrelevant case information, respectively, over time. For instance, learners may begin by proposing an incorrect diagnosis but develop a comprehensive representation of the case that evolves into an adequate final diagnosis over the course of the diagnostic process. Future studies should therefore examine other outcomes and process-related indicators of cognitive case representations in the diagnostic process when investigating structured reflection.

Furthermore, with regard to our exploratory analyses, the relatively brief learning phase in our experiment should be taken into account, as it could account for why structured reflection did not significantly affect early case representations. However, we adopted the reflection questions from Mamede et al.'s (2014) studies in which students also did not reflect for long periods of time. Future studies may investigate the moderating effects of reflection time on the effects of structured reflection on the development of complex skills.

### 5.2. Implications for educational practice

Some practical implications can be derived for higher educational practice. Specifically, our study provides practical insights for medical educators seeking to optimize the teaching of collaborative diagnostic reasoning using simulations. It is essential to adapt instructional support to the varying levels of prior content knowledge among advanced students for them to have effective learning experiences. Instructors can support students with a high level of prior content knowledge by incorporating structured reflection phases into the learning process. This scaffold encourages students to cognitively engage with the case in a

focused manner to prepare for subsequent collaborative diagnostic activities, resulting in improved collaborative performance. For students with low levels of prior content knowledge, instructors are advised to prioritize the use of collaboration scripts, to provide sufficient guidance without cognitively overwhelming the learners.

Moreover, it is crucial to understand the nuanced impact of instructional supports on subskills such as hypothesis sharing and evidence sharing. When fostering collaborative diagnostic reasoning, it seems important to consider not only students' prior knowledge but also the specific subskills students struggle with. Instructional support should be carefully selected and implemented accordingly.

Furthermore, recognizing the importance of collaboration partners is key to improving collaborative diagnostic reasoning. Collaboration in simulations seems to be a valuable source of knowledge in the collaborative diagnostic reasoning process that helps improve the final representation of the case and offers inherent learning potential. Therefore, we recommend that instructors actively encourage students to use their collaboration partners during the learning phase.

Overall, by adapting instructional support to meet learners' needs (e.g., their varying levels of prior knowledge), understanding the differential impact of different types of scaffolds on collaborative diagnostic activities, and strategically leveraging collaboration partners, medical instructors may create a comprehensive approach for enhancing the learning of collaborative diagnostic reasoning with simulations.

## 6. Conclusion

This study reveals that the effects of structured reflection and collaboration scripts in collaborative diagnostic reasoning simulations depend on learners' prior knowledge, and such practices can even have negative effects on learning. In contrast to other types of instructional support, such as collaboration scripts or worked examples, externally guided reflection on individual reasoning requires at least a certain level of prior knowledge. But before concluding that reflection support is generally not promising for learners with low prior knowledge, future research needs to explore the effect of increased guidance for structured reflection for learners with low prior knowledge, for example, by pre-structuring the reflection content, combining reflection with knowledge prompts, or supporting the cognitive linking of information. In contrast to previous explanations of the mechanisms behind reflection effects—especially on diagnostic accuracy in individual diagnostic reasoning—reflection does not change the diagnoses learners make (indicators of cognitive case representations) in collaborative diagnostic reasoning. However, interaction with a collaboration partner appears to do so.

Overall, providing advanced learners who have low prior content knowledge with collaboration scripts that offer high levels of external guidance and encouraging advanced learners who have high prior content knowledge to reflect before collaborating are promising approaches for facilitating collaborative diagnostic activities in simulations. These instructional supports are not necessarily suitable for facilitating diagnostic outcomes. In simulated collaborative settings, where the collaboration partner serves as an external source of information, engaging in collaboration itself contributes more to overall diagnostic outcomes by improving case representations than instructional support.

## Funding sources

This work was supported by a grant from DFG (Deutsche Forschungsgemeinschaft) Research Unit COSIMA [FOR 2385; FI 792/11-1].

## Research data availability

The dataset used for this study and the R script containing all inferential statistical analyses are stored in the open science framework

repository (OSF) and can be retrieved from [https://osf.io/fhd5b/?view\\_only=4aa9b3751a1845749d71a83015218d46](https://osf.io/fhd5b/?view_only=4aa9b3751a1845749d71a83015218d46).

## CRedit authorship contribution statement

**Constanze Richters:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft. **Matthias Stadler:** Conceptualization, Supervision, Writing – review & editing. **Anika Radkowitz:** Conceptualization, Data curation, Investigation, Methodology, Writing – review & editing. **Felix Behrmann:** Resources, Validation, Writing – review & editing. **Marc Weidenbusch:** Resources, Validation, Writing – review & editing. **Martin R. Fischer:** Conceptualization, Funding acquisition, Resources, Writing – review & editing. **Ralf Schmidmaier:** Conceptualization, Funding acquisition, Resources, Writing – review & editing. **Frank Fischer:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.

## Declaration of generative AI and AI-assisted technologies in the writing process

While preparing this work, the authors used *ChatGPT* to shorten text passages and *DeepL* to improve readability and language. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## Declaration of competing interest

No potential competing interest was reported by the authors.

## Acknowledgements

The authors would like to express their gratitude to Andreas Wildner, and Laura Brandl for their support before and during the data collection and for their support in coding the data.

## Appendix. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.learninstruc.2024.101912>.

## References

- 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](https://doi.org/10.1016/0364-0213(92)90022-M)
- Boud, D. (2001). Using journal writing to enhance reflective practice. *New Directions for Adult and Continuing Education*, *2001*(90), 9–18. <https://doi.org/10.1002/ace.16>.
- Bowen, J. L. (2006). Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, *355*(21), 2217–2225. <https://www.njcm.org/doi/10.1056/NEJMr054782>.
- Braun, L. T., Borrmann, K. F., Lottspeich, C., Heinrich, D. A., Kieseewetter, J., Fischer, M. R., & Schmidmaier, R. (2019a). Scaffolding clinical reasoning of medical students with virtual patients: Effects on diagnostic accuracy, efficiency, and errors. *Diagnosis*, *6*(2), 137–149. <https://doi.org/10.1515/dx-2018-0090>.
- Braun, L. T., Borrmann, K. F., Lottspeich, C., Heinrich, D. A., Kieseewetter, J., Fischer, M. R., & Schmidmaier, R. (2019b). Guessing right—whether or how medical students give incorrect reasons for their correct diagnoses. *GMS Journal for Medical Education*, *36*(6), Doc05. <https://doi.org/10.3205/ZMA001293>
- Charlin, B., Boshuizen, H. P. A., Custers, E. J., & Feltoch, 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., Lubarsky, S., Millette, B., Crevier, F., Audétat, M.-C., Charbonneau, A., et al. (2012). Clinical reasoning processes: Unravelling complexity through graphical representation. *Medical Education*, *46*(5), 454–463. <https://doi.org/10.1111/j.1365-2923.2012.04242.x>
- Chemikova, 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>
- Chemikova, 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>

### 3 Study 2: Reflection on Collaborative Action: Fostering Collaborative Diagnostic Reasoning in an Agent-Based Medical Simulation

*Constanze Richters \* Matthias Stadler \* Laura Brandl \* Ralf Schmidmaier \*  
Martin R. Fischer \* Frank Fischer*

**Reference:** Richters, C., Stadler, M., Brandl, L., Schmidmaier, R., Fischer, M. R., & Fischer, F. (2023). Reflection on collaborative action: Fostering collaborative diagnostic reasoning in an agent-based medical simulation. In C. Damsa, M. Borge, E. Koh, & M. Worsley (Eds.), *Proceedings of the 16th International Conference on Computer Supported Collaborative Learning - CSCL 2023* (pp. 209–212). International Society of the Learning Sciences. <https://doi.org/10.22318/csl2023.596913>

© 2023 International Society of the Learning Sciences. Presented at the International Conference of the Learning Sciences Annual Meeting (ISLS) 2023. Reproduced by permission.





International Society of  
the Learning Sciences

## Reflection on Collaborative Action: Fostering Collaborative Diagnostic Reasoning in an Agent-Based Medical Simulation

Constanze Richters, Matthias Stadler, Laura Brandl,  
constanze.richters@psy.lmu.de, matthias.stadler@psy.lmu.de, lbrandl@psy.lmu.de  
Department of Psychology, LMU Munich, Germany

Ralf Schmidmaier, Medizinische Klinik und Poliklinik IV, LMU Klinikum, LMU Munich, Germany,  
ralf.schmidmaier@med.uni-muenchen.de

Martin R. Fischer, Institut für Didaktik und Ausbildungsforschung in der Medizin, LMU Klinikum, LMU  
Munich, Germany, martin.fischer@med.uni-muenchen.de

Frank Fischer, Department of Psychology, LMU Munich, Germany, frank.fischer@psy.lmu.de

**Abstract:** Externally guided reflection on collaborative action is promising to foster collaborative diagnostic reasoning (CDR) for interdisciplinary practice. The interplay between the degree of external structure in reflection and learners' prior knowledge seems crucial for its effectiveness. In this study, we investigated the effects of low- and high-structured reflection on the learning of CDR in an agent-based medical simulation, depending on prior knowledge. We randomly assigned 195 medical students to one of three conditions: low-structured, high-structured, or no reflection support. We found positive effects of low-structured reflection on the learning of CDR for learners with low prior knowledge. For learners with high prior knowledge, both levels of structure seemed inappropriate. This study helps to individualize reflection support and lays the foundation for further empirical research on the effects of differently structured reflection as a function of prior knowledge.

### Theoretical background

*Collaborative diagnostic reasoning (CDR)* is critical to professional practice in many interdisciplinary fields such as medicine. Diagnostic accuracy often requires medical expertise from multiple subspecialties. Interdisciplinary collaboration can help physicians better understand the patient's illness and its underlying causes. In CDR, physicians are expected to *elicit* and *share* previously generated and evaluated *evidence* and *hypotheses* to reach final conclusions (Radkowsch et al., 2022). The extent that these cognitively demanding *collaborative diagnostic activities (CDAs)* are performed with high quality relies not only on content knowledge, but also on internal collaboration scripts (*Script Theory of Guidance*; Fischer et al., 2013). These scripts include *collaboration knowledge* (i.e., knowledge about collaborators' knowledge bases, roles, and tasks; Engelmann & Hesse, 2011).

Because the collaborative skills involved in CDR are complex, they need to be trained before entering professional practice, which can be accomplished through agent-based simulations (e.g., Radkowsch et al., 2022). Collaboration with a computer agent is particularly advantageous when the goal is to foster specific subskills. Furthermore, externally guided, respectively, structured reflection as a form of learning support has been shown to be beneficial for learning diagnostic skills (e.g., Mamede & Schmidt, 2017). Reflection in a broader sense describes an attentive, critical, and exploratory process of looking at one's thoughts and actions to potentially change them and improve one's understanding of the course of action (Nguyen et al., 2014). Structured reflection allows learners to reveal their thoughts and actions while diagnosing, making them more explainable and understandable to others, which could lead to better justified diagnoses (i.e., diagnostic justification).

However, previous studies investigating different types of reflection for learning diagnostic skills have yielded mixed results, which may be due to different specifications and degrees of structure within the reflection phases (Mamede & Schmidt, 2017). Moreover, recent meta-evidence suggests that reflection phases characterized by relatively low degrees of structure that are highly demanding in terms of self-regulation are more beneficial for learners with high levels of prior knowledge, as they simply do not provide enough structure for beginner learners (Chemikova et al., 2020). In the collaborative setting, Richters et al. (2022) found similar evidence of effectiveness for learners with high prior knowledge who improved their CDAs by reflecting on initial suspected diagnoses before collaborating in an agent-based medical simulation. However, it is still unclear whether a higher degree of structure in reflection phases can potentially influence their effectiveness for learners with low prior knowledge. To date, we are not aware of any studies that have systematically and theoretically varied the degree of structure in reflection, and examined its effectiveness as a function of prior knowledge, which may hold promise for the use of differently structured reflection across domains. In addition, Richters et al. (2022) did not examine structured reflection phases that directly addressed collaborative action, which may be even more effective in improving collaborative skills and preparing for interdisciplinary practice.



International Society of  
the Learning Sciences

### Research question

Depending on prior knowledge, to what extent can low- and high-structured reflection support, which stimulates learners to reflect on their collaborative action, improve their collaborative diagnostic reasoning, i.e., collaborative diagnostic activities (evidence sharing and hypotheses sharing) and diagnostic outcome (diagnostic accuracy and diagnostic justification) in an agent-based medical simulation?

### Methods

#### Sample and agent-based simulation

The sample consisted of 195 intermediate medical students ( $N_{\text{female}} = 130$ ;  $N_{\text{non-binary}} = 3$ ) between the third and fifth year of a six-year German medical school. We used a three-level one-factorial design, randomizing students to one of three conditions (low-structured, high-structured, or no reflection support). The study was conducted online via Zoom as part of the medical curriculum and was approved by the medical ethics committee. Study participation was mandatory for all students, but only student data voluntarily for research purposes were included in the study.

The learners' task was to diagnose five fictitious but realistic patient cases in an agent-based simulation developed by Radkowsch et al. (2022), in which they took on the role of an internal specialist in a hospital emergency department. For each patient case, learners were first given a medical record with relevant clinical information. Next, learners collaborated with an agent-based radiologist by requesting a radiological examination, sharing clinical information and suspected diagnoses from a long menu of 249 differential diagnoses to gain further insight into the patient's problem and reduce diagnostic uncertainty. As long as the learners provided sufficient justification for their requests, the radiologist shared their examination results and radiological assessment. Learners could use between three and ten request forms. Finally, learners completed each patient case by selecting a final diagnosis from the same long menu described above and justifying it in a free text field.

#### Structured reflection

After requesting radiological examinations in the patient case, learners in the low- and high-structured reflection support conditions received free-text questions that encouraged them to reflect on their actions in collaborating with the radiologist. The questions addressed the learners' CDAs. Specifically, learners were asked to what extent their requested examinations, shared clinical information, and shared suspected diagnoses helped them to diagnose and what they would improve about these activities in the future. In designing high and low degrees of structure based on theory, we followed the *principle of internal script guidance* (Fischer et al., 2013) and set the collaboration between the learner and the agent-based radiologist as the *play*. Further, we defined each CDA as a *scene*, i.e., requesting an examination: evidence elicitation, sharing clinical information: evidence sharing, and sharing suspected diagnoses: hypotheses sharing. Further, we defined the activities occurring within a scene (e.g., sharing clinical information: distinguishing between relevant and irrelevant clinical information) as *scriptlets*.

Learners with *low-structured reflection support* received *scene-level* questions with information about which scene to reflect on, e.g., reflection on sharing clinical information with radiology: *Was the sharing of clinical information with the radiologist helpful to your diagnostic process?*

Learners with *high-structured reflection support* received the same information about the scene to reflect on but with questions broken down to the *scriptlet-level*: e.g., *Did you share sufficient critical information with the radiologist? If you requested high-risk examinations, did you provide the radiologist with all the necessary information to perform your request? When sharing information, did you distinguish between information that was important to the radiologist and information that was not?*

#### Procedure

First, learners worked on an online survey consisting of questionnaires to measure prior knowledge and self-regulation skills. Next, all learners worked on one pretest patient case without reflection support. Afterward, learners worked on three more learning patient cases with or without reflection support according to their condition. Finally, all learners worked on one posttest patient case without reflection support.

#### Measures and analyses

We captured the quality of evidence sharing and hypotheses sharing as CDAs, and diagnostic accuracy and diagnostic justification as diagnostic outcomes. *Evidence sharing* was calculated by the proportion of shared relevant clinical information out of all shared clinical information, resulting in a range from 0 points indicating no relevant evidence was shared to 1 point indicating all shared evidence was relevant ( $\omega = .76$ ). *Hypotheses sharing* was calculated by the proportion of shared relevant diagnoses out of all shared diagnoses, resulting in a range from 0 points indicating no relevant hypotheses were shared to 1 point indicating all shared hypotheses were relevant ( $\omega = .75$ ). *Diagnostic accuracy* was assessed by the correctness of the final diagnosis, with 1 point





for a correct diagnosis and 0 points for an incorrect diagnosis ( $\omega = .44$ ). *Diagnostic justification* was calculated by the proportion of the relevant information mentioned out of all the relevant information that would have fully justified the final correct diagnosis. We obtained a range from 0 points, indicating an insufficiently justified or incorrect final diagnosis, to 1 point, indicating a correct and adequately justified final diagnosis ( $\omega = .79$ ).

As *prior knowledge*, we measured *collaboration knowledge* using seven text-based patient mini-cases on radiology ( $\omega = .87$ ). We calculated mean scores across all cases for each learner, which resulted in a range from 0 to 1 point indicating learners' prior knowledge. In addition, we assessed learners' prior content knowledge ( $\omega = .73$ ) and self-regulation skills ( $\omega = .82$ ) to control for potential effects on our results in the analysis.

We fitted linear regression models to test the effect of structured reflection as a function of prior knowledge on learners' CDAs and diagnostic outcomes. We fitted a binomial logistic regression model to test the effect on diagnostic accuracy. In each model, we used pretest score, content knowledge, and self-regulation as covariates; posttest score as the dependent variable; and assigned condition (dummy coded) and prior knowledge as predictors. We modeled the product of the two predictors as an interaction term to test the effect of structured reflection at different levels of learners' prior knowledge. For this purpose, prior knowledge was included as a deviation from the predictor levels of interest, namely low prior knowledge (one standard deviation below the sample mean) and high prior knowledge (one standard deviation above the sample mean). All continuous variables (covariates, predictors, and dependent variables) were z-standardized to facilitate interpretation of the results.

## Results

To address our research question, we looked more closely at the interaction effects between assigned condition and prior knowledge on each of the CDR indicators. An overall interaction effect on evidence sharing ( $F(2, 165) = 3.21, p = .043, \eta_p^2 = .04$ ) was found, indicating a large significant difference between low-structured reflection and no reflection support ( $b = -1.20, p = .019, d = -3.58$ ), but no difference between high-structured reflection and no reflection support ( $b = -0.03, p = .951$ ), or between both conditions with reflection support ( $b = 1.17, p = .057$ ). In addition, an overall interaction effect on diagnostic accuracy was found ( $\chi^2(2) = 9.6, p = .008, \Phi = 0.23$ ), indicating a large significant difference between low-structured reflection and no reflection support ( $b = -12.91, p = .011, OR = 0.00$ ), whereas no difference was found between high-structured reflection and no reflection support ( $b = -5.79, p = .267$ ), or between both conditions with reflection ( $b = -7.12, p = .214$ ). Further, an overall interaction effect on diagnostic justification was found ( $F(2, 146) = 3.66, p = .028, \eta_p^2 = .05$ ), indicating both a large significant difference between low-structured reflection and no reflection support ( $b = -4.01, p = .030, d = -2.05$ ) and between high-structured reflection and no reflection support ( $b = -4.70, p = .024, d = -2.03$ ), whereas no difference was found between the two conditions with reflection ( $b = 1.55, p = .474$ ). No overall interaction effect was found for hypotheses sharing ( $F(2, 165) = 0.32, p = .729$ ).

More specifically, the following patterns of results apply to these interaction effects: When working with low-structured reflection, learners with low prior knowledge scored significantly higher in evidence sharing ( $M = 0.11, SE = 0.05, 95\% \text{ CI } [0.01, 0.20]$ ) than when working without reflection support ( $M = -0.05, SE = 0.05, 95\% \text{ CI } [-0.14, 0.03]$ ). Further, learners with low prior knowledge were more likely to diagnose accurately when they worked with low-structured reflection ( $M_{\text{prob}} = 0.78, SE = 0.11, 95\% \text{ CI } [0.49, 0.92]$ ) compared to when they worked without reflection support ( $M_{\text{prob}} = 0.31, SE = 0.12, 95\% \text{ CI } [0.13, 0.57]$ ). Moreover, learners with low prior knowledge achieved higher diagnostic justification scores when they worked with low-structured ( $M = 0.41, SE = 0.18, 95\% \text{ CI } [0.05, 0.77]$ ) or high-structured reflection ( $M = 0.30, SE = 0.21, 95\% \text{ CI } [-0.12, 0.71]$ ) compared to when they worked without reflection support ( $M = -0.23, SE = 0.17, 95\% \text{ CI } [-0.55, 0.10]$ ). In contrast, learners with high prior knowledge performed significantly worse when they worked with high-structured reflection ( $M = -0.36, SE = 0.21, 95\% \text{ CI } [-0.77, 0.05]$ ) compared to when they worked without reflection support ( $M = 0.06, SE = 0.16, 95\% \text{ CI } [-0.25, 0.38]$ ).

## Discussion

The results of this study show that structured reflection support, which encourages learners to reflect on collaborative action in an agent-based medical simulation, can improve the CDR of learners with low prior knowledge. The effectiveness of structured reflection is particularly true for the low degree of structure, which enabled learners to improve their CDAs (i.e., evidence sharing) and diagnostic outcome. However, the low-structured reflection did not help these learners to improve their hypotheses sharing, which might be due to ceiling effects, as the learners generally scored high. Considering the *optimal scripting level principle* (Fischer et al., 2013), the present findings suggest that structured reflection at the scene level provides an optimal structure for learners who still lack sufficiently developed collaboration scripts (i.e., with low collaboration knowledge). For learners with high collaboration knowledge, on the other hand, this reflection seemed superfluous. Furthermore, structured reflection on the scriptlet-level seemed to be relatively superfluous for learners with low collaboration





knowledge and even taxing for learners with high collaboration knowledge. One possible explanation could be that the scriptlet-level might be more appropriate for fostering general collaboration skills, as suggested by Vogel et al. (2017); in addition, this fine-grained learning support might have possibly caused an expertise reversal effect (Kalyuga et al., 2003) for learners with high collaboration knowledge. However, although scriptlet-level reflection did not help learners with low prior knowledge to improve their evidence sharing or make a correct diagnosis, this learning support did help them improve their diagnostic justification. We assume that appropriate task-related specifications and information in structured reflection are particularly supportive for generating a coherent explanation of the experiences made in the diagnostic process (Mamede & Schmidt, 2017). Therefore, a higher degree of structure, including detailed information about task-related activities, may have helped learners with low prior knowledge who correctly solved the diagnostic case to coherently explain their successful diagnostic process. Thus, the effectiveness of different levels of structure in reflection phases may not only vary for learners with different prior knowledge, but also for different learning outcomes. While the scene-level may be fundamentally helpful for learners with low prior knowledge to solve the case correctly, the scriptlet-level may help these learners who solved the case correctly to adequately argue for the correct case solution.

Overall, our findings on the moderating role of prior knowledge in the effectiveness of structured reflection differ from previous findings that learners with high prior knowledge benefit most from structured reflection (Chernikova et al., 2020; Richters et al., 2022). These differences in the effects of reflection may be because different contexts (e.g., individual vs. collaborative), content, and learning outcomes (different forms of knowledge vs. task-related or general collaborative skills) of reflection require different specifications and degrees of structure. Furthermore, an objective scaling of different degrees of structure of reflection phases in different studies and thus the possibility to compare different effects is missing. Future research should replicate our findings and further investigate the effectiveness of differently structured reflection phases based on theory (e.g., more fine-grained at the scene-level). To this end, different reflection content and goals should be investigated, including learners with a wider range of prior knowledge.

In conclusion, for learners with low collaboration knowledge, scene-level reflection is most promising for improving interdisciplinary collaboration skills involved in CDR. For learners with more collaboration knowledge, structured reflection at the play-level may be a promising approach to explore in the future.

## References

- 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.
- Engelmann, T., & Hesse, F. W. (2011). Fostering sharing of unshared knowledge by having access to the collaborators' meta-knowledge structures. *Computers in Human Behavior*, 27(6), 2078–2087.
- Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. *Educational Psychologist*, 48(1), 56–66.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38(1), 23–31.
- Mamede, S., & Schmidt, H. G. (2017). Reflection in medical diagnosis: A literature review. *Health Professions Education*, 3(1), 15–25.
- 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.
- Radkowsitch, A., Sailer, M., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2022). Diagnosing collaboratively: A theoretical model and a simulation-based learning environment. In F. Fischer, & A. Opitz (Eds.), *Learning to diagnose with simulations: Teacher education and medical education* (pp. 123–141).
- Richters, C., Stadler, M., Radkowsitch, A., Behrmann, F., Weidenbusch, M., Fischer, M.R., Schmidmaier, R., & Fischer, F. (2022). Making the rich even richer? Interaction of structured reflection with prior knowledge in collaborative medical simulations. In A. Weinberger, W. Chen, D. Hernández-Leo, & B. Che (Eds.), *Proceedings of the 15th International Conference on CSCL* (pp. 155-162).
- Vogel, F., Wecker, C., Kollar, I., & Fischer, F. (2017). Socio-cognitive scaffolding with computer supported collaboration scripts: A meta-analysis. *Educational Psychology Review*, 29(3), 477-511.

## Acknowledgments

This study was supported by a grant from DFG (Deutsche Forschungsgemeinschaft) Research Unit COSIMA (FOR 2385; FI 792/11-2).

## 4 Study 3: Who is on the Right Track? Behavior-Based Prediction of Diagnostic Success in a Collaborative Diagnostic Reasoning Simulation

*Constanze Richters \* Matthias Stadler \* Anika Radkowitzsch \* Ralf Schmidmaier \* Martin R. Fischer \* Frank Fischer*

**Reference:** Richters, C., Stadler, M., Radkowitzsch, A., Schmidmaier, R., Fischer, M. R., & Fischer, F. (2023). Who is on the right track? Behavior-based prediction of diagnostic success in a collaborative diagnostic reasoning simulation. *Large-scale Assessments in Education*, 11(1), 3. <https://doi.org/10.1186/s40536-023-00151-1>

© 2023 The Authors. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original authors and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## RESEARCH

## Open Access



# Who is on the right track? Behavior-based prediction of diagnostic success in a collaborative diagnostic reasoning simulation

Constanze Richters<sup>1,2\*</sup>, Matthias Stadler<sup>1</sup> , Anika Radkowsch<sup>5</sup>, Ralf Schmidmaier<sup>2,3</sup>, Martin R. Fischer<sup>2,4</sup> and Frank Fischer<sup>1,2</sup>

\*Correspondence:  
constanze.richters@psy.lmu.de

<sup>1</sup> Department of Psychology,  
Ludwig-Maximilians-Universität  
München, Munich, Germany

<sup>2</sup> Munich Center of the Learning  
Sciences (MCLS), Ludwig-  
Maximilians-Universität  
München, Munich, Germany

<sup>3</sup> Medizinische Klinik und  
Poliklinik IV, University Hospital,  
Ludwig-Maximilians-Universität  
München, Munich, Germany

<sup>4</sup> Institute of Medical Education,  
University Hospital, Ludwig-  
Maximilians-Universität  
München, Munich, Germany

<sup>5</sup> Leibniz Institute for Science  
and Mathematics Education, Kiel,  
Germany

## Abstract

**Background:** Making accurate diagnoses in teams requires complex collaborative diagnostic reasoning skills, which require extensive training. In this study, we investigated broad content-independent behavioral indicators of diagnostic accuracy and checked whether and how quickly diagnostic accuracy could be predicted from these behavioral indicators when they were displayed in a collaborative diagnostic reasoning simulation.

**Methods:** A total of 73 medical students and 25 physicians were asked to diagnose patient cases in a medical training simulation with the help of an agent-based radiologist. Log files were automatically coded for collaborative diagnostic activities (CDAs; i.e., evidence generation, sharing and eliciting of evidence and hypotheses, drawing conclusions). These codes were transformed into bigrams that contained information about the time spent on and transitions between CDAs. Support vector machines with linear kernels, random forests, and gradient boosting machines were trained to classify whether a diagnostician could provide the correct diagnosis on the basis of the CDAs.

**Results:** All algorithms performed well in predicting diagnostic accuracy in the training and testing phases. Yet, the random forest was selected as the final model because of its better performance ( $\kappa = .40$ ) in the testing phase. The model predicted diagnostic success with higher precision than it predicted diagnostic failure (sensitivity = .90; specificity = .46). A reliable prediction of diagnostic success was possible after about two thirds of the median time spent on the diagnostic task. Most important for the prediction of diagnostic accuracy was the time spent on certain individual activities, such as evidence generation (typical for accurate diagnoses), and collaborative activities, such as sharing and eliciting evidence (typical for inaccurate diagnoses).

**Conclusions:** This study advances the understanding of differences in the collaborative diagnostic reasoning processes of successful and unsuccessful diagnosticians. Taking time to generate evidence at the beginning of the diagnostic task can help build an initial adequate representation of the diagnostic case that prestructures subsequent collaborative activities and is crucial for making accurate diagnoses. This information



could be used to provide adaptive process-based feedback on whether learners are on the right diagnostic track. Moreover, early instructional support in a diagnostic training task might help diagnosticians improve such individual diagnostic activities and prepare for effective collaboration. In addition, the ability to identify successful diagnosticians even before task completion might help adjust task difficulty to learners in real time.

**Keywords:** Simulations, Collaborative diagnostic reasoning processes, Learning process analysis, Medical education, Logfile analysis, Supervised machine learning

### Introduction

Training in collaborative diagnostic reasoning is important across various domains in higher education because, in practice, diagnosticians often work together in teams (e.g., in medical consultations, classrooms, scientific laboratories, therapeutical supervision, or industrial engineering). Previous research on collaborative problem solving (e.g., Graesser et al., 2018) has highlighted the need for training in collaboration skills, which form a key competence of the twenty-first century. For example, in order to assess a student's learning status or to diagnose a patient's health problem accurately, teachers or physicians, respectively, must be able to generate, elicit, and share evidence as well as come up with and share hypotheses and draw conclusions (so-called *collaborative diagnostic activities* [CDAs]; Fischer et al., 2014; Radkowsitch et al., 2022). The improvement of such complex skills is related to a constant increase in learners' current *zone of proximal development* (Vygotsky, 1978), which describes what learners are currently not able to solve on their own but could certainly solve with external help. Thus, for optimal learning outcomes, there is a need for learning environments that include problem-solving tasks that are slightly more difficult than what learners can already solve independently (Roosevelt, 2008).

Simulations are often used to train complex skills. They enable standardized repetitions of individual learning steps and deliberate practice (Ericsson, 2004) and training in rarely occurring or critical real-life situations (e.g., rare or deadly diseases). There is evidence that simulations are particularly effective when the embedded instructional support is adaptive (Chernikova et al., 2020). However, properly and immediately adjusting the appropriate instructional support to learners' individual needs represents a challenge for instructional designers and educators. Moreover, being able to identify at what point in time learners can already solve the task without additional support might also be helpful for removing or fading out (Pea, 2004) instructional support that might even hinder learning (Kalyuga et al., 2003). One starting point for such an adjustment involves using machine learning to analyze learners' behavior on the basis of process data that are recorded and stored by the computer system (e.g., log files). Previous studies have demonstrated that analyzing learners' behavior can help identify how learners approach certain problems (Griffin and Care, 2015) and can aid the understanding of specific misconceptions that arise in the learning process (e.g., Stadler et al., 2019). Earlier analyses showed that specific actions in the learning environment were associated with task completion success (Cirigliano et al., 2020). Thus, assessing behavioral indicators of diagnostic reasoning skills (e.g., CDAs) and relating them to the diagnostic outcome can provide insights into whether learners currently have adequate or inadequate representations

of the diagnostic problem. For instance, such behavioral indicators may be beneficial for assessing whether a patient's relevant signs and symptoms are adequately interpreted (Charlin et al., 2012). If a learner's performance can be predicted before the diagnostic task is completed, instructors may be able to take early action to improve learning outcomes. The information obtained from the analysis of CDAs could provide a promising starting point for performance-based individualized instructional support and could make a positive contribution to effective diagnostic training.

#### **Collaborative diagnostic reasoning as a complex skill**

The process of diagnosing can be considered the "goal-oriented collection and interpretation of case-specific or problem-specific information to reduce uncertainty" (Heitzmann et al., 2019, p. 4) to be able to make professional decisions. Specific diagnostic situations require planned or initiated actions based on observations of and information about the problem to meet the diagnostic goal. Building on the conceptual framework of scientific reasoning and argumentation (Fischer et al., 2014), Heitzmann et al. (2019) defined such actions as *epistemic diagnostic activities*, which consist of, for example, evidence generation, evidence evaluation, hypothesis generation, and drawing conclusions (see also Klahr & Dunbar, 1988). These activities are grouped into a framework but cannot be placed in a fixed general sequence or order. According to Fischer et al. (2014), *evidence generation* refers to generating evidence in favor of or against a claim. Next, *evidence evaluation* is aimed at assessing "the degree to which a certain piece of evidence supports a claim or theory" (Fischer et al., 2014, p. 34). *Hypothesis generation* refers to the process by which students frame possible answers to the question, hereby deriving them from plausible models, available theoretical frameworks, or empirical evidence that they have access to. Finally, in *drawing conclusions*, students integrate different pieces of evidence "by weighing every single piece according to the method by which it was generated and by the rules and criteria of the discipline" (Fischer et al., 2014, p. 35).

To ensure high diagnostic quality, practicing scientists, physicians, psychologists, teachers, and engineers often need to diagnose in teams. Collaborative diagnostic reasoning (and, more generally, collaborative problem solving) has some advantages over individual reasoning, such as dividing labor according to individual professions, different perspectives, and knowledge bases (OECD, 2017), plus higher diagnostic accuracy (Tschan et al., 2009). However, existing research has demonstrated that students often lack collaborative skills (e.g., Hall & Buzzwell, 2012; O'Neill et al., 2013; Pauli et al., 2008) and that practitioners lack collaborative diagnostic reasoning skills (e.g., physicians; Tschan et al., 2009). By extending Fischer et al.'s (2014) framework of individual diagnostic activities to collaborative contexts, Radkowsch and colleagues (2022) recently defined CDAs in their model of collaborative diagnostic reasoning. This model describes the diagnostic reasoning processes of two diagnosticians with different knowledge backgrounds. In doing so, Radkowsch and colleagues (2022) distinguished individual activities from social or collaborative activities, namely, sharing, elicitation, negotiation, and coordination. The model can also be viewed as an integration and extension of Liu et al.'s (2015) collaborative problem-solving framework and Klahr and Dunbar's (1988) scientific discovery as dual search (SDDS) model. More precisely, the collaborative diagnostic reasoning model combines individual and collaborative activities and integrates them



into CDAs referred to as *eliciting, sharing, negotiating, and coordinating evidence* as well as *hypotheses* (Radkowsch et al., 2022). During the diagnostic reasoning process, these activities help diagnosticians construct and maintain a shared conception of a problem (Roschelle & Teasley, 1995). The quality of CDAs is assumed to be crucial for the success of the collaboration (Radkowsch et al., 2022).

#### **Using process data analysis for individualized learning support in the context of simulation-based complex skills training**

To foster complex skills (e.g., collaborative diagnostic reasoning), simulations have been established in various domains in higher education. Flight simulators have been used in pilot training for many years (Landriscina, 2012) just as surgical simulations are common in the medical context (Al-Kadi & Donnon, 2013). Standardized training in simulations has different advantages over training in real-world scenarios. First, simulations can reduce the complexity of a situation while offering learners the opportunity to apply their knowledge to specific cases in standardized settings (Grossman et al., 2009). Second, simulations enable repetitive deliberate practice, which has been considered to be crucial for acquiring professional expertise (Ericsson, 2004). Third, unlike real-life scenarios, simulations enable training while ensuring ethical safety regarding mental or physical human conditions (Gegenfurtner et al., 2014; Grossman et al., 2009). Useful real-learning situations are often either rare (e.g., disruptive patient behavior) or too critical (e.g., amniotic fluid examination) to be used for training purposes. In real life, failure or complications would have serious unacceptable consequences (Ziv et al., 2003). A large number of primary studies and several meta-analyses have yielded positive effects of simulation-based learning and have provided recommendations for their implementation (e.g., Chernikova et al., 2020; Cook et al., 2013).

However, despite their potential, the effective use of simulations in training, especially in the field of collaborative diagnostic reasoning, remains challenging. To enhance highly effective learning that is based on complex and challenging problems, additional instructional support is often important (e.g., Hmelo-Silver et al., 2007). Instructional support is considered to be particularly effective when it is adapted to learners' individual needs (i.e., microlevel; e.g., Plass & Pawar, 2020). Dynamic assessment that can be realized by measuring learners' current performance in the problem-solving process (performance-based adaptation; e.g., VanLehn, 2011) can provide an adaptive basis for instructional support. One way to dynamically assess learners' performance is to analyze learners' behavior. This allows researchers to identify processes that are related to arriving at a successful solution to the problem (Griffin & Care, 2015) and to understand misconceptions in the learning process (e.g., Stadler et al., 2019). Compared with looking at only the summative outcome measure of a learning process, considering the learning process itself also offers the advantage of identifying subtler differences among learners that might not be reflected in the outcome measure (Stadler et al., 2020). To foster collaborative diagnostic reasoning skills, it might be useful to detect whether learners are currently leaning toward a correct or incorrect diagnosis—which is related to whether they have adequate or inadequate representations of the patient's problem—by predicting diagnostic accuracy. Following the hierarchical model of clinical reasoning processes (MOT; Charlin et al., 2012), which depicts the complex process of clinical reasoning as a

network, these cognitive representations of the patient's problem evolve and change as the diagnostic reasoning process unfolds.

In recent years, interest in predicting learners' performance with machine learning has increased considerably (e.g., Baker & Inventado, 2014; Hilbert et al., 2021). For instance, previous studies have predicted learners' performance to identify those at risk of failing a course (e.g., Tomasevic et al., 2020) or to support an intervention (e.g., San Pedro et al., 2013). The data for such an assessment can be collected automatically in real time while the learners are exploring the learning content (e.g., stealth assessment; Shute, 2011). However, the analysis of learners' behavior—especially during collaborative diagnostic reasoning procedures for automated assessments—based on wide, general behavioral indicators has not yet been sufficiently investigated or implemented in practice. First, previous studies that have analyzed learners' behavior have tended to focus on problem-solving strategies (e.g., Stadler et al., 2019) rather than on diagnostic activities. Second, the chosen behavioral indicators have been highly specific to the problem context presented in the learning environment (e.g., necessary and unnecessary actions for fixing a water pump; Zhu et al., 2016). A more general and replicable approach may be found in relating successful learning to more generic behavioral indicators that can be found across a broader range of diagnostic contexts (O'Neil et al., 2003). Predictions of diagnostic success could inform learners and instructors in real time whether or not learners are currently in need of instructional support in the collaborative diagnostic reasoning process and can thus help to individually address learners' zone of proximal development (Vygotsky, 1978). Supporting learners with individual instructional support in single diagnostic cases enables dynamic diagnostic training, which is important for the learning of collaborative diagnostic reasoning skills. Research on complex problem solving has shown that learners use problems that have been solved as blueprints for similar new problems to find new solutions (Richter & Weber, 2013). The opportunity to use learners' learning behavior to readjust instructional support for each diagnostic case would offer the advantage of being able to take learning progress into account.

However, beyond the ability to predict diagnostic success or failure, in order to effectively adapt instructional support, it is necessary to better understand the behavior of successful and unsuccessful diagnosticians. We consider the CDAs to be broad process-based indicators of collaborative diagnostic reasoning skills that can be used in various collaborative diagnostic contexts—from diagnosing diseases to assessing a student's current learning status—to identify differences in successful and unsuccessful diagnostic reasoning processes.

### **This study**

The goals of this study were twofold. First, to provide a general and replicable approach for analyzing diagnostic reasoning processes, we aimed to link diagnostic accuracy to broad behavioral indicators by analyzing the CDAs displayed in a medical training simulation using log files. We aimed to investigate differences in successful and unsuccessful diagnostic reasoning processes and to determine the extent to which CDAs could predict diagnostic accuracy. Second, we aimed to investigate how early diagnostic accuracy could be predicted from CDAs on the basis of behavior exhibited before, during, and



after collaboration. In this way, we aimed to exploratively identify early starting points for effective ways to adapt instructional support.

We addressed the following research questions:

1. To what extent can CDAs predict diagnostic accuracy in a medical training simulation using machine learning classification models?
2. How early in the process of making a diagnosis can diagnostic accuracy be reliably predicted from CDAs in a medical training simulation using machine learning classification models?

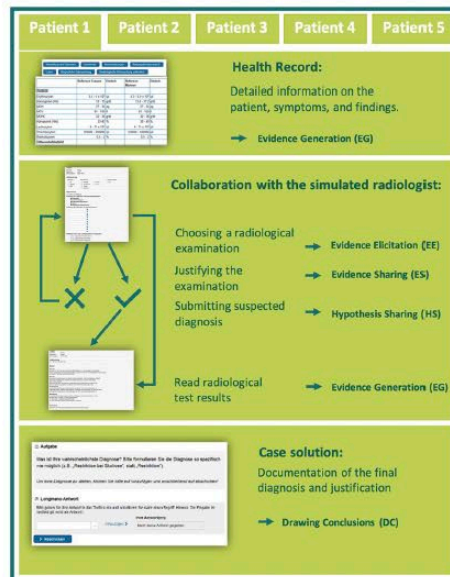
## Methods

### Sample, simulation, and procedure

To predict diagnostic accuracy, we selected a sample with sufficiently high variance in prior knowledge. Participants were 73 medical students ( $N_{\text{female}}=51$ ) in their clinical years from the 5th semester and higher ( $M=8.32$  semesters,  $SD=2.80$ ) of a 6-year study program and 25 physicians from internal medicine ( $N_{\text{female}}=11$ ) with a minimum of 3 years of clinical experience ( $M=13.6$  years of clinical work,  $SD=10.5$ ). Participation for medical students was limited to those in their clinical years because we assumed that, in principle, students in their preclinical years have not yet generated systematic prior knowledge of radiology and internal medicine. Participation was voluntary. The mean age of the participating medical students was  $M=24.9$  ( $SD=4.23$ ); for the participating physicians, it was  $M=42.0$  ( $SD=11.7$ ).

In the text-based simulation, participants acted in the role of an internal specialist in the emergency department of a hospital. Figure 1 presents an overview of the structure of the simulation. Five patient cases that all had the same structure had to be processed. Sequentially, participants received an electronic health record of five fictitious patients who all presented with a fever of unknown origin. The electronic health record was implemented as an electronic folder that contained information about the patients' admission, their medical history, findings from a physical examination, and laboratory results. Participants could navigate between these sections by clicking on representatively named buttons (e.g., medical history), which led to texts with the respective information. The health record could be accessed during the entire diagnostic procedure. After individually processing the information presented in the health record, participants were asked to collaboratively generate further evidence by requesting a radiological examination from an agent-based radiologist.

Participants filled out a request form by choosing a radiological examination and by sharing evidence of the suspected disease and hypotheses with the agent-based radiologist. The agent-based radiologist conducted the radiological examination only when the request was appropriately justified by the shared evidence and hypotheses. Participants then received a detailed document containing the radiological evidence they requested. Otherwise, participants were asked to revise their requests. After requesting the radiological examination, participants could request up to 10 additional radiological examinations. Finally, participants solved the patient case by indicating the diagnosis they thought was most likely. In sum, a participant's task was to collect evidence and generate



**Fig. 1** Overview of the Structure of the Simulation With Corresponding Assignment of Activities to the CDAs

hypotheses about a patient's illness to reduce uncertainty about the final diagnosis. The simulation was implemented in the learning platform CASUS ([www.instruct.eu](http://www.instruct.eu)). For further information about the development, implementation, and validation of the simulation, see Radkowsch et al. (2022). The study was conducted in a laboratory setting. Participants could work on the cases without time constraints but were asked to work efficiently. They were prompted to offer a solution to a case after 15 min. The total processing time per case was  $Mdn_{\min} = 15.26$ . The minimum median processing time was 6.77 min, and the maximum was 26.03 min. Participants received 25€ as compensation for their participation.

#### Coding collaborative diagnostic activities and measuring diagnostic accuracy

Participants' activities (i.e., their clicks and text entries) during the diagnostic reasoning process were automatically recorded and assigned to the five abovementioned previously specified CDAs (Radkowsch et al., 2022). Due to the implementation of the simulation, some activities were individual diagnostic activities (e.g., evidence generation), whereas other activities were collaborative diagnostic activities with the agent-based radiologist (e.g., evidence sharing). The overview of the structure of the simulation in Fig. 1 contains the corresponding assignment of activities to the CDAs within each section. The ways in which the activities were assigned to the activity categories is described below in more detail.

**Evidence generation (EG)**

Any individual activity by which learners directly received additional information about a patient's health status was coded as evidence generation. This included any clicking within the health record as well as reading the results from the radiological examination.

**Evidence elicitation (EE)**

An activity was coded as evidence elicitation whenever participants asked the agent-based radiologist to generate further evidence about a patient's health status. The specific activities included choosing a body part about whose status participants required further evidence as well as choosing a radiological examination (e.g., computer tomography [CT] scan) to examine the respective body part using the request form.

**Evidence sharing (ES)**

Anytime participants used the request form to share evidence about a patient's health status (e.g., main symptoms, course of the disease) with the agent-based radiologist to help them interpret the radiological evidence, an activity was coded as evidence sharing.

**Hypothesis sharing (HS)**

Anytime participants used the request form to share a differential diagnosis with the agent-based radiologist, an activity was coded as hypothesis sharing.

**Drawing conclusions (DC)**

Learners concluded a patient case by choosing a final diagnosis from a long menu containing over 200 entries. To do so, participants were asked to type in the initial letters of a diagnosis, after which matching entries popped up, and from which they could select a fitting diagnosis. In addition, participants were asked to justify their diagnosis using a free text field. This activity and the previous one were coded as drawing conclusions. The quality of the final diagnosis was used as an indicator of diagnostic accuracy.

**Diagnostic accuracy**

We used the final diagnoses proposed by the participants as indicators of diagnostic accuracy, which we used as an easy-to-interpret summative measure of diagnostic reasoning skills. The final diagnoses were coded by researchers from the learning sciences based on sample solutions developed by medical experts as either 1 (*correct*) or 0 (*incorrect*). Two trained raters independently coded 20% of the data set. They achieved perfect interrater agreement ( $ICC = 1$ ). The remaining data set was split in half, and each half was coded by one of the trained raters.



### Statistical analyses

All analyses described below were conducted in R 4.0.2 (R Core Team, 2020). The data sets, R script, and formulas are available from the open science framework (OSF) repository at [https://osf.io/2ne3y/?view\\_only=13ae84318f164875a67b7919cf85fd21](https://osf.io/2ne3y/?view_only=13ae84318f164875a67b7919cf85fd21).

### Feature extraction

To analyze participants' activities during the diagnostic reasoning process, the total time participants worked on the patient cases was split into seconds for each patient case. We logged the collaborative diagnostic activity that was being performed for each second. This procedure resulted in 490 individual strings of activities (98 participants with five patient cases each) with the length of the total time-on-task measured in seconds. Subsequently, 14 of the strings had to be removed due to missing values in the case solution, resulting in a final number of  $N = 476$  strings. For the subsequent feature extraction, we opted to apply an exploratory approach.

An approach that was created for applying an exploratory search of repetitive patterns within long sequences is the  $n$ -gram method (Damashek, 1995). The  $n$ -gram method summarizes a long string of entries (e.g., individual diagnostic steps in a diagnostic reasoning process) as sequences of  $n$  consecutive elements. To limit the number of features, we split the strings of activities into  $n$ -grams of length 2 (bigrams), using the "ngram" R package (Schmidt & Heckendorf, 2017), resulting in 25 variables, each representing the frequency of the occurrence of a unique combination of activities (see He & von Davier, 2016). More precisely, the resulting bigrams included two types: bigrams consisting of one activity (e.g., EE.EE) and bigrams consisting of two activities (e.g., EE.ES). The more frequently bigrams of two identical activities occurred, the more time was spent on that activity. The more frequently bigrams of two different activities occurred, the more frequently the transition from the first to the second activity occurred. Bigrams that occurred in only a maximum of one participant's string of activities were not included in the following analyses.

To identify bigrams that led to correct or incorrect diagnoses, we employed the Chi-Square feature selection model proposed by He and von Davier (2016). Using this approach, we conducted a weighted Chi-Square test for each bigram to determine whether its occurrence and nonoccurrence were independent for participants who came up with the correct versus the incorrect diagnosis. We used the weighted frequencies of the bigrams in correct and incorrect diagnoses to calculate whether the bigrams were more typical of correct or incorrect diagnoses (more details can be found in Oakes et al., 2001).

### Machine learning approaches

To investigate our research questions, we trained three different supervised machine learning models to classify whether a participant would provide the correct diagnosis for any specific patient on the basis of the bigrams. Specifically, we trained support vector machine (SVM) models with linear kernels, random forest (RF) models, and gradient boosting machine (GBM) models. We chose these models because they are widely used in educational data mining and are viewed, among others, as representatives of the state-of-the-art methods for predicting binary or categorical outcome variables inside

and outside of educational assessment (e.g., Costa et al., 2017; Fernández-Delgado et al., 2014; Qiao & Jiao, 2018). Detailed insights into the calculation principles (including formulas) can be found in Bonaccorso (2017).

SVMs classify data into two classes by finding the hyperplane that captures the largest distance between the data points in one class and those in the other class. The maximum width of the slab parallel to the hyperplane, which has no inner data points, is called the margin (Cortes and Vapnik, 1995). The data points at the left and right sides of the margin closest to the hyperplane (support vectors) are used as the starting point for maximizing the margin. With the help of the so-called kernel function, which is applied to the predictor variables, SVMs raise the variable space to a higher dimension and can thus also identify nonlinear relationships (Hilbert et al., 2021). Previous studies have shown that SVMs achieve better performance than other algorithms such as RFs or naïve bayes (e.g., Costa et al., 2017). Moreover, SVMs offer the advantage of being suitable for smaller data sets (Hussain et al., 2019). For the application of SVMs to our data set, we chose linear kernels to map linear relationships in the data in addition to nonlinear relationships that we captured with RFs and GBMs. RFs are based on decision trees and are used in classification and regression problems.

RFs constructs a certain number of single decision trees using random parts of the data to be classified. The procedure uses the test data on all constructed trees and assigns the most frequently occurring outcomes as labels to the test data (Breiman, 2001). As ensembles of single decision trees, RFs have advantages over single trees in terms of predictive power (Fernández-Delgado et al., 2014). Due to the large number of trees (law of large numbers), RFs barely overfit compared with single decision trees or other tree-based ensemble methods, such as GBMs (Breiman, 2001). Moreover, RFs are easier to tune and less time-consuming than GBMs, as well as easier to interpret than other supervised machine learning models, such as SVMs (Hilbert et al., 2021).

In contrast to RF models, which train trees independently, GBMs construct decision trees sequentially so that each new tree can help compensate for errors in previous trees (gradient descent method). By limiting the maximum number of leaves and splits, each decision tree acts as a weak learner (a model that performs slightly better than a random classifier/regressor) and does not dominate the prediction. GBM models allow high flexibility (Natekin & Knoll, 2013) and often achieve better performance than RFs (e.g., Qiao & Jiao, 2018) due to various hyperparameter options. Moreover, a strength of GBM models is that they can easily handle plenty of features and unbalanced data sets (Schröders et al., 2022).

#### **Model development and evaluation**

To train the models, we used the R packages “caret” (Kuhn, 2020), “ranger” (Wright & Ziegler, 2017), and “gbm” (Greenwell et al., 2020).

For all methods, the same data were used to train and test the models. First, we randomly split the data set into a training set (70% of the data) and a testing set (30% of the data). This resampling strategy is also called the holdout estimator (Pargent et al., 2022). The training set was then used to fit the predictive models. Unlike more conventional statistical models (e.g., linear regression), machine learning algorithms involve hyperparameters that have to be set before they are run (Probst et al., 2019). For SVM models with linear kernels, only



one hyperparameter (the cost value, which specifies how much the algorithm is “punished” for incorrect assignments) has to be tuned. The RF models were tuned to optimize minimal node size (the minimum number of data points required in any given node to split it), splitrule (gini or extra trees), and the number of predictors considered for splitting at each node (mtry). Important hyperparameters for GBM models include the basis of the number of trees (total number of trees in the ensemble), the interaction depth (maximum nodes per tree), the shrinkage (learning rate), and the minimal number of observations in a node (n.minobsinnode).

While training, the abovementioned hyperparameters were tuned automatically for each model on the basis of model performance using  $10 \times 3$  cross-validation (Fushiki, 2011). The cross-validation resulted in 30 iterations (10 folds, three repetitions) of training for each model, thus allowing us to determine the optimal hyperparameters and estimate the stability of each model to avoid over- or underfitting.

The optimal model was selected automatically for each of the algorithms on the basis of the largest kappa value (degree of agreement between the classifications and the real data, taking into account the agreement that occurred by chance). To check whether the diagnostic accuracy could be predicted on the basis of unseen data (RQ1), the optimal model was evaluated in the testing data set. To evaluate the algorithms, the classification accuracy (proportion of correct classifications out of all classifications), sensitivity (proportion of true classified correct diagnoses), specificity (proportion of true classified incorrect diagnoses), positive predictive value (PPV; proportion of true classified correct diagnoses out of all diagnoses classified as correct), negative predictive value (NPV; proportion of true classified incorrect diagnoses out of all diagnoses classified as incorrect), and F1 value (weighted average of sensitivity and positive predictive value) were calculated in addition to kappa.

The algorithm with the best average kappa value resulting from the cross-validation (training phase) was selected for further analysis and interpretation. For this model, we estimated the relative importance (Chen et al., 2020) of each bigram with the R package “caret” (Kuhn, 2020), which indicates how each feature affected the model’s performance (total classification accuracy). The higher the variable importance score, the more important the feature was for the overall prediction (Fisher et al., 2019). This provided some measure of how relevant any specific combination of activities was for the total prediction in relation to the others but could not be interpreted concerning size or direction. Machine learning models can become highly complex and are therefore sometimes referred to as black boxes (Yarkoni and Westfall, 2017), which make it difficult to interpret the individual contribution of each feature. However, for this study, we were mainly interested in the total prediction rather than in individual feature interpretation.

To address RQ2, the algorithm was then applied to 10 subsets of the original complete data, created by splitting the first 1200 s of the total processing time into time intervals of 120 s before extracting the features (bigrams). The data sets contained the behaviors (bigrams) that participants exhibited at the corresponding time points.

## Results

### Descriptive statistics

Table 1 presents the numbers of incorrect and correct diagnoses across the behavioral strings of physicians and medical students. Physicians and medical students came up



**Table 1** Distributions of Incorrect and Correct Diagnoses Across Behavioral Strings of Physicians and Medical Students

	Number of behavioral strings		Total
	Incorrect diagnoses	Correct diagnoses	
Physicians	34	91	125
Medical students	128	223	351
Total	162	314	476

with correct diagnoses in 73% and 64% of the cases, respectively. However, this difference was not statistically significant,  $\chi^2(1) = 3.52$ ,  $p = .061$ . Overall, there was a higher proportion of correct diagnoses.

#### Research question 1

To investigate whether diagnostic accuracy could be predicted from observed behavior (RQ1), we first took a closer look at differences in the CDAs between the incorrect and correct diagnoses.

Table 2 summarizes the numbers of strings of incorrect and correct diagnoses in which the bigrams occurred and the total frequencies in those strings. The three bigrams that occurred in only one string of activities in either correct or incorrect diagnoses (HS.DC, DC.ES, and DC.HS) were excluded from the following analyses, leaving a total of 22 bigrams. Further, Table 2 presents the results of the Chi-Square feature selection model, which shows the differences in the probabilities of the bigrams for participants who correctly diagnosed the patient case and those who did not. Bigrams with higher Chi-Square values were better at discriminating between the two groups.

When looking at the bigrams with only one activity (i.e., the bigrams that indicated how much time was spent on that activity), the bigram DC.DC (i.e., spending more time drawing conclusions) was by far the most discriminative bigram for participants who gave an incorrect diagnosis versus those who gave a correct diagnosis. Spending more time drawing conclusions occurred more often among participants who gave a correct diagnosis. Next was EE.EE (spending more time eliciting evidence), which was more typical of participants who gave an incorrect diagnosis, followed by HS.HS (spending more time sharing hypotheses) and EG.EG (spending more time generating evidence), both of which were more typical of participants who gave a correct diagnosis. For the bigrams with two activities (i.e., the bigrams that indicated more frequent transitions from the first to the second activity), EE.EG (switching back from the radiological request to the health record or to reading radiological test results), ES.EE, and HS.EE (both representing setbacks during the radiological request) were the most discriminative behaviors, all of which were more typical of participants who submitted an incorrect final diagnosis. Moreover, both switching between submitting the final diagnosis and requesting the agent-based radiologist (DC.EE, EE.DC, ES.DC) and studying the health record (DC.EG) were among the most discriminative behaviors, all of which were more typical of participants who gave an incorrect diagnosis. All of the described bigrams were statistically significantly able to discriminate between the two groups.

**Table 2** Frequency of Occurrence of Bigrams in Incorrect and Correct Diagnoses

Bigram	Frequency In strings		Weight	Total frequency of bigrams				Chi-Square test		
	Incorrect diagnoses	Correct diagnoses		Incorrect diagnoses		Correct diagnoses		$\chi^2$	p	Dir
				Raw	Wgt	Raw	Wgt			
EG.EG	162	314	0.03	71,931	410.50	110,662	631.53	144.17	<.001	+
EG.EE	159	312	0.04	405	8.83	521	11.36	0.11	.735	-
EG.ES	38	47	2.27	49	110.59	54	121.87	49.80	<.001	-
EG.HS	14	21	3.04	18	54.56	29	87.91	31.62	<.001	+
EG.DC	156	309	0.06	230	9.63	376	15.74	6.44	.011	+
EE.EG	82	81	1.79	113	200.39	83	147.19	766.49	<.001	-
EE.EE	162	313	0.03	11,727	113.68	8801	85.32	403.93	<.001	-
EE.ES	151	283	0.17	360	57.84	464	74.55	0.67	.414	-
EE.HS	48	60	2.06	73	149.57	77	157.76	101.46	<.001	-
EE.DC	9	3	3.33	9	29.93	3	9.98	410.76	<.001	-
ES.EG	70	82	1.63	87	141.08	97	157.29	56.34	<.001	-
ES.EE	54	39	2.22	74	163.26	54	119.14	633.63	<.001	-
ES.ES	157	298	0.14	29,944	3588.27	39,799	4769.22	0.43	.514	-
ES.HS	146	280	0.20	301	57.41	464	88.49	19.51	<.001	+
ES.DC	6	2	3.39	6	20.27	2	6.76	278.01	<.001	-
HS.EG	147	270	0.24	265	59.37	423	94.77	31.10	<.001	+
HS.EE	54	41	2.16	68	146.03	48	103.08	620.17	<.001	-
HS.ES	52	78	1.81	58	104.47	97	174.71	88.15	<.001	+
HS.HS	159	311	0.06	14,990	537.64	23,578	845.66	257.95	<.001	+
HS.DC	1	4	3.37	1	3.36	4	13.45	89.06	<.001	+
DC.EG	46	60	2.12	83	174.87	85	179.08	149.63	<.001	-
DC.EE	10	4	3.34	11	36.68	4	13.34	462.98	<.001	-
DC.ES	1	2	3.26	1	3.25	2	6.50	9.10	.003	+
DC.HS	0	2	3.10	0	0.00	2	6.19	114.52	<.001	+
DC.DC	160	313	0.04	20,863	441.81	40,842	864.89	1203.06	<.001	+

Note. EG = Evidence generation, EE = Evidence elicitation, ES = Evidence sharing, HS = Hypothesis sharing, DC = Drawing conclusions. Higher Chi-Square values indicate more discriminative bigrams. Dir = Direction of the difference in the occurrence of bigrams between learners who diagnosed the case correctly and those who diagnosed the case incorrectly, "+" represents a more frequent occurrence of the bigram in the strings of learners who correctly diagnosed the case, "-" represents a more frequent occurrence of the bigram in the strings of learners who incorrectly diagnosed the case

**Table 3** Mean Classification Accuracy and Kappa From the Cross-Validation for All Algorithms

Measures	SVM	RF	GBM
Mean accuracy	.73	.75	.74
CI <sub>Accuracy</sub>	[.71-.76]	[.70-.79]	[.70-.76]
Mean kappa	.33	.37	.36
CI <sub>kappa</sub>	[.24-.42]	[.31-.49]	[.30-.43]

Note. CI 95% confidence interval

Subsequently, we trained three different machine learning models to classify whether a participant would provide the correct diagnosis for any specific patient case on the basis of the 22 remaining bigrams. Table 3 summarizes the results for all models from the training phase (cross-validation) by presenting the average classification



accuracy and kappa across all 30 repetitions. Generally, the different model iterations did not differ much, thus suggesting no substantial overfitting. All algorithms showed significantly higher average classification accuracy than the no information rate (NIR), which indicates how many observations out of all observations would have been correctly classified if only the label “correct diagnosis” (the larger class) would have been assigned. The NIR of .66 corresponds to the proportion of all correct diagnoses in all observations (see Table 1). Considering an ideal NIR of .50 (equally distributed classes; Batista et al., 2004), .66 deviates somewhat from this value but does not indicate a substantial skewness in favor of one of the classes. Beyond accuracy, the algorithms reached acceptable kappa values (Fleiss et al., 2003). Moreover, the models did not differ significantly in their average classification accuracy values,  $F(2, 87) = 0.56, p = .559, \eta^2 = .01$ , or in their average kappa values,  $F(2, 87) = 0.72, p = .491, \eta^2 = .02$ . However, since the RF showed descriptively a slightly better average kappa, it was selected to finally answer RQ1 and RQ2.

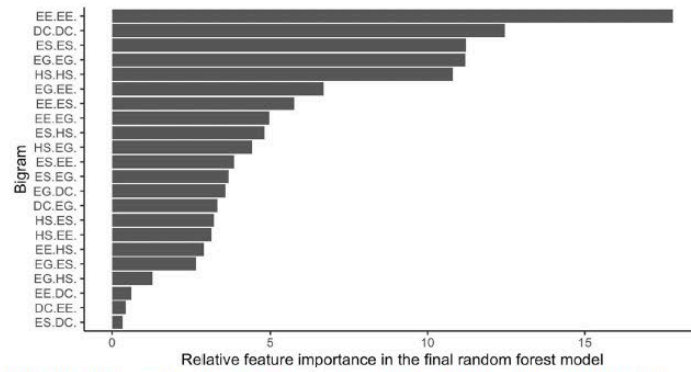
Table 4 presents the evaluation results of all algorithms in the testing data set. As can be seen, RF (final tuning parameters: min node size = 1, mtry = 2, and splitrule = gini), GBM (final tuning parameters: n.trees = 50, interaction.depth = 1, shrinkage = 0.1, and n.minobsinnode = 10), and SVM (final tuning parameter: cost value = 0.25) all achieved significantly higher classification accuracy than the NIR as well as acceptable kappa values (Fleiss et al., 2003). Strikingly, all models showed high sensitivity, and good PPV and F1 values but rather low specificity, indicating that correct diagnoses were substantially better predicted than incorrect diagnoses. However, all models reached acceptable NPV values, indicating precision in classifying incorrect diagnoses (many of the diagnoses classified as “incorrect” were indeed incorrect diagnoses). Overall, the algorithms did not differ greatly in their performance. The final selected algorithm, the RF model, achieved acceptable to good values on all measures (classification accuracy = .75, kappa = .40, sensitivity = .90, specificity = .46, PPV = .77, NPV = .71, and F1 = .83) and was therefore selected for further interpretation and analyses.

Figure 2 illustrates the bigrams’ relative importance in the RF model. By far most important for the overall prediction was how much time was spent eliciting evidence (EE.EE) followed by the amount of time spent drawing conclusions (DC.DC), sharing evidence (ES.ES), generating evidence (EG.EG), and sharing hypotheses (HS.HS).

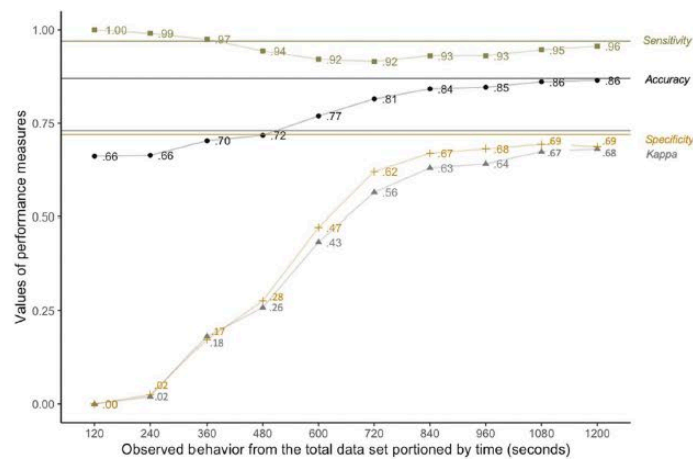
**Table 4** Results of the Evaluation of the Algorithms in the Testing Data Set

Measures	SVM	RF	GBM
NIR (.66)			
Acc	.75	.75	.74
p-value [Acc > NIR]	.012	.011	.029
Kappa	.39	.40	.40
Sensitivity	.91	.90	.84
Specificity	.44	.46	.54
PPV	.76	.77	.78
NPV	.72	.71	.63
F1	.83	.83	.81

Note. NIR = Proportion of correct diagnoses in all observations, Acc = Classification accuracy, PPV = Positive predictive value, NPV = Negative predictive value



**Fig. 2** Relative Importance of Each Bigram for the Final RF Model. EG Evidence generation, EE Evidence elicitation, ES Evidence sharing, HS Hypothesis sharing, DC Drawing conclusions



**Fig. 3** Performance Measures for the Random Forest Model Applied to Increasing Amounts of Data. The horizontal lines represent the final values for the RF algorithm based on the original complete data set

Moreover, the analysis revealed that the most important bigrams with two activities were the frequency of switching between evidence generation and evidence elicitation (EG.EE; EE.EG) as well as the frequency of transitions from evidence elicitation to evidence sharing (EE.ES).

**Research question 2**

To investigate how early during diagnosing it is possible to reliably predict diagnostic accuracy on the basis of CDAs (RQ2), we applied the final RF model to a sequence of

subsets of the complete data that included only the actions observed in the first 120 to 1200 s. As can be seen in Fig. 3, classification accuracy, kappa, sensitivity, and specificity approximated the values estimated for the complete data (horizontal lines) after 1200 s. In the first 120 s, the model did not perform better than the NIR of .66 (classification accuracy = .66, sensitivity = 1, kappa = 0, specificity = 0). From second 240, the performance slowly increased and asymptotically approached the final values in the complete data set. More precisely, in second 360, the accuracy exceeded the NIR until it reached approximately its final value in the complete data set at second 1200 with 0.86. Similarly, the kappa value increased over time (largest increase between seconds 600 and 840). At second 120, the RF began with a sensitivity (correct classification of correct diagnoses) of 1 (100%) because, in the beginning, the model classified all observations as "correct." Up to second 720, the sensitivity slowly decreased, while kappa and specificity increased, until sensitivity approximately reached its final value in the complete data set with .96 after 1200 s. By contrast, at second 120, the model began with a specificity (correct classification of incorrect diagnoses) of 0 (0%) but approximately approached the final value over time with .69. Overall, it can be seen from the graph that the model's performance took on acceptable predictive values from about second 840. Correct diagnoses could be predicted particularly well after 600 s (10 min) or after two thirds (66%) of the median time (15 min) had been spent on the patient case.

### Discussion

This study examined the extent to which and how quickly diagnostic accuracy could be predicted from learners' engagement in CDAs based on log file data from a medical simulation with the help of machine learning. Three different classification algorithms (SVM, RF, GBM) reached acceptable overall prediction quality. Due to slightly better performance, the RF model was selected for further interpretation and analysis of how early it is possible to achieve a reliable prediction of diagnostic accuracy during diagnosing on the basis of CDAs. The results showed that after approximately two thirds of the median time learners spent on the diagnostic task, the RF algorithm was able to reliably predict diagnostic success. Moreover, the time spent on CDAs was especially important for predicting diagnostic accuracy and was the best at distinguishing between correct and incorrect diagnoses. While spending more time engaged in individual activities (e.g., generating evidence and drawing conclusions) was more typical of successful diagnosticians, spending more time engaged in collaborative activities (e.g., eliciting and sharing evidence; i.e., interaction with the agent-based radiologist) tended to be behavior that was more typical of unsuccessful diagnosticians. These findings are aligned with previous work that showed somewhat similar results in the context of complex problem solving. For example, Stadler et al. (2019) found that successful problem solvers spent more time reflecting on the task (i.e., they spent more time drawing conclusions), whereas unsuccessful problem solvers spent more time performing activities that involved gathering information. However, the equivalent results for unsuccessful diagnosticians in the context of our simulation apply only to collaborative engagement with the evidence (i.e., spending more time eliciting and sharing evidence as opposed to spending more time generating evidence).



Previous research found that time spent on tasks was moderated by prior knowledge level (e.g., Goldhammer et al., 2014). Our study adds to this line of research by qualifying the types of activities within the task. Considering the MOT model (Charlin et al., 2012), in contrast to unsuccessful diagnosticians, successful diagnosticians should be able to identify early case cues, have more specific initial representations, and be better able to determine the relevant objectives of the encounter. Applied to our simulation, when successful diagnosticians have a concrete suspected diagnosis, they are able to make a more specific radiological request that they know will help them find support for or falsify their diagnosis. As a consequence, they consult the radiologist less often and elicit less evidence. Instead, they spend more time carefully processing the information from the health record and radiological test results, and at the end, they spend more time drawing conclusions before settling on a final diagnosis. On the other hand, unsuccessful diagnosticians might have trouble identifying early cues in the patient case and determining the appropriate objectives of the patient encounter (Bowen, 2006). Compared with diagnosticians who have a proper initial patient representation, they urgently require further radiological information to be able to diagnose the case but might have trouble further processing this large amount of weakly organized information (Stadler et al., 2019), as they lack a proper initial representation. Thus, these diagnosticians have trouble making optimal use of collaboration as a source of information (Radkowsitch et al., 2022) because they have both no clue about what additional information to look for in the patient and problems with sharing *relevant* information with the collaboration partner (Tschan et al., 2009), leading to an increasing amount of time spent selecting appropriate examinations and sharing evidence from the health record. This interpretation would be supported by the frequent transitions and setbacks typically encountered by unsuccessful diagnosticians while working in the simulation. One reason for frequent transitions within the radiological request is that these diagnosticians request a larger number of examinations, supporting the assumption that they have a greater need for additional radiological evidence. Diagnosticians who displayed frequent switches from the radiological request form to the health record may have lacked a concrete idea about the patient's problem at that time, had several possible suspected diagnoses in mind, and were unable to retain information from the health record in their working memory while simultaneously implementing the requirements of collaboration. Further, switching back and forth between submitting the final diagnosis (drawing conclusions) and dealing with evidence by either requesting the radiologist and studying the health record (generating and eliciting evidence) or sharing patient information with the radiologist (sharing evidence) are typical behaviors of unsuccessful diagnosticians. This finding most likely indicates that these diagnosticians have problems using the evidence appropriately to validate or exclude a particular hypothesis from their set of suspected hypotheses (evidence evaluation).

Notably, the Chi-Square feature selection model revealed that the above described transitions from one CDA to another and switching between CDAs, both of which are related to incorrect diagnoses, better distinguish between successful and unsuccessful diagnosticians than the time spent on these activities. However, in the RF model, the time spent on CDAs was clearly most important for the overall prediction. Thus, we assume that beyond the Chi-Square test, the prediction of the RF model may have



revealed additional nonlinear relationships between CDAs and diagnostic accuracy (black box problem; Yarkoni and Westfall, 2017).

Taken together, these findings on the differences between successful and unsuccessful diagnosticians suggest that, at least in the context of our simulation, an adequate initial representation of the case is crucial for diagnostic success. The information on adequate or inadequate initial representations of the case could be used to provide adaptive process-based feedback on whether learners are heading toward correct diagnoses. On the other hand, an inadequate representation can hardly be compensated for by subsequent collaboration with the agent-based radiologist. Thus, in the context of our simulation, the agent tended not to be helpful to diagnosticians who were on the wrong track. Further, deviations from the intended structure of the simulation were more likely to be indicators of misdiagnoses, thus applying to a wide range of expertise. However, referring to the high sensitivity but low specificity achieved by our model, we were able to reliably predict correct diagnoses better and earlier than incorrect ones. We assume that one reason for the low specificity compared with the high sensitivity is that in our sample successful diagnosticians may not differ in their behavior as much as unsuccessful diagnosticians. After reading the health record, successful diagnosticians enter the collaboration with an adequate mental representation, through which they can make targeted radiologic requests to reduce diagnostic uncertainty regarding suspected diagnoses, and solve the diagnostic case correctly. In contrast, the misdiagnoses of unsuccessful diagnosticians could be due to cognitive misbehavior of various causes, which manifests itself at the simulation level in different behavior. For example, recent analyses on the behavior after impasses in the context of the same simulation show that diagnosticians differ in their success in identifying and subsequently compensating for errors in the diagnostic reasoning process (Heitzmann et al., 2023). Future research may follow this line of research and examine the behaviors that lead to an incorrect diagnosis in more detail.

Our study represents a “proof of concept” for one way in which the prediction of successful and unsuccessful diagnosticians using the behavior displayed in the simulation could be used in microadaptive learning environments. Yet, further research will be necessary. Early predictions of learners heading toward a correct diagnosis can inform instructors and educators to remove instructional support in real time before it has negative effects on learning (Kalyuga et al., 2003). Our prediction of correct diagnoses was successful only after two thirds of the diagnostic reasoning process and thus cannot necessarily be considered an early prediction, for example, as shown by Ulitzsch et al. (2022), when they used only about one third of their examined clickstream data in the context of complex problem-solving. However, because we obtained the information on diagnostic success before learners completed the diagnostic task, it is still possible to adjust the task difficulty in real time or in the upcoming task (Roosevelt, 2008) to address learners’ zone of proximal development (Vygotsky, 1978). Moreover, to increase the likelihood of building a correct initial case representation that prepares and pre-structures the individual diagnostic reasoning process for collaborating with the agent-based radiologist, learners could receive prompts that remind them to review the health record and radiological test results properly and help them integrate the information into hypotheses. Conceivable types of scaffolding may be reflection prompts (Mamede

and Schmidt, 2017), which encourage learners to reflect on the evidence they generated in terms of potential hypotheses.

#### Limitations and further research

In interpreting these findings, there are some limitations to be considered. The first relates to the prediction of diagnostic success after beginning the diagnostic reasoning process, which was possible only after 10 min because the behavior in the earlier minutes was probably not diverse enough.

The reason for this finding can be seen in the rather coarse granulation level of the coded log files of the CDAs, which might not have been fine enough to identify early subtle differences in the behaviors of successful and unsuccessful diagnosticians. However, the use of broad diagnostic indicators is also one of the strengths of this study, as they can be applied to other diagnostic contexts for generalization at a low threshold. Nevertheless, future process analyses could investigate diagnostic behavior at finer coding levels to uncover further latent differences between successful and unsuccessful diagnosticians.

Second, at least to some extent, the use of bigrams limited the insights that could have been gained about the behavior of successful and unsuccessful diagnosticians if trigrams (e.g., EE.ES.HS), which would have included two transitions, had been used. Alternatively, unigrams (e.g., EE) might have been interpretable in a more straightforward way. However, trigrams would have extensively increased the number of possible features ( $k=125$ ), and unigrams would have indicated only the time spent on CDAs without considering transitions from one to another. To verify our choice of bigrams, we repeated the Chi-Square test with trigrams to control for possible significant sequences of two transitions. We found that the ranking of the most important indicators of diagnostic success and failure did not change such that, for each strong discriminative bigram (e.g., EE.EG.), both possible trigrams (EE.EE.EG; EE.EG.EG) discriminated equally well. Interested readers can find these analyses on the OSF. Moreover, our approach to feature extraction did not consider participants' pauses between activities, even though pausing behavior may provide a valuable source of information (e.g., Tenison and Arslan, 2020). Pausing behavior, for instance, may indicate reflective thinking about the diagnostic reasoning process or may be linked to behavioral responses following errors or impasses. Since the n-gram approach is not necessarily the best one to capture pausing behavior, approaches more appropriate for timing data may be considered in future research.

Third, we did not consider case difficulty, case typicality, or the prior knowledge or expertise level of diagnosticians in our prediction models. However, the fact that our algorithm was able to reliably predict diagnostic accuracy across different cases and expertise levels is a strong sign of robustness. Further, another study with the same tasks found that changes in difficulty across tasks led to changes in time on task regardless of participants' level of expertise (Stadler et al., 2021), further supporting their equivalence in typicality. However, our interpretations of the behavior of successful versus unsuccessful diagnosticians were mainly valid for cases in which early cues already pointed to the correct diagnosis (typical cases). The extent to which the algorithms can predict similar results exclusively for atypical cases needs to be investigated in further studies.



Moreover, the present analysis focused on diagnostic accuracy and not on learning as a change in knowledge and skills. It is possible that our participants who “gambled the radiologist” by sharing and requesting a lot of information may be among those who still failed to reach a correct conclusion but still learned a lot from the simulation. Exploring complex problem-solving tasks with the goal of finding out as much as possible, without the goal of establishing a well-supported solution or diagnosis may be an effective approach to learning, as it is connected to lower cognitive load (goal-free instruction; Sweller et al., 2019). Finally, the study participants in our setting interacted with an agent. A recent study found no differences between agents and human collaborators in the assessment of collaborative problem solving in PISA (Herborn et al., 2020), yet agent-based collaboration carries the risk of being a poor substitute for natural collaboration. However, we chose agent-based collaboration for one significant advantage: In contrast to human-to-human collaboration, it enabled the standardized measurement of collaborative diagnostic reasoning processes by holding the agent’s behavior and knowledge level constant. In addition, the simulation’s interface (request form) and its structure were carefully developed by learning scientists and medical experts on the basis of real clinical situations in which an internist collaborates with a radiologist, who serves as a potential additional source of evidence, to reduce further diagnostic uncertainty. Yet, future research should address the transfer to human-to-human collaboration in diagnostic settings.

### Conclusion

Even though having the competence to provide a correct diagnosis collaboratively is relevant in many domains, the fostering of collaborative diagnostic reasoning has yet to be thoroughly investigated. Simulations with dynamic individual learning support are a promising approach for fostering such complex skills. The present study identified behavioral characteristics for successful and unsuccessful diagnosticians in a collaborative medical training simulation based on CDAs—broad theoretical indicators that can be found in various diagnostic contexts. We used these indicators to develop a model that enabled a reliable and robust prediction of diagnostic accuracy across diagnosticians with varying expertise levels and different diagnostic cases. The study provides preliminary evidence that (a) the individual diagnostic reasoning process controls the collaborative diagnostic reasoning process and is thus crucial for overall diagnostic success and that (b) diagnostic success can be predicted better than diagnostic failure, and after only 66% of the average time spent on the diagnostic case, which might be due to the fact that diagnostic failure underlies more heterogeneous behavior than diagnostic success.

Our study is an example of how log-file-based process data analyses could be further used in adaptive learning environments to individually foster collaborative diagnostic reasoning skills in a targeted manner. These insights can open up new ways to conduct collaborative diagnostic training both within and outside of higher education.

### Abbreviations

CDAs	Collaborative diagnostic activities
DC	Drawing conclusions
EE	Evidence elicitation

EG	Evidence generation
ES	Evidence sharing
GBM	Gradient boosting machine
HS	Hypothesis sharing
OECD	Organization for Economic Co-operation and Development
OSF	Open Science Framework
RF	Random forest
SDDS	Scientific discovery as dual search
SVM	Support vector machine

#### Acknowledgements

Not applicable.

#### Author contributions

CR: writing—original draft, conceptualization, methodology, project administration. MS: methodology, formal analysis, writing—original draft, writing—review and editing, conceptualization. AR: resources, writing—original draft, writing—review and editing, conceptualization. RS: funding acquisition, conceptualization, writing—review and editing. MF: funding acquisition, conceptualization, writing—review and editing. FF: conceptualization, funding acquisition, supervision, writing—review and editing. All authors read and approved the final manuscript.

#### Funding

This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, FOR 2385) for a subproject conducted by the CoSiMed research group (FI 792/11-1).

#### Availability of data and materials

The datasets analysed during the current study are available in the open science repository (OSF), [https://osf.io/2ne3y/?view\\_only=13ae84318f164875a67b7919cf85fd21](https://osf.io/2ne3y/?view_only=13ae84318f164875a67b7919cf85fd21).

#### Declarations

##### Ethics approval and consent to participate

Ethical clearance was declared by the Ethics Committee at the Medical Faculty of LMU Munich prior to data collection.

##### Consent for publication

The authors consent to the publication of the manuscript in *Large-scale Assessments in Education*.

##### Competing interests

The authors declare no competing interests.

Received: 13 January 2022 Accepted: 4 January 2023

Published online: 20 January 2023

#### References

- Al-Kadi, A. S., & Donnon, T. (2013). Using simulation to improve the cognitive and psychomotor skills of novice students in advanced laparoscopic surgery: a meta-analysis. *Medical Teacher*, 35(sup1), S47–S55. <https://doi.org/10.3109/0142159X.2013.765549>
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics* (pp. 61–75). Springer. [https://doi.org/10.1007/978-1-4614-3305-7\\_4](https://doi.org/10.1007/978-1-4614-3305-7_4)
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. <https://doi.org/10.1145/1007730.1007735>
- Bonaccorso, G. (2017). *Machine learning algorithms: A reference guide to popular algorithms for data science and machine learning*. Packt Publishing.
- Bowen, J. L. (2006). Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21), 2217–2225. <https://doi.org/10.1056/NEJMr054782>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Charlin, B., Lubarsky, S., Millette, B., Crevier, F., Audétat, M.-C., Charbonneau, A., Caire Fon, N., Hoff, L., & Bourdy, C. (2012). Clinical reasoning processes: Unravelling complexity through graphical representation. *Medical Education*, 46(5), 454–463. <https://doi.org/10.1111/j.1365-2923.2012.04242.x>
- Chen, R.-C., Dewi, C., Huang, S.-W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*. <https://doi.org/10.1186/s40537-020-00327-4>
- 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>
- Cirigliano, M. M., Guthrie, C. D., & Pusic, M. V. (2020). Click-level learning analytics in an online medical education learning platform. *Teaching and Learning in Medicine*, 32(4), 410–421. <https://doi.org/10.1080/10401334.2020.1754216>
- Cook, D. A., Hamstra, S. J., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., Erwin, P. J., & 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>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.



- Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73, 247–256. <https://doi.org/10.1016/j.chb.2017.01.047>
- Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199), 843–848. <https://doi.org/10.1126/science.267.5199.843>
- Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, 79(10), S70–S81. <https://doi.org/10.1097/00001888-200410001-00022>
- Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1), 3133–3181.
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dörner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, 2(3), 28–45. <https://doi.org/10.14786/flr.v2i2.96>
- Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81. <http://jmlr.org/papers/v20/18-760.html>
- Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Wiley. <https://doi.org/10.1002/0471445428>
- Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21(2), 137–146. <https://doi.org/10.1007/s11222-009-9153-8>
- 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>
- Goldhammer, F., Naumann, J., Stelter, A., Töth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106(3), 608–626. <https://doi.org/10.1037/a0034716>
- Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, 19(2), 59–92. <https://doi.org/10.1177/1529100618808244>
- Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers. (2020). *Package gbm* (Version 2.1.8) [Computer software]. <https://cran.r-project.org/web/packages/gbm/gbm.pdf>
- Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61, 36–46. <https://doi.org/10.1016/j.chb.2016.02.095>
- Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence*, 50, 100–113. <https://doi.org/10.1016/j.intell.2015.02.007>
- Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills*. Dordrecht: Springer. <https://doi.org/10.1007/978-94-017-9395-7>
- 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.
- Hall, D., & Buzwell, S. (2012). The problem of free-riding in group projects: Looking beyond social loafing as reason for non-contribution. *Active Learning in Higher Education*, 14(1), 37–49. <https://doi.org/10.1177/1469787412467123>
- He, Q., & Von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749–776). IGI Global. <https://doi.org/10.4018/978-1-4666-9441-5>
- Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., Schmidmaier, R., Neuhaus, B., 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(4), 1–24. <https://doi.org/10.14786/flr.v7i4.384>
- Heitzmann, N., Stadler, M., Richters, C., Radkowsky, A., Schmidmaier, R., Weidenbusch, M., & Fischer, M. R. (2023). Learners' adjustment strategies following impasses in simulations—effects of prior knowledge. *Learning and Instruction*. <https://doi.org/10.1016/j.learninstruc.2022.101632>
- Herborn, K., Stadler, M., Mustafic, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*. <https://doi.org/10.1016/j.chb.2018.07.035>
- Hilbert, S., Coors, S., Kraus, E. B., Bischl, B., Frei, M., Lindl, A., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*. <https://doi.org/10.1002/rev3.3310>
- Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). 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>
- Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., & Ali, S. (2019). Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review*, 52(1), 381–407. <https://doi.org/10.1007/s10462-018-9620-8>
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38(1), 23–31. [https://doi.org/10.1207/s15326985EP3801\\_4](https://doi.org/10.1207/s15326985EP3801_4)
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12(1), 1–48. [https://doi.org/10.1207/s15516709cog1201\\_1](https://doi.org/10.1207/s15516709cog1201_1)
- Kuhn, M. (2020). *caret: Classification and Regression Training* (Version 6.0–86) [Computer software]. <https://CRAN.R-project.org/package=caret>
- Landriscina, F. (2012). Simulation and learning The role of mental models. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning*. Springer. [https://doi.org/10.1007/978-1-4419-1428-6\\_1874](https://doi.org/10.1007/978-1-4419-1428-6_1874)
- Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, J. D. (2015). A tough nut to crack. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Advances in higher education and professional development (AHEPD) book series. Handbook of research on technology tools for real-world skill development*. IGI Global. <https://doi.org/10.4018/978-1-4666-9441-5.ch013>



- 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>
- Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. <https://doi.org/10.3389/fnbot.2013.00021>
- 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>
- Oakes, M., Gaizauskas, R., Fowkes, H., Jonsson, A., Wan, V., & Beaulieu, M. (2001). A method based on the chi-square test for document classification. In D. H. Kraft, W. B. Croft, D. J. Harper, & J. Zobel (Eds.), *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 440–441). ACM Press. <https://doi.org/10.1145/383952.384080>
- OECD. (2017). PISA 2015 Assessment and analytical framework: Science, reading, mathematics, financial literacy and collaborative problem solving. *PISA, OECD Publishing*. <https://doi.org/10.1787/9789264281820-en>
- O’Neil, H. F., Chuang, S.-H., & Chung, G. K. W. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *National Center for Research on Evaluation, Standards, and Student Testing*, 10(3), 361–373. <https://doi.org/10.1080/0969594032000148190>
- O’Neill, T. A., Allen, N. J., & Hastings, S. E. (2013). Examining the “Pros” and “Cons” of Team Conflict: A Team-Level Meta-Analysis of Task, Relationship, and Process Conflict. *Human Performance*, 26(3), 236–260. <https://doi.org/10.1080/08959285.2013.795573>
- Pargent, F., Schoedel, R., & Stachl, C. (2022). An introduction to machine learning for psychologists in R. *PsyArXiv*. <https://doi.org/10.31234/osf.io/89snd>
- Pauli, R., Mohiyeddini, C., Bray, D., Michie, F., & Street, B. (2008). Individual differences in negative group work experiences in collaborative student learning. *Educational Psychology*, 28(1), 47–58. <https://doi.org/10.1080/01443410701413746>
- Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences*, 13(3), 423–451. [https://doi.org/10.1207/s15327809jls1303\\_6](https://doi.org/10.1207/s15327809jls1303_6)
- Plass, J. L., & Pawar, S. (2020). 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>
- Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(1), 1–32. <https://www.jmlr.org/papers/volume20/18-444.pdf>
- Qiao, X., & Jiao, H. (2018). Data Mining Techniques in Analyzing Process Data: A Didactic. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2018.02231>
- Radkowsitch, A., Sailer, M., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2022). Diagnosing collaboratively: A theoretical model and a simulation-based learning environment. In F. Fischer & A. Opitz (Eds.), *Learning to diagnose with simulations: Teacher education and medical education* (pp. 123–141). Springer Nature. <https://doi.org/10.1007/978-3-030-89147-3>
- Richter, M. M., & Weber, R. O. (2013). Case-Based Reasoning. *Springer*. <https://doi.org/10.1007/978-3-642-40167-1>
- R Core Team. (2020). *R: A Language and environment for statistical computing* (Version R4.0.2) [Computer software]. <https://www.R-project.org/>
- Roosevelt, F. D. (2008). Zone of proximal development. In N. J. Salkind (Ed.), *Encyclopedia of educational psychology* (pp. 1017–1022). SAGE Publications. <https://doi.org/10.4135/9781412963848.n282>
- Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O. Malley (Ed.), *Computer supported collaborative learning* (pp. 69–97). Springer. [https://doi.org/10.1007/978-3-642-85098-1\\_5](https://doi.org/10.1007/978-3-642-85098-1_5)
- San Pedro, M., Baker, R. S., Bowers, A. J., & Heffernan, N. T. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In S. D’Mello R. Calvo, & A. Oldey (Eds.), *Proceedings of the 6th international conference on educational data mining* (pp. 177–184).
- Schmidt, D., & Heckendorf, C. (2017). *Guide to the ngram package: Fast n-gram tokenization* (Version 3.0.4) [Computer software]. <https://cran.r-project.org/package=ngram>
- Schröders, U., Schmidt, C., & Gnams, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, 82(1), 29–56. <https://doi.org/10.1177/00131644211004708>
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, 55(2), 503–524.
- 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*. <https://doi.org/10.3389/fpsyg.2019.00777>
- Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. *Computers in Human Behavior*. <https://doi.org/10.1016/j.chb.2020.106442>
- Stadler, M., Radkowsitch, A., Schmidmaier, R., Fischer, M., & Fischer, F. (2021). Take your time: Invariance of time-on-task in problem-solving tasks across expertise levels. *Psychological Test and Assessment Modeling*, 65(4), 517–525.
- Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292.
- Tenison, C., & Arslan, B. (2020). Characterizing pause behaviors in a science inquiry task. In T. C. Stewart (Ed.), *Proceedings of the 18th International Conference on Cognitive Modeling* (pp. 283–298). Applied Cognitive Science Lab.
- Tomasovic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*. <https://doi.org/10.1016/j.compedu.2019.103676>
- Tschan, F., Semmer, N. K., Gurtner, A., Bizzari, L., Spychiger, M., Breuer, M., & Marsch, S. U. (2009). Explicit reasoning, confirmation bias, and illusory transactive memory: A simulation study of group medical decision making. *Small Group Research*, 40(3), 271–300. <https://doi.org/10.1177/1046496409332928>
- Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2022). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-022-01844-1>

- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. <https://doi.org/10.1080/00461520.2011.611369>
- Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press. <https://doi.org/10.2307/j.ctv9f9vz4>
- Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17. <https://doi.org/10.18637/jss.v077.i01>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. <https://doi.org/10.1177/1745691617693393>
- Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment: Network analysis for process data. *Journal of Educational Measurement*, 53(2), 190–211. <https://doi.org/10.1111/jedm.12107>
- Ziv, A., Wolpe, P. R., Small, S. D., & Glick, S. (2003). Simulation-based medical education: An ethical imperative. *Academic Medicine*, 78(8), 783–788. <https://doi.org/10.1097/00001888-200308000-00006>

#### Publisher's Note

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

Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

## **5 General Discussion**

*Constanze Catharina Richters*

The present dissertation pursued the overarching goal of establishing foundations of adaptive instructional support for learning collaborative diagnostic reasoning. Two subgoals were addressed: to investigate the conditions under which (a) reflection guidance and collaboration scripts are effective, with a particular focus on reflection guidance and (b) process data can inform the adaptive simulation-based learning of collaborative diagnostic reasoning using machine learning. To achieve these goals, three studies were conducted using different methodologies. Whereas Studies 1 and 2 focused on adaptive reflection guidance at the macro level while considering prior knowledge, Study 3 focused on the micro level by exploring learner behavior as indicated by process data. This chapter first summarizes and interprets the findings of the three studies. Then, the resulting new theoretical implications are discussed in relation to the current state of research, the limitations of the studies are addressed, and future research directions are suggested. Finally, practical implications and a final conclusion are drawn for this dissertation.

## 5.1 Summary and Interpretation of Central Results

*Study 1* (Richters et al., submitted) investigated how guidance for reflection on individual activities and collaboration scripts separately and synergistically affected collaborative diagnostic reasoning as a function of learners' prior content and collaboration knowledge in an agent-based simulation. Furthermore, the study explored how engagement in individual reflection and collaboration contributed to the diagnostic process. A sample consisting of 151 advanced medical students was used for this study. Students were randomly assigned to receive either structured reflection questions, external collaboration scripts, both, or no scaffolding while working on patient cases in the agent-based simulation. Students first worked on a pretest case without scaffolding, then on three learning cases with scaffolding in accordance with their assigned experimental condition, and finally on one posttest case without scaffolding.

The results revealed that reflection guidance, which provides less guidance, is effective for learners with high levels of prior content knowledge, whereas collaboration scripts, which provide more guidance, are effective for learners with low levels of prior content knowledge. These findings are in line with previous research both within and outside of simulations (Chernikova, Heitzmann, Fink, et al., 2020; Chernikova, Heitzmann, Stadler, et al., 2020; Jiang et al., 2018; Kalyuga, 2007; Simonsmeier et al., 2021). Effects of the scaffolding were found on collaborative diagnostic activities but not on the diagnostic outcomes (diagnostic accuracy and diagnostic justification). Reflection guidance positively

affected the hypothesis-sharing performance of learners with high levels of prior content knowledge, whereas it negatively affected the performance of learners with low levels of prior content knowledge. Furthermore, collaboration scripts positively affected the evidence-sharing performance of learners with low levels of prior content knowledge, whereas such scripts negatively affected the performance of learners with high levels of prior content knowledge. These findings suggest that learners with high levels of prior content knowledge were able to effectively activate their content knowledge by reflecting in writing on individual activities before collaborating, whereas learners with low levels of prior content knowledge were unable to do so (see similar effects for note-taking; Wetzels et al., 2011). For learners with low levels of prior content knowledge, collaboration scripts provided an optimal level of guidance, at least leading to the sharing of more relevant evidence, whereas for learners with high prior content knowledge, scripts appeared to provide unnecessary guidance (expertise reversal effect; Kalyuga et al., 2003), possibly limiting their autonomy in the collaborative diagnostic reasoning process (Wise & Schwarz, 2017).

Moreover, there was no evidence of a synergistic effect. Because not all learners benefited from reflection and collaboration scripts, it was not surprising that the combination of the two did not generally have a positive effect on learning. It seems likely that a prerequisite for a synergistic effect of the two forms of scaffolding (Tabak, 2004) would be that learners also benefited from the individual forms of scaffolding, or at least that they did not perform worse than without scaffolding, which was not the case for all learners. Precollaboration reflection prevented a positive effect of the script on the evidence-sharing performance of learners with low levels of prior content knowledge and compensated for the negative effect of the script for learners with high levels of prior content knowledge. This finding suggests that this form of scaffolding, combined with the additional script, likely cognitively overwhelmed learners with low levels of prior content knowledge (Eckhardt et al., 2013), and they could not benefit from reflection because of their insufficient prior knowledge base. For the learners for whom the script was conducive to learning, to benefit from additional reflection on the content, as suggested by Vogel et al. (2017), the reflection may need to be more structured or may need to provide additional content to counteract the cognitive overload. Furthermore, for learners with high levels of prior knowledge, reflection seemed to provide them with the opportunity to critically evaluate their individual activities, activate their existing knowledge, and foster cognitive flexibility, which helps these learners adaptively adjust their use of collaboration scripts, thereby mitigating the negative effects of collaboration scripts.



Regarding the contributions of reflection and collaboration to the diagnostic process, the study revealed that reflection did not change learners' early suspected diagnosis (indicator of cognitive case representation). By contrast, collaborative engagement improved the final diagnosis (diagnostic accuracy) regardless of learners' prior content knowledge or whether learners received external collaboration scripts. Thus, collaborating with the computer agent appeared to be generally helpful for the overall diagnostic outcome because learners could choose how often to consult their partner and thus gain access to external knowledge sooner or later. In addition, collaboration may have inherently stimulated reflection, as learners may have realized that they needed to think about the case before they started collaborating.

Overall, Study 1 indicated that guidance that helps learners reflect on individual activities and collaboration scripts are beneficial for fostering collaborative diagnostic reasoning with agent-based simulations, as long as the guidance and scripts are aligned with learners' prior content knowledge. Guidance that helps learners reflect on individual activities has the potential to activate prior content knowledge and thereby enhance collaboration, provided that learners have a high level of prior content knowledge. The overall diagnostic outcome is fostered by the collaboration itself, regardless of learners' prior content knowledge or additional collaboration support.

*Study 2* (Richters, Stadler, Brandl, et al., 2023) followed up on the conditions under which reflection guidance is beneficial for learning collaborative diagnostic reasoning by examining the effects of different types of reflection guidance (low- and high-structured) that directly addressed collaborative activities as a function of learners' prior collaboration knowledge. A sample consisting of 195 mid-level medical students was used for this study. Students were randomly assigned to receive either low-structured reflection guidance (scriptlet level), high-structured reflection guidance (scene level), or no reflection guidance while working on patient cases in the agent-based simulation. Again, students first worked on a pretest case without scaffolding, then on three learning cases with scaffolding in accordance with their assigned experimental condition, and finally on a posttest case without scaffolding.

The results indicated that reflection guidance was exclusively effective for learners with low prior collaboration knowledge. Effects on both collaborative diagnostic activities and diagnostic outcomes were found. Low-structured reflection guidance improved learners' performance in evidence sharing, diagnostic accuracy, and diagnostic justification, whereas the high-structured reflection improved the quality of diagnostic justification only for learners with low prior knowledge (see Appendix F). The only subskill that did not improve with reflection for learners with low levels of prior collaboration knowledge was performance in

hypothesis sharing. As there was also little variance in hypothesis sharing across learners with different levels of prior knowledge, this lack of effect for hypothesis sharing suggests that hypothesis sharing likely depends more on collaboration knowledge than on content knowledge (Study 1). For learners with high prior knowledge, low-structured reflection had a negative effect on evidence-sharing performance and no effect on diagnostic accuracy, and both low- and high-structured reflection had negative effects on the quality of diagnostic justification. Structured reflection at the scene level (low-structured reflection) seems to be the optimal structure for learners with low levels of prior collaboration knowledge, whereas it was either unnecessary or even detrimental for learners with well-developed collaboration scripts (Kalyuga et al., 2003). In contrast to structured reflection at the scene level, structured reflection at the scriptlet level (high-structured) was too detailed—sometimes even harmful—for all learners. These findings suggest on the one hand that low-structured reflection offered the optimal instructional support without overloading the working memory of learners with low levels of prior knowledge (Sweller, 2005) while fostering learner autonomy to independently explore and critically evaluate the diagnostic process (Nguyen et al., 2014; R. M. Ryan & Deci, 2000; Strauß et al., 2023). On the other hand, however, high-structured reflection may limit learner autonomy, which may result in learners with low levels of prior collaboration knowledge still relying heavily on the external reflection guidance rather than developing and using their own reflection strategies, thus hindering their learning (Wise & Schwarz, 2017). Furthermore, the performance of learners with high levels of prior collaboration knowledge may even suffer from this high degree of guidance, as it induces working memory overload that is detrimental to learning (Kalyuga et al., 2003). Interestingly, however, the scriptlet level did foster learning, at least for learners with low prior collaboration knowledge, by helping them justify their diagnoses. It seems that the high level of structure in reflection helped learners explain their diagnostic process as coherently as the low level of structure did. Thus, in order to externalize the diagnostic process or diagnostic decisions, a higher level of structure does not seem unnecessary, but may potentially be helpful.

Overall, Study 2 indicated that guidance for reflection on collaborative activities is beneficial for fostering collaborative diagnostic reasoning with agent-based simulations, provided it is aligned with learners' prior collaboration knowledge. Low-structured reflection on collaborative activities is beneficial for learners with low levels of prior collaboration knowledge but detrimental for learners with high levels of prior collaboration knowledge. High-structured reflection on collaborative activities is on average less helpful across all

subskills for learners with low levels of prior collaboration knowledge and even detrimental for learners with high levels of prior collaboration knowledge.

*Study 3* (Richters, Stadler, Radkowsch, et al., 2023) investigated whether and how quickly diagnostic accuracy (correct and incorrect diagnoses indicating diagnostic success or failure) could be predicted from collaborative diagnostic activities in an agent-based simulation using machine learning. To do so, a diverse sample consisting of 73 medical students and 25 physicians working on five consecutive patient cases was used. Log files were automatically coded for collaborative diagnostic activities, including evidence generation, evidence elicitation, evidence sharing, hypothesis sharing, and drawing conclusions. For each participant working on a case, a behavior string was created from the log files, resulting in a total of  $N = 476$  behavior strings after missing values were excluded. From these strings, bigrams containing information about the time spent on and transitions between collaborative diagnostic activities were created and used to train three different algorithms. Support vector machines, random forests, and gradient boosting machines classified the diagnosticians' final diagnoses as either correct or incorrect on the basis of the collaborative diagnostic activities. Furthermore, a Chi-Square test for each bigram was performed to determine which bigrams were more typical of diagnostic success and which were more typical of diagnostic failure.

Results indicated that all algorithms performed well in predicting diagnostic accuracy, but the random forest model was selected for the final interpretation because it performed slightly better in the testing phase ( $\kappa = .40$ ). The results indicated a more reliable prediction of diagnostic success (sensitivity = .90) than diagnostic failure (specificity = .46). Diagnostic success could be predicted before the case was completed. This result suggests that successful diagnosticians in this sample may have exhibited less behavioral variation than unsuccessful diagnosticians, who may differ greatly in their cognitive misbehavior as manifested by diverse behavior at the simulation level. Moreover, dedicating more time to individual activities, such as evidence generation and drawing conclusions, was indicative of diagnostic success. By contrast, dedicating more time to collaborative activities, such as evidence elicitation, setbacks in collaborative activities (e.g., returning from hypothesis sharing to evidence elicitation), and transitions between individual and collaborative activities (progressing from evidence sharing to drawing conclusions) were indicative of diagnostic failure. These findings highlight the importance of an appropriate initial cognitive case representation (Charlin et al., 2007) as a prerequisite for successful collaboration and the diagnostic outcome. Successful diagnosticians are able to generate a clear suspected diagnosis

and make targeted radiological requests. They appear to spend more time cognitively processing information from the health record and radiologic test results before arriving at a final diagnosis. By contrast, unsuccessful diagnosticians struggle with early cue identification and thus lack an adequate initial cognitive representation of the case. These diagnosticians urgently seek more radiologic information but struggle with collaborating (requesting numerous tests) and processing the information effectively (using evidence to validate or exclude hypotheses). These struggles manifest in long collaboration times, frequent transitions, and setbacks.

Overall, Study 3 clearly indicated that the time spent on collaborative diagnostic activities during the collaborative diagnostic reasoning process can be effectively used as a source of data to predict the diagnostic outcome, particularly diagnostic success. The prediction of diagnostic success is possible before task completion.

## **5.2 Theoretical Implications for Fostering Collaborative Diagnostic Reasoning Through Reflection in Agent-Based Simulations**

The first subgoal of this dissertation was to identify conditions under which scaffolding, especially reflection guidance, is effective for learning collaborative diagnostic reasoning. The positive effects of guidance for reflection on individual activities for learners with high levels of prior content knowledge and the negative effects for learners with low levels of prior content knowledge found in Study 1 are consistent with previous research (Chernikova, Heitzmann, Fink, et al., 2020; Chernikova, Heitzmann, Stadler, et al., 2020). The positive effects for learners with high levels of prior content knowledge appear to be due to knowledge activation, which is partly consistent with recent theoretical discussions of the empirical findings on reflection in the context of diagnostic reasoning (Mamede & Schmidt, 2022). However, reflection on individual activities only improved hypothesis sharing but did not change learners' initial case representations (Charlin et al., 2007), as reflected by the lack of effect on the overall diagnostic outcome. These findings contrast with previous studies that have demonstrated positive effects of reflection on the outcomes of individual diagnostic reasoning outside of simulations (cf. Ibiapina et al., 2014; cf. Mamede et al., 2014), effects that have been attributed primarily to knowledge reorganization (Mamede et al., 2014). These contrasting results suggest that the effects of reflection guidance on individual problem solving outside of simulation-based learning might not transfer readily to collaborative problem solving within or outside of simulation-based learning.



When practicing tasks that must be performed as part of specific professions, simulations allow learners to engage in behaviors that are not possible or would have serious consequences in other learning environments or in real life (e.g., trial and error or the repeated performance of certain activities), making simulations inherently effective for learning and possibly even more effective than scaffolding (Chernikova, Heitzmann, Stadler, et al., 2020). This notion was further supported by the Study 1 finding that collaborative engagement aids the diagnostic outcome by reducing diagnostic uncertainty, whereas reflection does not. In collaborative problem solving—or speaking more generally, in collaboration—collaborators can provide additional sources of knowledge and perspectives (Clark & Sampson, 2007; OECD, 2017; Radkowsch et al., 2022). Furthermore, collaboration partners provide mutual scaffolding (De Wever et al., 2010). Thus, collaboration has the potential to offer inherent learning potential by helping learners develop knowledge and skills (Vogel et al., 2017). The agent-based simulation provided learners with the opportunity to interact with the collaborator (agent) multiple times and, sooner or later, to access the knowledge that resulted from these interactions and ultimately benefit from the learning potential that collaboration offers. Because collaborative engagement generates new knowledge, it appears to have been more beneficial than individually reflecting on content. In a broader sense, these results can be linked to Vogel et al.'s (2017) findings that collaboration scripts are particularly beneficial for domain-specific learning when combined with additional content-specific support. Study 1 did not find evidence that reflection was a useful additional support, as indicated by the lack of synergistic effects between reflection guidance and collaboration scripts. However, unlike collaboration, reflection did not provide additional content. Instead of reflection as content support, collaboration itself provided additional content, rendering it generally beneficial for learning collaborative diagnostic reasoning.

Notably, this finding held true for all learners, not just those with low levels of prior content knowledge or those who were additionally supported by collaboration scripts. This effect can be explained by the following: Despite different levels of prior knowledge, all learners (medical students) in the role of internists were still in an intermediate stage of skill development, in contrast to the collaboration partner (agent-based radiologist), which was programmed as an expert colleague. Therefore, this main effect suggests that collaboration with an expert colleague, at least in an agent-based simulation, has inherent learning potential for intermediate learners who are learning complex problem solving or diagnostic reasoning. Similarly, Zambrano et al. (2019) found different benefits from collaboration in CL, depending on the composition of the team in terms of prior knowledge. Specifically, the

authors found that learners with low levels of prior knowledge benefited more from collaboration compared with individual learning, whereas learners with high levels of prior knowledge did not necessarily benefit from collaboration. The variance in prior content knowledge measured in Study 1 does not appear to be sufficient to determine differential benefits of collaboration among learners, as all participants appeared to benefit in similar ways. Examining broader skill scales (e.g., novice to advanced) in future studies could potentially reveal the differential benefits of collaboration and provide a more nuanced understanding of its effects. Thus, future studies could examine the extent to which learners at more advanced skill levels benefit from an expert collaboration partner in diagnostic reasoning.

Another explanation could be that collaboration may have left more room for inherent reflection, similar to what Fink et al. (2021) found with serial cue cases. The particular importance of collaboration for diagnostic outcomes was also indirectly demonstrated in Study 2, where learners with low levels of prior collaboration knowledge improved their diagnostic outcomes by reflecting on collaborative activities.

Study 2 identified other conditions under which guidance for reflection on collaborative activities is effective for learning collaborative diagnostic reasoning. The positive effects for learners with low levels of prior collaboration knowledge and the negative effects for learners with high levels of prior collaboration knowledge contrast with Study 1 and previous meta-analytic findings (cf. Chernikova, Heitzmann, Fink, et al., 2020) that have suggested that reflection guidance is particularly beneficial for learners with high levels of prior knowledge. The differences between the effect of the interaction between reflection and content knowledge in Study 1 and the effect of the interaction between reflection and collaboration knowledge in Study 2 suggest that content and collaboration knowledge are structured and organized differently. Taken together, the results of Studies 1 and 2 show that reflection is not generally appropriate for learners with high levels of prior knowledge (cf. Chernikova, Heitzmann, Stadler, et al., 2020) but that its effectiveness depends on the content that the learner is reflecting on and the fit between the level of structure in the reflection and the learner's prior knowledge level.

Furthermore, the findings from Study 2 contradict the previously stated hypotheses that high-structured reflection would be beneficial for learners with low levels of prior collaboration knowledge and low-structured reflection for learners with high levels of prior collaboration knowledge. High-structured reflection was not effective for learners with low levels of prior collaboration knowledge and was to some extent even detrimental for learners

with high levels of prior collaboration knowledge. Low-structured reflection was not effective or was even detrimental for learners with high levels of prior collaboration knowledge, but it was effective for learners with low levels of prior collaboration knowledge. One possible explanation for these patterns of findings could be that the hierarchical relationship between the scene level and the scriptlet level, as postulated in F. Fischer et al.'s (2013) script theory of guidance, is not necessarily valid. More specifically, although the script theory of guidance makes the assumption that internal collaboration scripts are highly flexible configurations of knowledge components, it assumes a hierarchical relationship between the components, such as the scene and scriptlet levels (F. Fischer et al., 2013). The findings of this dissertation may indicate that the relationship between the components is also highly flexible and possibly nonlinear. Future research could therefore benefit from exploring and empirically testing this hierarchical relationship between script levels. Moreover, a possible explanation for the lack of effect of high-structured reflection may be based on the meta-analysis by Vogel et al. (2017). Vogel et al. suggested that the detailed scriptlet level (high-structured reflection) is particularly appropriate for fostering general collaboration skills, such as argumentation skills (e.g., Noroozi et al., 2012), whereas the scene level (low-structured reflection) may be more appropriate for fostering domain-specific knowledge. Because collaborative diagnostic reasoning involves both collaborative and domain-specific aspects, the scriptlet level might not have been an appropriate choice for fostering collaborative diagnostic reasoning.

An exception to this finding, however, was diagnostic justification, on which a positive effect of the scriptlet level was found for learners with low levels of prior knowledge. There were no effects on any of the other subskills, but the lack of effects on diagnostic accuracy compared with diagnostic justification is particularly interesting in this context. The disparity in the effects on the two facets of diagnostic outcomes is consistent with Bauer et al. (2022), who found that preservice teachers differed significantly in the ability to make accurate diagnoses and adequately justify them. Such differences were attributed to the different knowledge bases underlying the two subskills, emphasizing that collaborative diagnostic reasoning involves several complex subskills that are more or less interrelated (Bauer et al., 2022). In continuing this line of research, Study 2 suggested that learners benefit from different levels of support for different subskills of (collaborative) diagnostic reasoning because of the different types of knowledge that are involved and the different levels of competence that learners have in different subskills. For example, learners with low levels of prior knowledge scored lower on diagnostic justification than on diagnostic accuracy (see Appendix F, Figures F2 and F3), which may explain why the scriptlet level helped them

better justify their diagnoses. However, the fact that the scriptlet level did not help these learners more than the scene level and even burdened the learners who had a high level of prior knowledge may indicate that the potential stand-alone benefits of increased structure in reflection only become apparent when the learner has substantially limited prior knowledge and also scores substantially low, and the potential harm becomes relevant only when the learner has exceeded a certain level of competence in a subskill (Kalyuga, 2007; Kirschner et al., 2006). Future studies could investigate thresholds at which a certain level of structure in reflection guidance becomes detrimental to the learning of certain subskills.

In sum, the amount of structure that learners need to guide their reflection seems to depend on which collaborative diagnostic reasoning subskill is being promoted, which particular type of knowledge the subskill involves, and how much competence learners already have in it.

Jointly, the findings from Studies 1 and 2 support the notion that the design of effective scaffolding is more about variation in cognitive and self-regulatory demands, namely, the fit between the structure of the instruction and learners' prior knowledge, than about the choice of the scaffold itself (Chernikova, Heitzmann, Fink, et al., 2020; Simonsmeier et al., 2021). Furthermore, the findings imply foundations for macro-adaptive reflection guidance: Guidance for reflection on individual activities primarily activates existing content knowledge, thus helping learners improve their collaboration (i.e., hypothesis sharing). To benefit from this guidance, learners require a high level of prior content knowledge (see Mamede & Schmidt, 2022). For learners with low levels of prior content knowledge, it may be necessary to provide initial support to help them build a cognitive case representation or to provide knowledge prompts along with reflection questions before pure reflection guidance becomes beneficial. By contrast, helping learners reflect on collaborative activities seems promising for helping them internalize collaboration scripts (F. Fischer et al., 2013) and restructure their content knowledge (Boshuizen & Schmidt, 1992). This process leads to improved collaboration (i.e., evidence sharing) and diagnostic outcomes (i.e., diagnostic accuracy and diagnostic justification). To support learners with insufficient prior collaboration knowledge, the use of guidance with less detailed questions (e.g., scene-level questions) effectively encourages thoughtful reflection on their collaborative performance. For learners with sufficient collaboration knowledge, an even less detailed prompt for reflection (e.g., at the play level) seems promising. Future research could investigate conditions under which reflection on collaborative activities is effective for learners with high levels of prior collaboration knowledge.



Overall, the findings suggest that guiding individual reflection in collaborative diagnostic reasoning offers a promising instructional approach for supporting the learning of collaborative diagnostic reasoning in agent-based simulations. Reflection guidance seems to be a flexible and autonomy-enhancing instructional approach (Nguyen et al., 2014; Strauß et al., 2023) that can be focused on different content areas and structured to a greater or lesser extent to meet the diverse needs of learners with different levels of prior knowledge or current skills. The effects of reflection support on collaborative diagnostic reasoning in simulation-based learning may differ from the effects on individual diagnostic reasoning found outside of simulation-based learning. The reasons for these differences include the learning opportunities that are inherently created by collaboration (Chi & Wylie, 2014; Clark & Sampson, 2007; De Wever et al., 2010; Kirschner et al., 2018; OECD, 2017; Radkowsch et al., 2022; Vogel et al., 2017) and simulation-based learning (Chernikova, Heitzmann, Stadler, et al., 2020), especially agent-based simulation (Graesser et al., 2018). Furthermore, when learning collaborative diagnostic reasoning with agent-based simulations, reflection on individual activities seems less helpful than reflection on collaborative activities for overall diagnostic outcomes. Whereas collaboration is generally helpful for improving diagnostic outcomes, collaboration and diagnostic outcomes can be improved a great deal by reflection on collaborative activities, at least for learners with low levels of prior knowledge. Thus, guidance for reflection on collaborative activities is particularly promising for fostering a wide range of collaborative diagnostic reasoning subskills. However, overly detailed guidance for reflection, such as the high-structured reflection in Study 2, might not be beneficial (cf. Renner et al., 2016), as it could compromise learner autonomy (R. M. Ryan & Deci, 2000), reminiscent of concerns associated with collaboration scripts (cf. Radkowsch et al., 2021; see Wise & Schwarz, 2017), and potentially overload working memory, especially for learners with high levels of prior knowledge (Kalyuga, 2007). Future research could examine the conditions under which reflection instruction may impede learning to provide valuable insights for refining its design.

### **5.3 Theoretical Implications for Adaptive Simulation-Based Learning of Collaborative Diagnostic Reasoning Using Process Data**

The second subgoal of this thesis was to go beyond macro-adaptivity and pure product data such as prior knowledge and to investigate conditions under which process analysis is suitable for informing adaptive simulation-based learning of collaborative diagnostic reasoning. To address this goal, Study 3 examined collaborative diagnostic reasoning

processes and the extent to which they could predict diagnostic accuracy. Whereas the first two studies provided evidence and suggestions about which and how scaffolding is or could be appropriate for learners with different levels of prior knowledge, Study 3 revealed differences in collaborative diagnostic reasoning processes between successful and unsuccessful diagnosticians. Notably, these differences were identified irrespective of the diagnosticians' prior knowledge and experience levels. Successful diagnosticians spend more time on individual activities, which prepares them for effective collaboration and leads to shorter collaboration times, whereas unsuccessful diagnosticians spend less time on individual activities, which leads to longer collaboration times and collaboration problems without meaningful processing of additional information gained through collaboration. More precisely, successful diagnosticians spend more time on existing case information in the beginning, rather than requesting additional information from the collaboration partner in an unfocused way. Stadler et al. (2019) found comparable results in the individual problem-solving context: Effective problem solvers prioritized thinking about the task, whereas their less successful counterparts spent more time on activities focused on gathering additional information, often without sufficient processing. Along with the findings from Studies 1 and 2, these findings highlight the critical role of an appropriate initial case representation (Charlin et al., 2007, 2012) for collaboration quality and overall diagnostic outcomes. The critical role of an appropriate initial case representation was also pointed out in previous analyses related to collaborative diagnostic reasoning in agent-based simulations (Vogel et al., 2023). Furthermore, in line with Studies 1 and 2, the findings emphasize the importance of collaboration for the overall diagnostic outcome (Radkowsch et al., 2022). Thus, all three studies somewhat emphasize the importance of an initial case representation and collaboration for overall diagnostic outcomes. However, because Studies 1 and 3 both used process analysis, it is particularly worthwhile to compare and integrate their findings. Study 1 suggested that one reason why reflection on individual activities prior to collaboration does not help learners achieve diagnostic success is that they struggle with restructuring their existing internal knowledge, as indicated by the unchanged suspected diagnoses. Instead, all learners achieved diagnostic success with the help of additional external knowledge gained through collaboration. Whereas Study 3 suggested that an appropriate case representation is a necessary condition for effective and efficient collaboration and subsequent diagnostic success, Study 1 suggested that collaborative engagement leads to diagnostic success regardless of whether or not learners begin with appropriate initial case representations. Because only time spent on the activities was considered as an indicator of diagnostic success

or failure in the process analyses in Study 3, Study 1 therefore provided additional insights into collaborative diagnostic reasoning processes. Jointly, these findings indicate that collaboration is generally helpful for improving diagnostic outcomes and fostering learning. The function of collaboration depends on the initial case representation: If the initial case representation is correct, collaboration tends to serve to confirm previous assumptions. In the best case, collaboration is efficient (i.e., fast and with few requests). If the initial case representation is incorrect, collaboration serves to introduce new knowledge into the diagnostic process and fundamentally change the process. In this case, however, collaboration runs the risk of being inefficient and unfocused (i.e., slow and with many requests).

Taken together, these findings can inform adaptive instructional support at the meso level, namely, in the upcoming case. Taking into account the experimental findings of Studies 1 and 2, a first implication of the process-analytical findings of Studies 1 and 3 concerns learners who struggled in the previous collaborative diagnostic reasoning process and therefore failed to correctly solve the case. Learners with low levels of prior content knowledge who struggle with the initial case representation could be given prompts to help them integrate information into hypotheses, or they could be given a list of relevant hypotheses to increase their likelihood of building a correct initial case representation that prestructures and prepares the individual diagnostic reasoning process for collaboration. Subsequently, they could receive collaboration support that encourages the concrete use of the collaboration partner as an external source of knowledge and guides the collaboration process. For example, an appropriate way to support learners with low levels of prior collaboration knowledge during collaboration is through externally guided reflection on collaborative activities with scene-level questions. Moreover, learners with a high level of prior content knowledge and an adequate initial case representation who still failed the case could benefit from reflection guidance to sharpen their existing representation. They could also benefit from collaboration support to help them make focused and efficient requests that are based on the correct case representation, thus helping them keep their collaboration effective and efficient. More precisely, learners with high levels of prior collaboration knowledge could benefit from a broad reflection prompt.

Another implication arises from the reliable prediction of diagnostic success before the case is completed, which allows for dynamic adaptivity at a micro level, namely, in the current case. For instance, learners could be given feedback that they are on the right track, or the difficulty of the task could be increased. However, because diagnostic success was predicted reliably only after about two thirds of the median time spent in the diagnostic

process, it did not qualify as early prediction compared with other studies (e.g., Ulitzsch et al., 2022). Thus, the prediction might not necessarily serve as a basis for removing scaffolding before it has a negative impact on learning (see Kalyuga, 2007). Furthermore, based on the findings from Study 3, such micro-level adaptivity would not be possible for learners who are on the wrong diagnostic track because diagnostic failure was not predicted nearly as reliably and quickly as diagnostic success, which may be due to greater behavioral variation among unsuccessful diagnosticians, at least in the sample that was used in the study.

The observed difference in predictive performance between diagnostic success and failure in Study 3 using the random forest model may be due to several factors. Features that are correlated with diagnostic success may inherently be of greater importance to the algorithm, resulting in improved predictive performance. In addition, the imbalance between cases of success and failure (162 failures and 314 successes) as well as the sensitivity of the algorithm to the class distribution may hinder the model's ability to effectively detect failure patterns. Furthermore, given the data structure of the log files, which are processed using *n-grams* (specifically, bigrams), there may be subtle variations in how these sequential patterns capture success- and failure-related information, potentially affecting the algorithm's generalization across outcomes. Predicting failure may be more challenging due to its varied and complex nature, resulting in different data patterns. Success patterns, on the other hand, may tend to be more consistent, perhaps making them easier for machine learning algorithms to identify and generalize. However, in a study by Brandl et al. (2021), both diagnostic success and failure were reliably predicted by using a random forest model on collaborative activities alone. These reliable results suggest that collaborative activities are better predictors of diagnostic failure than the combination of individual and collaborative activities, again emphasizing the central role of collaboration in the diagnostic outcome (Radkowsch et al., 2022). Because diagnostic failure was less reliably predicted than diagnostic success, additional research is needed to analyze the reasons for diagnostic failure at the behavioral level in order to identify and adapt to learners' needs.

Overall, the findings imply that the analysis of process data is a promising basis for meso- and micro-adaptivity when learning collaborative diagnostic reasoning with agent-based simulations. The analysis of learner behavior in collaborative diagnostic reasoning processes within a case provided reliable indications of where learners were struggling and where they were at risk of failing the case (e.g., insufficient focus on individual activities), indicating a lack of an appropriate initial case representation, subsequently leading to difficulties in collaboration. This information allows for meso-adaptive scaffolding that could

be offered in the next case when the learner failed the previous one. In addition, it is possible to determine whether learners are on the right diagnostic track even before the case is completed, and such information can be used to microadapt features of simulation-based learning, such as feedback or task difficulty.

#### **5.4 Limitations**

The various conceptual and methodological approaches used in this dissertation to provide foundations for adaptively fostering collaborative diagnostic reasoning in agent-based simulations are not without limitations. First, the use of agent-based collaboration may limit the applicability of our results to human collaboration. However, the simulation interface was carefully designed to closely resemble real-life collaboration between internists and radiologists, and previous studies have provided evidence of its validity (Radkowsch, F. Fischer, et al., 2020). Furthermore, no significant differences between agents and humans were found in a recent assessment of collaborative problem solving (Herborn et al., 2020). Future studies may wish to explore the transferability of the results to human-to-human collaboration.

Additional limitations concern the intervention studies (Studies 1 and 2). First, the test of collaboration knowledge focused exclusively on the exchange of information in medical collaboration and ignored other important aspects of collaboration, such as negotiation or regulation. However, this focus was justified given the importance of information sharing in medical practice (Tschan et al., 2009). The collaboration between radiology and internal medicine in the simulation also focused on information sharing, as a previous study showed that students and physicians have particular problems in sharing information (Tschan et al., 2009). In addition, radiologists often take on a service provider role, performing examinations on the basis of the internist's input, making collaborative aspects such as negotiation less central. However, it is still possible that the lack of interaction effects with the collaboration scripts in Study 1 is also related to the exclusive focus of the collaboration knowledge test on information sharing. In particular, the agent-based radiologist rejected learners when they made errors in evidence sharing; therefore, this rejection may mean that the prompts were adapted to evidence-sharing skills. However, the script was effective for learners with low levels of content knowledge. Thus, the script that was adapted to collaboration knowledge was particularly effective when learners had low levels of prior content knowledge. This finding further emphasizes the need to consider different learning characteristics at the same



time (Tetzlaff et al., 2023) and indirectly the assumption that content knowledge is a prerequisite for collaboration knowledge.

Second, the relatively brief learning phase in our experiments could also be taken into account. In such a short period of time, it is not unlikely that the learners' knowledge and skills were not yet strongly developed. For instance, the short reflection times in the learning phase may account for why reflection on individual activities did not change learners' early case representation (see Study 1). However, in previous studies with the same reflection questions, learners also did not reflect for longer periods of time (Ibiapina et al., 2014; Mamede et al., 2014). Therefore, future research could investigate longer reflection times. Furthermore, the significant results, especially those from Study 2, indicate that learners can improve their performance in collaborative diagnostic reasoning in a short period of time when supported by reflection guidance. Further research could investigate the long-term effects of reflection guidance on learning collaborative diagnostic reasoning with agent-based simulations.

Third, none of the studies in this dissertation examined the extent to which content knowledge and collaboration knowledge are related or how much of the other kinds of prior knowledge learners had. However, as the results of this dissertation suggest, learners with low levels of both content and collaboration knowledge may need help forming correct initial case representations and guidance in collaborating effectively before they can access external knowledge. Conversely, learners with high levels of content knowledge but low levels of collaboration knowledge may primarily need help refining existing representations and making targeted, efficient collaboration requests. Future research could therefore benefit from studies that can consider several learner characteristics at once in order to provide more valid results than regression, such as latent profile analyses (e.g., Tetzlaff et al., 2023).

Moreover, there is one more limitation concerning the lack of evidence for the restructuring of cognitive case representation through reflection in Study 1. The accuracy of a suspected diagnosis at a given point in time was used as a summative indicator of the case representation. However, the accuracy of suspected diagnoses does not fully capture the complexity and dynamics involved in a complete case representation, which additionally involves more details such as the inclusion and exclusion of relevant and irrelevant case information over time (see Braun et al., 2018). To learn more about the mechanisms of reflection effects, future studies could examine other outcomes and process-related indicators of cognitive case representations.

Finally, limitations concern the implications for adaptive simulation-based learning of collaborative diagnostic reasoning derived from Study 3. First, whereas the focus of Study 3 was on diagnostic accuracy, which was used as a measure of the task solution, it is questionable whether diagnostic accuracy is a sufficient valid and reliable measure of diagnostic competence (Klug et al., 2013), which is a much larger and more complex construct. In addition to indicators of diagnostic quality, such as diagnostic accuracy, diagnostic competence also includes professional knowledge and diagnostic activities (Heitzmann et al., 2019). However, diagnostic accuracy is the central goal of diagnostic reasoning (Chinn et al., 2011), and despite ongoing discussions about alternative measures of diagnostic competence (see Klug et al., 2013), accuracy remains the predominant metric for assessing diagnostic competence (Braun et al., 2019; Mamede et al., 2014; Pickal, Engelmann, Chinn, Neuhaus, et al., 2023), particularly in the medical field. The accuracy of diagnoses is of great importance due to the potentially serious consequences for patients when a diagnosis is not accurate (National Academies of Sciences, Engineering, and Medicine, 2015). Also, the diagnosis of a patient has a profound effect on subsequent procedures, including the formulation of treatment plans (Cook et al., 2019). For Study 3, diagnostic accuracy was deliberately chosen as an indicator of competence to predict diagnostic success and failure on the basis of collaborative diagnostic activities. For the prediction, several cases were used to ensure reliability at least to some extent. In addition, as Studies 2 and 3 examined instructional effects on different subskills of collaborative diagnostic reasoning, collaborative diagnostic competence was captured more comprehensively by considering both diagnostic quality and diagnostic activities.

Second, learning was not examined directly in Study 3. For example, someone who makes an incorrect diagnosis may still have learned something, or someone who makes a correct diagnosis might not have learned anything at all. Finally, whereas performance indicators were successfully derived from process data to inform future adaptive simulation-based learning, the proposed adaptive support approaches themselves were not implemented in Study 3. Similar studies in the context of simulation-based learning have been criticized for this issue and linked to the “from description to prescription” problem (Vermunt, 2023). For this reason, the results of Studies 1 and 2, which specifically examined scaffolding, were included in the suggestions for meso-adaptive scaffolding in order to make more valid statements.

## 5.5 Transferability to Other Fields and Contexts

An important question in the context of this dissertation is the transferability of the findings to other fields of higher education, such as engineering, psychology, or teacher education, where collaborative diagnostic reasoning plays an important role. Previous research on diagnostic reasoning has already looked at the comparison between medical and teacher education (e.g., Bauer et al., 2020; Chernikova, Heitzmann, Fink, et al., 2020). In principle, it is assumed that, at least to a certain extent, the results are transferable to other fields, as this dissertation was concerned with cross-field collaborative diagnostic activities (Radkowsch et al., 2022) and utilized broad concepts, such as the script theory of guidance (F. Fischer et al., 2013).

However, when considering transferability, it is important to recognize field-specific standards and practices in (collaborative) diagnostic reasoning as well. The nuances of collaborative diagnostic reasoning may differ depending on field-specific knowledge, problem-solving context, decision factors, stakes, collaboration dynamics, and time frames. For example, in medicine, collaborative diagnostic reasoning commonly involves high-stakes and rapid decision making as well as interdisciplinary collaboration (e.g., in emergency rooms), whereas in teacher education, collaborative diagnostic reasoning often unfolds over longer time frames with lower stakes. Instructional support for learning collaborative diagnostic reasoning that is tailored to the specific demands of each field therefore seems promising. As a method for providing instructional support that helps students learn collaborative diagnostic reasoning, reflection guidance holds promise across fields, as reflection is a flexible and autonomous process (Nguyen et al., 2014; Strauß et al., 2023) that can be adapted to various content and problem-solving processes. For example, the low-structured guidance for reflection on collaborative activities, which was shown to be particularly effective, leaves enough room for adaptation across fields. Such flexibility in instructional support is particularly important because diagnostic reasoning—whether individually or collaboratively applied—is less standardized in some fields, such as teacher education, than in medical education (Bauer et al., 2020). Transferring evidence from highly standardized fields to less standardized fields can be challenging. Whereas medical education benefits from well-defined procedures and sets of rules for solving specific problems or making a clinical decision, as well as associated sample solutions to specific problems, teacher education lacks such standardized resources. This difference can affect the accessibility and clarity of models for reflection in these fields. To address this issue, there is a need to develop standardized frameworks or collections of exemplars that are tailored to the

teacher education context. These resources can provide clear reference points for learners to effectively guide their reflective processes.

Moreover, when considering how collaborative diagnostic reasoning might be related to collaborative problem solving, the question that arises is whether the findings on the effects of reflection in this dissertation can be generalized not only across different fields but also across different collaborative problem-solving contexts. For example, learning processes that are associated with reflection processes, such as knowledge activation or restructuring, may be similar in other collaborative problem-solving contexts. Reflecting on individual activities may be promising not only for refining initial cognitive case representations in (collaborative) diagnostic reasoning but more generally for refining initial problem representations. Furthermore, the learning processes underlying reflection on collaborative activities in collaborative diagnostic reasoning may also be similar in other collaborative problem-solving contexts. Thus, reflecting on one's own collaborative contribution (self-reflection in collaboration) may be promising across contexts. However, even in medical education, empirical research on concrete reflection processes and the learning processes associated with them is still scarce.

Overall, the findings of this dissertation are promising for fostering collaborative diagnostic reasoning in different fields of higher education and across diverse collaborative problem-solving skills. However, further research is needed to test the generalizability across fields and collaborative problem-solving contexts.

## **5.6 Practical Implications**

Beyond its theoretical implications, this dissertation also offers valuable information for educational practice in how to adaptively foster collaborative diagnostic reasoning with agent-based simulations. First, this dissertation focused on reflection guidance as a scaffolding approach that can be applied to help medical students learn collaborative diagnostic reasoning in agent-based simulations. Reflection processes are fundamental to the development of professional competence, autonomy, and self-regulation (Nguyen et al., 2014; Strauß et al., 2023). The importance of reflective thinking for professional practice is also recognized in higher education, such as in teacher education (Beauchamp, 2015) and medical education (Sandars, 2009), where programs increasingly aim to develop students' skills by supporting reflection on practical experiences (Grossman & McDonald, 2008).

Reflection in collaborative contexts has mainly been conceptualized as collaborative reflection (e.g., Prilla et al., 2020; Zhang et al., 2023). In medical practice, collaborative

reflection is often referred to as team-based reflection (e.g., Schmutz et al., 2018, 2021). However, in collaboration in fields such as medicine, where a common problem is information sharing (e.g., Tschan et al., 2009), individually reflecting on one's own contribution to the collaboration seems to be an important step for improvement. The findings of this dissertation emphasize this perspective by demonstrating that individual reflection guidance can support the learning of different subskills involved in collaborative diagnostic reasoning when adapted to learners' prior knowledge. Therefore, medical education programs are likely to benefit from integrating reflection guidance to enhance collaborative diagnostic reasoning skills but also to develop reflection skills in a targeted manner. As opposed to the term *collaborative reflection*, which refers to joint reflection activities, these skills could be referred to as *self-reflection skills in collaboration*. In designing effective reflection guidance, the key challenge for medical educators is to ensure that while appropriate reflection guidance is provided, learners also do not become cognitively overwhelmed (Sweller, 2005). Such balance can be achieved by adapting the structure in the reflection phase to the learner's prior knowledge before simulation-based learning (macro-level adaptivity). Educators can determine the content and level of structure in reflection phases and thus design effective reflection support by considering which particular subskill medical students need to develop, what knowledge is associated with that subskill, and how much of that knowledge the student has. It is recommended that reflection phases do not include overly detailed questions to avoid cognitive overload (Sweller, 2005) and that the level of structure is reduced with increasing prior knowledge (Jiang et al., 2018; Kalyuga, 2007).

Furthermore, this dissertation showed that analyzing collaborative diagnostic reasoning processes by using data from interactions with agent-based simulations and machine learning can provide concrete insights into where diagnosticians face challenges in the process. For instance, diagnosticians tend to face challenges in building up an initial problem representation (see Charlin et al., 2007), or they intensively collect data without continuing to engage in inferential processes (see Stadler et al., 2019), thus leading to inaccurate diagnoses. Beyond product data (e.g., prior knowledge), analyzing collaborative diagnostic reasoning processes in real time seems promising as a basis for dynamically adapting scaffolding (e.g., reflection phases), to learners' current needs. Such adaptivity is realized during simulation-based learning between cases (meso-level adaptivity) or within a case (micro-level adaptivity). To develop effective agent-based simulations that meet the needs of medical students, it is therefore advisable to consider not only product data, such as prior knowledge, but also process data as a basis for adaptation. For process data, such as log



files, to provide valid insights into difficulties in the collaborative diagnostic reasoning process, it is advisable for medical educators to link the process data to theoretical models (Gašević et al., 2015), as was done in this dissertation. However, further research is warranted to directly explore the effectiveness of dynamically adapting scaffolding or feedback on the basis of collaborative diagnostic reasoning processes or performance in agent-based simulations.

Overall, integrating reflection support in higher education, particularly in simulation-based learning environments, can positively impact collaborative diagnostic reasoning and reflection skills. Medical educators are encouraged to consider the learning opportunities offered by simulations when integrating guidance for reflection to optimize the development of complex skills. Guidance for reflection on collaborative diagnostic reasoning processes seems particularly effective for helping medical students develop different subskills. In the reflection phases, it is advisable to pay attention to an appropriate structure that corresponds to the level of the student's prior knowledge. With increasing prior knowledge, the level of structure could be reduced (Jiang et al., 2018; Kalyuga, 2007). By implementing these practical recommendations, medical educators can enhance the effectiveness of instructional support, such as providing simulation-based learning and reflection guidance, fostering collaborative diagnostic reasoning skills, and facilitating adaptive learning experiences.

### **5.7 Directions for Future Research**

In addition to deriving theoretical and practical implications for adaptively fostering collaborative diagnostic reasoning with agent-based simulations, promising directions for future research can be derived from the findings of this dissertation. A first direction for future research concerns the conditions under which reflection is effective. For instance, Study 2's findings suggest that guidance for reflection on collaborative activities is particularly promising for fostering collaborative diagnostic reasoning, including collaboration and task outcomes. However, this approach was not effective for learners with high levels of collaboration knowledge. The dissertation provides a theoretical explanation for this lack of effect and suggests that a less detailed reflection prompt may be more effective for fostering collaborative diagnostic reasoning in learners with high levels of collaboration knowledge, a hypothesis that could be examined in future studies.

Moreover, Study 2 showed that learners with low levels of prior collaboration knowledge benefited from guidance for reflecting on collaborative activities with a relatively low level of structure, whereas Study 1 showed that learners with low levels of prior content

knowledge did not benefit from guidance for reflecting on individual activities at all. These findings suggest that the two different types of reflection guidance are not easy to compare in terms of the types of knowledge involved, their underlying mechanisms, and the levels of guidance. Previous research has emphasized that, by its very nature, reflective practice makes its quantification difficult but that systematic research with rigorous study designs, such as the designs used in this dissertation, is needed to evaluate different approaches to foster reflection (Mann et al., 2009). Future research could continue to strive to objectively scale different levels of structure in reflection support to allow reliable comparisons of different effects in the future. Such objective scaling may also increase the validity of potential meta-analyses that investigate the conditions under which reflection guidance is beneficial for learning complex skills in simulations, which has yet to be addressed in detail.

Furthermore, a promising direction for future research is to focus on investigating reflection processes. The analysis of reflection processes can provide information about the mechanisms behind reflection effects. The findings in this dissertation provide a theoretical starting point for understanding the conditions of reflection effects. However, there is still a need for further research on the mechanisms. In particular, even in individual diagnostic reasoning, there is a lack of empirical evidence of the extent to which reflection affects the initial cognitive case representation and restructures knowledge, such as through the use of illness scripts (Mamede & Schmidt, 2022). The findings from Study 1 suggest that knowledge is solely activated through reflection but not substantially reorganized or restructured, a finding that stands in some contrast to previous findings on individual diagnostic reasoning outside of simulation contexts (Mamede et al., 2014). Possible explanations for this discrepancy, such as the collaboration (Radkowsch et al., 2022) and the nature of simulations (Chernikova, Heitzmann, Stadler, et al., 2020), were already mentioned in this discussion. Follow-up studies could focus on exploring mechanisms by analyzing reflection processes. Possible methods could include coding and analyzing written reflection answers or think-aloud protocols. One approach to coding written reflection responses was suggested by Kember et al. (2008). The authors suggested categories for the level of reflection, namely, habitual action/nonreflection, understanding, reflection, and critical reflection (Kember et al., 2008). Such process analyses could also be used to compare the processes of learners with different levels of prior knowledge as they reflect on different activities in order to identify differences in reflection approaches and strategies and to gain more insight into the reasons for differences in reflection effects.

Considering other learner characteristics in reflection effects is another potential direction for research. The findings of the two intervention studies emphasize the complexity of reflection (e.g., Boud, 2001). Thus, the effects of reflection guidance are likely to depend on other factors beyond prior knowledge, such as motivation, interest, or self-regulation skills. Future research could examine the effectiveness of reflection guidance as a function of the interplay of different learner characteristics in order to derive more valid results (see Tetzlaff et al., 2023).

A direction for future research with respect to Study 3 concerns the effects of meso- or microadaptive instructional support, such as scaffolding or feedback based on collaborative diagnostic reasoning processes, on the learning of collaborative diagnostic reasoning. Study 3 identified differences in collaborative diagnostic reasoning processes between successful and unsuccessful diagnosticians. Furthermore, engagement in collaborative diagnostic activities reliably predicted diagnostic success even before the case was completed. Future studies could integrate dynamic assessments of learner engagement into these activities and predictions of diagnostic outcomes based on these activities into agent-based simulations to implement adaptive scaffolding.

Finally, as previously mentioned, future research could test both the generalizability of the findings of this dissertation across different fields other than medicine and the transferability to other types of collaboration in medicine and other collaborative problem-solving contexts. In this dissertation, collaborative diagnostic reasoning was conceptualized in accordance with the CDR model (Radkowsch et al., 2022). Due to the generic activities, the model promises transferability to collaborative diagnostic reasoning in other fields within and outside of higher education, such as in teacher education (e.g., Pickal et al., 2022) or automotive automechanics (e.g., Abele, 2018). In addition, the applicability to other interdisciplinary or even interprofessional collaborations in medicine (Hansen et al., 2023) could be of interest, such as the collaboration of different medical professionals in cardiac resuscitation. The findings on the effects of reflection could also be tested in other collaborative problem-solving contexts in different fields to advance the understanding of the benefits of reflection and its underlying mechanisms in a broader sense in the context of collaborative problem solving.

## **6 Conclusion**

*Constanze Catharina Richters*

Interdisciplinary collaboration skills, such as those involved in collaborative diagnostic reasoning, are central to professional practice in different fields. Collaborative reasoning is particularly crucial and important in high-stakes fields such as medicine, where diagnostic problems require careful consideration of different knowledge backgrounds in order to gain a comprehensive understanding of the underlying disease and take appropriate action (Shafran et al., 2017). Given the complexity of collaborative diagnostic reasoning and the challenges observed in practice in performing certain subskills (Tschan et al., 2009), learners are likely to benefit from support in learning how to diagnose collaboratively so that they can become proficient future diagnosticians (Radkowsch, F. Fischer, et al., 2020). Adapting simulation-based learning and scaffolding to individual learner's needs appears to offer a promising approach involving instructional support that helps students develop specific subskills of complex competencies, such as collaborative reasoning (F. Fischer et al., 2022; Plass & Pawar, 2020; Tetzlaff et al., 2021).

This dissertation aimed to provide various conceptual and methodological foundations for adaptively fostering collaborative diagnostic reasoning in agent-based simulations, with a particular focus on reflection guidance. To achieve this goal, three studies using different methodological approaches (conventional regression analysis and machine learning) were conducted. The findings of this dissertation provide robust foundations for macro-adaptive reflection support in simulation-based learning as well as starting points for instructional support at the meso and micro levels.

The findings of this dissertation make a theoretical contribution to research on individual reflection processes and how they can be facilitated in collaborative problem-solving contexts. More precisely, guidance for reflection on individual activities has the potential to activate prior content knowledge (Mamede & Schmidt, 2022) and thereby enhance collaboration, provided that learners have a high level of prior content knowledge. It is the task of future research to empirically clarify the extent to which this type of reflection can also restructure knowledge in (collaborative) diagnostic reasoning in medicine, in other domains, and more broadly in other (collaborative) problem-solving contexts. Guidance for reflection on collaborative activities has the potential to foster the learning of a wide range of collaborative diagnostic reasoning subskills by helping learners internalize collaboration scripts (F. Fischer et al., 2013) and restructure their knowledge (Boshuizen & Schmidt, 1992).

Overall, the effectiveness of guiding individual reflection in collaborative diagnostic reasoning in the agent-based simulation suggests that reflection guidance is a flexible and autonomy-enhancing instructional approach (Nguyen et al., 2014; Strauß et al., 2023) that can



be adapted to different levels of learners' prior knowledge. Reflecting on collaborative activities is particularly promising for improving different subskills of collaborative diagnostic reasoning. However, caution is advised against overly detailed reflection guidance (cf. Renner et al., 2016), which may compromise learner autonomy and lead to cognitive overload (Sweller, 2005). Furthermore, speaking more broadly, the effectiveness of reflection guidance on learning collaborative problem solving in simulations appears to result from a complex interplay of the degree of structure in reflection, the type and amount of learners' prior knowledge (Chernikova, Heitzmann, Fink, et al., 2020), the specific focus of the reflection (individual vs. collaborative activities), and the learning outcome or subskill that is targeted (e.g., problem solution such as diagnostic accuracy vs. externalization of problem-solving processes such as diagnostic justification). As collaborative diagnostic reasoning is a complex skill, and certain subskills (e.g., diagnostic accuracy and justification) are more or less interdependent (Bauer et al., 2022), it is likely that learners' skill levels in different subskills will vary.

Furthermore, the findings of this dissertation highlight the importance of theory-based process data (i.e., log files) beyond product data—such as prior knowledge—to identify subtle differences in collaborative diagnostic reasoning processes between successful and unsuccessful diagnosticians (Brandl et al., 2021; Goldhammer et al., 2017). Whereas more time spent on individual activities—such as evidence generation—predicted diagnostic success, more time spent on collaborative activities—such as evidence elicitation, jumping back and forth between collaborative activities, and jumping back and forth between collaborative and individual activities—predicted diagnostic failure. Combined with the reliable prediction of diagnostic success prior to task completion, these findings allow for more fine-grained and dynamic instructional support in the future, which is expected to improve the overall effectiveness of simulation-based learning.

Beyond the theoretical and practical implications for adaptively fostering collaborative diagnostic reasoning in agent-based simulations that this dissertation provides, it contributes significantly to the validation of the CDR model proposed by Radkowsch et al. (2022) by presenting diverse evidence supporting the relationships between collaborative diagnostic activities and diagnostic outcomes. Using multiple methodologies, the sources of evidence provided by the findings of this dissertation meet APA standards for validity, with a particular focus on evidence derived from relationships between test scores and other variables (American Educational Research Association et al., 2014). The findings demonstrate not only the reliability of predicting diagnostic outcomes from collaborative diagnostic activities using

machine learning techniques but also that external support for these activities (i.e., reflection guidance) improves diagnostic outcomes. These two sources of evidence increase the robustness of the validation of the relationships postulated in the CDR model.

Going beyond the CDR model, the findings underscore the critical role of a well-established initial cognitive case representation (Charlin et al., 2007) for successful collaboration and positive diagnostic outcomes (Vogel et al., 2023) as well as the importance of collaboration for positive diagnostic outcomes. Therefore, one might expect the effect of individual activities on the diagnostic outcome to be mediated by collaborative activities. However, the circumstances under which collaborative activities mediate the effects of individual activities on diagnostic outcomes, as postulated in the CDR model (Radkowsch et al., 2022), remain unclear. Initial studies that were designed to jointly validate the relationships postulated in the CDR model found relationships between individual characteristics, such as prior knowledge and collaborative activities, and between collaborative activities and the diagnostic outcome, but no mediation effect (Brandl et al., sub.). However, regardless of the potential mediating role of collaboration in the effects of individual characteristics and activities on diagnostic outcomes, collaborative engagement in agent-based simulations seems to offer inherent learning potential. Therefore, further research on the relationships in the CDR model and the learning potential of collaboration could aid the further development of adaptive instructional support for learning collaborative diagnostic reasoning.

In conclusion, building on the findings presented in this dissertation in future research and higher education practice has the potential to better prepare future diagnosticians—or more broadly, problem solvers—for interdisciplinary collaboration while fostering the autonomy that is critical for professional growth.

## 7 References

- Abele, S. (2018). Diagnostic problem-solving process in professional contexts: Theory and empirical investigation in the context of car mechatronics using computer-generated log-files. *Vocations and Learning, 11*(1), 133–159. <https://doi.org/10.1007/s12186-017-9183-x>
- Al-Kadi, A. S., & Donnon, T. (2013). Using simulation to improve the cognitive and psychomotor skills of novice students in advanced laparoscopic surgery: A meta-analysis. *Medical Teacher, 35*(Suppl.1), S47–S55. <https://doi.org/10.3109/0142159X.2013.765549>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition – Implications for the design of computer-based scaffolds. *Instructional Science, 33*(5–6), 367–379. <https://doi.org/10.1007/s11251-005-1272-9>
- Bauer, E., Fischer, F., Kiesewetter, J., Shaffer, D. W., Fischer, M. R., Zottmann, J. M., & Sailer, M. (2020). Diagnostic activities and diagnostic practices in medical education and teacher education: An interdisciplinary comparison. *Frontiers in Psychology, 11*, 562665. <https://doi.org/10.3389/fpsyg.2020.562665>
- Bauer, E., Sailer, M., Kiesewetter, J., Fischer, M. R., & Fischer, F. (2022). Diagnostic argumentation in teacher education: Making the case for justification, disconfirmation, and transparency. *Frontiers in Education, 7*, 977631. <https://doi.org/10.3389/educ.2022.977631>
- 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., 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>
- Bosch, B., & Mansell, H. (2015). Interprofessional collaboration in health care: Lessons to be learned from competitive sports. *Canadian Pharmacists Journal/Revue Des Pharmaciens Du Canada, 148*(4), 176–179. <https://doi.org/10.1177/1715163515588106>
- Boshuizen, H. P. A., Gruber, H., & Strasser, J. (2020). Knowledge restructuring through

- case processing: The key to generalise expertise development theory across domains? *Educational Research Review*, 29, 100310.  
<https://doi.org/10.1016/j.edurev.2020.100310>
- 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/10.1207/s15516709cog1602\\_1](https://doi.org/10.1207/s15516709cog1602_1)
- Boud, D. (Ed.). (1985). *Reflection: Turning experience into learning*. Routledge.
- Boud, D. (2001). Using journal writing to enhance reflective practice. *New Directions for Adult and Continuing Education*, 2001(90), 9–18. <https://doi.org/10.1002/ace.16>
- Boud, D., Cressy, P., & Docherty, P. (2006). *Productive reflection at work: Learning for changing organizations*. Routledge.
- Bowen, J. L. (2006). Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, 355(21), 2217–2225.  
<https://doi.org/10.1056/NEJMra054782>
- Brady, A., Laoide, R. Ó., McCarthy, P., & McDermott, R. (2012). Discrepancy and error in radiology: Concepts, causes and consequences. *The Ulster Medical Journal*, 81(1), 3–9.
- Brandl, L., Richters, C., Radkowsch, A., Obersteiner, A., Fischer, M. R., Schmidmaier, R., Fischer, F., & Stadler, M. (2021). Predicting performance with theoretically derived process features. *Psychological Test and Assessment Modeling*, 63(4), 542–560.
- Brandl, L., Stadler, M., Richters, C., Radkowsch, A., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2023). *Collaborative problem solving in knowledge rich domains: A multi study structural equation model* [Unpublished manuscript]. Department of Psychology, Ludwig-Maximilians-Universität München.
- Braun, L. T., Borrmann, K. F., Lottspeich, C., Heinrich, D. A., Kiesewetter, J., Fischer, M. R., & Schmidmaier, R. (2019). Scaffolding clinical reasoning of medical students with virtual patients: Effects on diagnostic accuracy, efficiency, and errors. *Diagnosis*, 6(2), 137–149. <https://doi.org/10.1515/dx-2018-0090>
- Braun, L. T., Lenzer, B., Kiesewetter, J., Fischer, M., & Schmidmaier, R. (2018). How case representations of medical students change during case processing – Results of a qualitative study. *GMS Journal for Medical Education*, 35(3).  
<https://doi.org/10.3205/zma001187>
- Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In N. J. Castellan, Jr. (Ed.), *Individual and group decision*



- making: Current issues.* (pp. 221–246). Lawrence Erlbaum Associates, Inc.
- Charlin, B., Boshuizen, H. P. A., Custers, E. J., & 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., Lubarsky, S., Millette, B., Crevier, F., Audétat, M.-C., Charbonneau, A., Caire Fon, N., Hoff, L., & Bourdy, C. (2012). Clinical reasoning processes: Unravelling complexity through graphical representation through graphical representation. *Medical Education, 46*(5), 454–463. <https://doi.org/10.1111/j.1365-2923.2012.04242.x>
- 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>
- Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105.  
<https://doi.org/10.1111/j.1756-8765.2008.01005.x>
- Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243.  
<https://doi.org/10.1080/00461520.2014.965823>
- Chinn, C. A., Buckland, L. A., & Samarapungavan, A. (2011). Expanding the dimensions of epistemic cognition: Arguments from philosophy and psychology. *Educational Psychologist, 46*(3), 141–167. <https://doi.org/10.1080/00461520.2011.587722>
- Cirigliano, M. M., Guthrie, C. D., & Pusic, M. V. (2020). Click-level learning analytics in an online medical education learning platform. *Teaching and Learning in Medicine, 32*(4), 410–421. <https://doi.org/10.1080/10401334.2020.1754216>
- Clark, D. B., & Sampson, V. D. (2007). Personally-seeded discussions to scaffold online argumentation. *International Journal of Science Education, 29*(3), 253–277.  
<https://doi.org/10.1080/09500690600560944>
- Cook, D. A., Durning, S. J., Sherbino, J., & Gruppen, L. D. (2019). Management reasoning: Implications for health professions educators and a research agenda. *Academic Medicine, 94*(9), 1310–1316. <https://doi.org/10.1097/ACM.0000000000002768>
- Cook, D. A., Hamstra, S. J., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., Erwin,

- P. J., & 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>
- Coulson, D., & Harvey, M. (2013). Scaffolding student reflection for experience-based learning: A framework. *Teaching in Higher Education*, 18(4), 401–413. <https://doi.org/10.1080/13562517.2012.752726>
- Cressey, P., & Boud, D. (2006). The emergence of productive reflection. In P. Boud, P. Cressey, & P. Docherty (Eds.), *Productive reflection at work: Learning for changing organisations* (pp. 11–26). Routledge.
- Csanadi, A., Kollar, I., & Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: Better together? *European Journal of Psychology of Education*, 36(1), 147–168. <https://doi.org/10.1007/s10212-020-00467-4>
- Custers, E. J. F. M. (2015). Thirty years of illness scripts: Theoretical origins and practical applications. *Medical Teacher*, 37(5), 457–462. <https://doi.org/10.3109/0142159X.2014.956052>
- 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(6), 902–912. <https://doi.org/10.1097/ACM.0000000000002618>
- Darmawansah, D., Lin, C.-J., & Hwang, G.-J. (2022). Empowering the collective reflection-based argumentation mapping strategy to enhance students' argumentative speaking. *Computers & Education*, 184, 104516. <https://doi.org/10.1016/j.compedu.2022.104516>
- Davies, S., George, A., Macallister, A., Barton, H., Youssef, A., Boyle, L., & Sequeiros, I. (2018). “It’s all in the history”: A service evaluation of the quality of radiological requests in acute imaging. *Radiography*, 24(3), 252–256. <https://doi.org/10.1016/j.radi.2018.03.005>
- Davis, E. A. (2000). Scaffolding students' knowledge integration: Prompts for reflection in KIE. *International Journal of Science Education*, 22(8), 819–837. <https://doi.org/10.1080/095006900412293>
- De Wever, B., Van Keer, H., Schellens, T., & Valcke, M. (2010). Structuring asynchronous discussion groups: Comparing scripting by assigning roles with regulation by cross-

- age peer tutors. *Learning and Instruction*, 20(5), 349–360.  
<https://doi.org/10.1016/j.learninstruc.2009.03.001>
- Dede, C. (2010). Comparing frameworks for 21st century skills. In J. Bellanca & R. Brandt (Eds.), *21st century skills: Rethinking how students learn* (pp. 51-76). Solution Tree Press.
- Desmarais, M. C., & Baker, R. S. J. D. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1–2), 9–38. <https://doi.org/10.1007/s11257-011-9106-8>
- Dewey, J. (1910). *How we think*. D.C. Heath & Co Publishers.
- Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada & P. Reiman (Eds.), *Learning in humans and machine: Towards an interdisciplinary learning science* (pp. 189-211). Elsevier.
- Dillenbourg, P., & Hong, F. (2008). The mechanics of CSCL macro scripts. *International Journal of Computer-Supported Collaborative Learning*, 3(1), 5–23.  
<https://doi.org/10.1007/s11412-007-9033-1>
- Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology*, 8, 1153. <https://doi.org/10.3389/fpsyg.2017.01153>
- Driscoll, M. R. (1987). Aptitude-treatment interaction research revisited. In M. R. Simonson & S. M. Zvacek (Eds.), *Proceedings of selected research paper presentations at the 1987 convention of the association for educational communications and technology and sponsored by the research and theory division* (pp. 172–182). Educational resources information center (ERIC).
- Eckhardt, M., Urhahne, D., Conrad, O., & Harms, U. (2013). How effective is instructional support for learning with computer simulations? *Instructional Science*, 41(1), 105–124. <https://doi.org/10.1007/s11251-012-9220-y>
- Emons, G., Steiner, E., Vordermark, D., Uleer, C., Bock, N., Paradies, K., Ortmann, O., Aretz, S., Mallmann, P., Kurzeder, C., Hagen, V., Van Oorschot, B., Höcht, S., Feyer, P., Egerer, G., Friedrich, M., Cremer, W., Prott, F.-J., Horn, L.-C., ... Erdogan, S. (2018). Interdisciplinary diagnosis, therapy and follow-up of patients with endometrial cancer. Guideline (S3-Level, AWMF Registry Nummer 032/034-OL, April 2018) – Part 1 with recommendations on the epidemiology, screening, diagnosis and hereditary factors of endometrial cancer. *Geburtshilfe und Frauenheilkunde*, 78(10), 949–971. <https://doi.org/10.1055/a-0713-1218>
- Engelmann, K., & Bannert, M. (2021). Analyzing temporal data for understanding the

- learning process induced by metacognitive prompts. *Learning and Instruction*, 72, 101205. <https://doi.org/10.1016/j.learninstruc.2019.05.002>
- Engelmann, T., & Hesse, F. W. (2010). How digital concept maps about the collaborators' knowledge and information influence computer-supported collaborative problem solving. *International Journal of Computer-Supported Collaborative Learning*, 5(3), 299–319. <https://doi.org/10.1007/s11412-010-9089-1>
- Engelmann, T., & Hesse, F. W. (2011). Fostering sharing of unshared knowledge by having access to the collaborators' meta-knowledge structures. *Computers in Human Behavior*, 27(6), 2078–2087. <https://doi.org/10.1016/j.chb.2011.06.002>
- Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *JAMA*, 287(2), 226. <https://doi.org/10.1001/jama.287.2.226>
- Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains: *Academic Medicine*, 79(Supplement), S70–S81. <https://doi.org/10.1097/00001888-200410001-00022>
- Eva, K. W. (2005). 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: Clinical expertise. *Medical Education*, 41(12), 1152–1158. <https://doi.org/10.1111/j.1365-2923.2007.02923.x>
- Feltovich, P. J., & Barrows, H. S. (1984). Issues of generality in medical problem solving. In H. G. Schmidt, & M.L. de Volder (Eds.), *Tutorials in problem-based learning: A new direction in teaching the health professions* (pp. 128–142). Van Gorcum.
- 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(1), 523. <https://doi.org/10.1186/s12909-021-02937-9>
- Fiore, S. M, Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., Massey, C., O'Neil, H., Pellegrino, J., Rothman, R., Soulé, H., & von Davier, A. (2017). *Collaborative problem solving: Considerations for the national assessment of educational progress*. National Center for Education Statistics.
- Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., Letsky, M., & Warner, N. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors: The Journal of the Human Factors*

- and Ergonomics Society*, 52(2), 203–224. <https://doi.org/10.1177/0018720810369807>
- Fischer, F., Bauer, E., Seidel, T., Schmidmaier, R., Radkowsch, A., Neuhaus, B. J., Hofer, S. I., Sommerhoff, D., Ufer, S., Kuhn, J., Küchemann, S., Sailer, M., Koenen, J., Gartmeier, M., Berberat, P., Frenzel, A., Heitzmann, N., Holzberger, D., Pfeffer, J., ... Fischer, M. R. (2022). Representational scaffolding in digital simulations – Learning professional practices in higher education. *Information and Learning Sciences*, 123(11/12), 645–665. <https://doi.org/10.1108/ILS-06-2022-0076>
- Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. *Educational Psychologist*, 48(1), 56–66. <https://doi.org/10.1080/00461520.2012.748005>
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M. R., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, 2(3), 28–45. <https://doi.org/10.14786/flr.v2i2.96>
- Fischer, G. (2001). User modeling in human–computer interaction. *User Modeling and User-Adapted Interaction*, 11(1/2), 65–86. <https://doi.org/10.1023/A:1011145532042>
- Fleck, R., & Fitzpatrick, G. (2010). Reflecting on reflection: Framing a design landscape. *Proceedings of the 22nd conference of the computer-human interaction special interest group of australia on computer-human interaction - OZCHI'10*, November 22-26, 2010 (pp. 216–223). <https://doi.org/10.1145/1952222.1952269>
- Foong, L. Y. Y., Nor, M. B. M., & Nolan, A. (2018). The influence of practicum supervisors' facilitation styles on student teachers' reflective thinking during collective reflection. *Reflective Practice*, 19(2), 225–242. <https://doi.org/10.1080/14623943.2018.1437406>
- 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. (2018). Systematizing professional knowledge of medical doctors and teachers: Development of an interdisciplinary framework in the context of diagnostic competences. *Education Sciences*, 8(4), 207. <https://doi.org/10.3390/educsci8040207>
- Gardner, A. K., & Ahmed, R. A. (2014). Transforming trauma teams through transactive memory: Can simulation enhance performance? *Simulation & Gaming*, 45(3), 356–370. <https://doi.org/10.1177/1046878114547836>
- Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not



- promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28, 68–84.  
<https://doi.org/10.1016/j.iheduc.2015.10.002>
- Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64–71. <https://doi.org/10.1007/s11528-014-0822-x>
- 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>
- Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education* (pp. 407–425). Springer International Publishing. [https://doi.org/10.1007/978-3-319-50030-0\\_24](https://doi.org/10.1007/978-3-319-50030-0_24)
- Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106(3), 608–626. <https://doi.org/10.1037/a0034716>
- Graber, M. L. (2009). Educational strategies to reduce diagnostic error: Can you teach this stuff? *Advances in Health Sciences Education*, 14(S1), 63–69.  
<https://doi.org/10.1007/s10459-009-9178-y>
- Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, 19(2), 59–92.  
<https://doi.org/10.1177/1529100618808244>
- Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61, 36–46. <https://doi.org/10.1016/j.chb.2016.02.095>
- Griffin, P., & Care, E. (Eds.). (2015). *Assessment and teaching of 21st century skills: Methods and approach*. Springer Netherlands. <https://doi.org/10.1007/978-94-017-9395-7>
- 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. <https://doi.org/10.1177/016146810911100905>

- Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal*, *45*(1), 184–205. <https://doi.org/10.3102/0002831207312906>
- Gustafsson, C., & Fagerberg, I. (2004). Reflection, the way to professional development? *Journal of Clinical Nursing*, *13*(3), 271–280. <https://doi.org/10.1046/j.1365-2702.2003.00880.x>
- Hall, D., & Buzwell, S. (2013). The problem of free-riding in group projects: Looking beyond social loafing as reason for non-contribution. *Active Learning in Higher Education*, *14*(1), 37–49. <https://doi.org/10.1177/1469787412467123>
- Hansen, N. L., Precht, H., Larsen, P., & Noehr-Jensen, L. (2023). Interprofessional diagnostic management teams: A scoping review protocol. *Systematic Reviews*, *12*(1), 223. <https://doi.org/10.1186/s13643-023-02391-2>
- Hautz, W. E., Kämmer, J. E., Schaubert, S. K., Spies, C. D., & Gaissmaier, W. (2015). Diagnostic performance by medical students working individually or in teams. *JAMA*, *313*(3), 303. <https://doi.org/10.1001/jama.2014.15770>
- Heitzmann, N., Seidel, T., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., Fischer, F., & Opitz, A. (2019). Facilitating diagnostic competences in simulations in higher education: A framework and a research agenda for medical and teacher education. *Frontline Learning Research*, 1–24. <https://doi.org/10.14786/flr.v7i4.384>
- Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, *104*, 105624. <https://doi.org/10.1016/j.chb.2018.07.035>
- Herrmann, W., Beckmann, J. F., & Kretzschmar, A. (2023). The role of learning in complex problem solving using MicroDYN. *Intelligence*, *100*, 101773. <https://doi.org/10.1016/j.intell.2023.101773>
- Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A Framework for Teachable Collaborative Problem Solving Skills. In P. Griffin & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills: Methods and Approach* (pp. 37–56). Springer Netherlands. [https://doi.org/10.1007/978-94-017-9395-7\\_2](https://doi.org/10.1007/978-94-017-9395-7_2)
- Hetmanek, A., Engelmann, K., Opitz, A., & Fischer, F. (2018). Beyond intelligence and domain knowledge. In F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation* (1st ed., pp. 203–226). Routledge.

- <https://doi.org/10.4324/9780203731826-12>
- Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). 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>
- Hommel, M., Fürstenau, B., & Mulder, R. H. (2023). Reflection at work – A conceptual model and the meaning of its components in the domain of VET teachers. *Frontiers in Psychology*, 13, 923888. <https://doi.org/10.3389/fpsyg.2022.923888>
- Hood, A. V. B., Whillock, S. R., Meade, M. L., & Hutchison, K. A. (2023). Does collaboration help or hurt recall? The answer depends on working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(3), 350–370. <https://doi.org/10.1037/xlm0001155>
- Houldin, A. D., Naylor, M. D., & Haller, D. G. (2004). Physician-nurse collaboration in research in the 21st century. *Journal of Clinical Oncology*, 22(5), 774–776.  
<https://doi.org/10.1200/JCO.2004.08.188>
- Ibiapina, C., Mamede, S., Moura, A., Elói-Santos, S., & van Gog, T. (2014). Effects of free, cued and modelled reflection on medical students' diagnostic competence. *Medical Education*, 48(8), 796–805. <https://doi.org/10.1111/medu.12435>
- Ifenthaler, D., Eseryel, D., & Ge, X. (Eds.). (2012). *Assessment in game-based learning: Foundations, innovations, and perspectives*. Springer. <https://doi.org/10.1007/978-1-4614-3546-4>
- Jiang, D., Kalyuga, S., & Sweller, J. (2018). The curious case of improving foreign language listening skills by reading rather than listening: An expertise reversal effect. *Educational Psychology Review*, 30(3), 1139–1165. <https://doi.org/10.1007/s10648-017-9427-1>
- Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development*, 48(4), 63–85.  
<https://doi.org/10.1007/BF02300500>
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9), 697–720. <https://doi.org/10.1037/0003-066X.58.9.697>
- Kahneman, D., & Frederick, S. (2001). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases* (1st ed., pp. 49–81). Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511808098.004>

- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases* (1st ed.). Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511809477>
- Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*(4), 509–539.  
<https://doi.org/10.1007/s10648-007-9054-3>
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23–31. [https://doi.org/10.1207/S15326985EP3801\\_4](https://doi.org/10.1207/S15326985EP3801_4)
- Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424.  
<https://doi.org/10.1080/07370000802212669>
- Kember, D., McKay, J., Sinclair, K., & Wong, F. K. Y. (2008). A four-category scheme for coding and assessing the level of reflection in written work. *Assessment & Evaluation in Higher Education, 33*(4), 369–379. <https://doi.org/10.1080/02602930701293355>
- Kim, P., Hong, J.-S., Bonk, C., & Lim, G. (2011). Effects of group reflection variations in project-based learning integrated in a Web 2.0 learning space. *Interactive Learning Environments, 19*(4), 333–349. <https://doi.org/10.1080/10494820903210782>
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. [https://doi.org/10.1207/s15326985ep4102\\_1](https://doi.org/10.1207/s15326985ep4102_1)
- Kirschner, P. A., Sweller, J., Kirschner, F., & Zambrano R., J. (2018). From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning, 13*(2), 213–233. <https://doi.org/10.1007/s11412-018-9277-y>
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–48. [https://doi.org/10.1207/s15516709cog1201\\_1](https://doi.org/10.1207/s15516709cog1201_1)
- Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*, 38–46.  
<https://doi.org/10.1016/j.tate.2012.10.004>
- Kmieciak, R. (2020). Critical reflection and innovative work behavior: The mediating role of individual unlearning. *Personnel Review, 50*(2), 439–459.  
<https://doi.org/10.1108/PR-10-2018-0406>
- Kolb, D. A. (1984). *Experimental learning: Experience as the source of learning and*

- development*. Prentice-Hall.
- Kollar, I., Wecker, C., & Fischer, F. (2018). Scaffolding and scripting (computer-supported) collaborative learning. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (1<sup>st</sup> ed., pp. 340–350). Routledge. <https://doi.org/10.4324/9781315617572-33>
- Kolodner, J. L. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review*, 6, 3–34. <https://doi.org/10.1007/BF00155578>
- Körkkö, M., Kyrö-Ämmälä, O., & Turunen, T. (2016). Professional development through reflection in teacher education. *Teaching and Teacher Education*, 55, 198–206. <https://doi.org/10.1016/j.tate.2016.01.014>
- Korthagen, F., & Vasalos, A. (2005). Levels in reflection: Core reflection as a means to enhance professional growth. *Teachers and Teaching*, 11(1), 47–71. <https://doi.org/10.1080/1354060042000337093>
- Krogstie, B. R., Prilla, M., & Pammer, V. (2013). Understanding and supporting reflective learning processes in the workplace: The CSRL model. In D. Hernández-Leo, T. Ley, R. Klamka, & A. Harrer (Eds.), *Scaling up learning for sustained impact* (Vol. 8095, pp. 151–164). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-642-40814-4\\_13](https://doi.org/10.1007/978-3-642-40814-4_13)
- Landriscina, F. (2012). Simulation and learning: The role of mental models. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 3072–3075). Springer US. [https://doi.org/10.1007/978-1-4419-1428-6\\_1874](https://doi.org/10.1007/978-1-4419-1428-6_1874)
- Leitner, P., Khalil, M., & Ebner, M. (2017). Learning analytics in higher education—A literature review. In A. Peña-Ayala (Ed.), *Learning analytics: Fundamentals, applications, and trends* (Vol. 94, pp. 1–23). Springer International Publishing. [https://doi.org/10.1007/978-3-319-52977-6\\_1](https://doi.org/10.1007/978-3-319-52977-6_1)
- Lin, X., Hmelo, C., Kinzer, C. K., & Secules, T. J. (1999). Designing technology to support reflection. *Educational Technology Research and Development*, 47(3), 43–62. <https://doi.org/10.1007/BF02299633>
- Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching*, 36(7), 837–858. [https://doi.org/10.1002/\(SICI\)1098-2736\(199909\)36:7<837::AID-TEA6>3.0.CO;2-U](https://doi.org/10.1002/(SICI)1098-2736(199909)36:7<837::AID-TEA6>3.0.CO;2-U)
- Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, D. (2016). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M.



- Mosharraf (Eds.), *Handbook of Research on Technology Tools for Real-World Skill Development* (pp. 344–359). IGI Global. <https://doi.org/10.4018/978-1-4666-9441-5.ch013>
- Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918. <https://doi.org/10.1037/a0037123>
- Mamede, S. (2020). What does research on clinical reasoning have to say to clinical teachers? *Scientia Medica, 30*(1), e37350. <https://doi.org/10.15448/1980-6108.2020.1.37350>
- 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., & Schmidt, H. G. (2022). Deliberate reflection and clinical reasoning: Founding ideas and empirical findings. *Medical Education, 57*(1), 76–85. <https://doi.org/10.1111/medu.14863>
- Mamede, S., van Gog, T., Sampaio, A. M., de Faria, R. M. D., Maria, J. P., & 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>
- Mann, K., Gordon, J., & MacLeod, A. (2009). Reflection and reflective practice in health professions education: A systematic review. *Advances in Health Sciences Education, 14*(4), 595–621. <https://doi.org/10.1007/s10459-007-9090-2>
- Mansilla, V., Gardner, H., & Miller, W. (2000). On disciplinary lenses and interdisciplinary work. In S. S. Wineburg & P. L. Grossman (Eds.), *Interdisciplinary curriculum: Challenges to implementation* (pp. 17–38). Teachers College Press.
- Mezirow, J. (1990). *Fostering critical reflection in adulthood: A guide to transformative and emancipatory learning* (1st ed). Jossey-Bass Publishers.
- Mezirow, J. (1998). On critical reflection. *Adult Education Quarterly, 48*(3), 185–198. <https://doi.org/10.1177/074171369804800305>
- Monteiro, S. D., Sherbino, J. D., Ilgen, J. S., Dore, K. L., Wood, T. J., Young, M. E., Bandiera, G., Blouin, D., Gaissmaier, W., Norman, G. R., & Howey, E. (2015). Disrupting diagnostic reasoning: Do interruptions, instructions, and experience affect the diagnostic accuracy and response time of residents and emergency physicians? *Academic Medicine, 90*(4), 511–517.

- <https://doi.org/10.1097/ACM.0000000000000614>
- Moon, J. A. (1999). *Reflection in learning and professional development* (1st ed.). Routledge. <https://doi.org/10.4324/9780203822296>
- Munshi, A., Biswas, G., Baker, R., Ocumpaugh, J., Hutt, S., & Paquette, L. (2023). Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. *Journal of Computer Assisted Learning*, 39(2), 351–368. <https://doi.org/10.1111/jcal.12761>
- National Academies of Sciences, Engineering, and Medicine. (2015). *Improving diagnosis in health care* (E. P. Balogh, B. T. Miller, & J. R. Ball, Eds.). National Academies Press. <https://doi.org/10.17226/21794>
- 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>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- Nokes-Malach, T. J., Richey, J. E., & Gadgil, S. (2015). When is it better to learn together? Insights from research on collaborative learning. *Educational Psychology Review*, 27(4), 645–656. <https://doi.org/10.1007/s10648-015-9312-8>
- Norman, G. R., Monteiro, S. D., Sherbino, J., Ilgen, J. S., Schmidt, H. G., & Mamede, S. (2017). The causes of errors in clinical reasoning: Cognitive biases, knowledge deficits, and dual process thinking. *Academic Medicine*, 92(1), 23–30. <https://doi.org/10.1097/ACM.0000000000001421>
- Norman, G. R., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education*, 41(12), 1140–1145. <https://doi.org/10.1111/j.1365-2923.2007.02914.x>
- Noroozi, O., Biemans, H. J. A., Weinberger, A., Mulder, M., & Chizari, M. (2013). Scripting for construction of a transactive memory system in multidisciplinary CSCL environments. *Learning and Instruction*, 25, 1–12. <https://doi.org/10.1016/j.learninstruc.2012.10.002>
- Noroozi, O., Weinberger, A., Biemans, H. J. A., Mulder, M., & Chizari, M. (2012). Argumentation-based computer supported collaborative learning (ABCSCCL): A synthesis of 15 years of research. *Educational Research Review*, 7(2), 79–106. <https://doi.org/10.1016/j.edurev.2011.11.006>

- Nückles, M., Hübner, S., & Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction, 19*(3), 259–271.  
<https://doi.org/10.1016/j.learninstruc.2008.05.002>
- OECD. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*. OECD.  
<https://doi.org/10.1787/9789264281820-en>
- O’Neill, T. A., Allen, N. J., & Hastings, S. E. (2013). Examining the “pros” and “cons” of team conflict: A team-level meta-analysis of task, relationship, and process conflict. *Human Performance, 26*(3), 236–260. <https://doi.org/10.1080/08959285.2013.795573>
- Paas, F., & Van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction, 16*(2), 87–91.  
<https://doi.org/10.1016/j.learninstruc.2006.02.004>
- Partnership for 21st Century Skills. (2009). *P21 Framework Definitions*. ERIC Clearinghouse.
- Pauli, R., Mohiyeddini, C., Bray, D., Michie, F., & Street, B. (2008). Individual differences in negative group work experiences in collaborative student learning. *Educational Psychology, 28*(1), 47–58. <https://doi.org/10.1080/01443410701413746>
- Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences, 13*(3), 423–451. [https://doi.org/10.1207/s15327809jls1303\\_6](https://doi.org/10.1207/s15327809jls1303_6)
- Pelaccia, T., Tardif, J., Tribby, E., & Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: The dual-process theory. *Medical Education Online, 16*(1), 5890. <https://doi.org/10.3402/meo.v16i0.5890>
- Peng, H., Ma, S., & Spector, J. M. (2019). Personalized adaptive learning: An emerging pedagogical approach enabled by a smart learning environment. *Smart Learning Environments, 6*(1), 9. <https://doi.org/10.1186/s40561-019-0089-y>
- Pickal, A. J., Engelmann, K., Chinn, C. A., Girwidz, R., Neuhaus, B. J., & Wecker, C. (2023). Fostering the collaborative diagnosis of cross-domain skills in video-based simulations. In C. Damsa, M. Borge, E. Koh, & M. Worsley (Eds.), *Proceedings of the 16th International Conference on Computer Supported Collaborative Learning -CSCL 2023* (pp. 139–146). International Society of the Learning Sciences.  
<https://doi.org/10.22318/csc12023.638463>
- Pickal, A. J., Engelmann, K., Chinn, C. A., Neuhaus, B. J., Girwidz, R., & Wecker, C. (2023). The diagnosis of scientific reasoning skills: How teachers’ professional

- knowledge predicts their diagnostic accuracy. *Frontiers in Education*, 8, 1139176.  
<https://doi.org/10.3389/educ.2023.1139176>
- Pickal, A. J., Engelmann, K., Girwidz, R., Neuhaus, B. J., & Wecker, C. (2022). Using simulations to foster pre-service teachers' diagnostic competences: What aspect of authenticity matters? In C. Chinn, E. Tan, C. Chan, & Y. Kali (Eds.), *Proceedings of the 16th International Conference on the Learning Sciences - ICLS 2022* (pp. 1353–162). International Society of the Learning Sciences.  
<https://doi.org/10.22318/icls2022.1353>
- Plass, J. L., & Pawar, S. (2020). 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>
- Prilla, M., Blunk, O., & Chounta, I.-A. (2020). How does collaborative reflection unfold in online communities? An analysis of two data sets. *Computer Supported Cooperative Work (CSCW)*, 29(6), 697–741. <https://doi.org/10.1007/s10606-020-09382-0>
- Prilla, M., Pammer, V., & Krogstie, B. (2013, September). Fostering collaborative redesign of work practice: Challenges for tools supporting reflection at work. In O. W. Bertelsen, L. Ciolfi, M. A. Grasso, & G. A. Papadopoulos (Eds.), *Proceedings of the 13th European Conference on Computer Supported Cooperative Work* (pp. 21–25) Springer.
- Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences*, 13(3), 337–386.  
[https://doi.org/10.1207/s15327809jls1303\\_4](https://doi.org/10.1207/s15327809jls1303_4)
- Quirk, M. E. (2006). *Intuition and metacognition in medical education: Keys to developing expertise*. Springer Pub. Co.
- Radkowsch, 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>
- Radkowsch, A., Sailer, M., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2022). Diagnosing collaboratively: A theoretical model and a simulation-based learning environment. In F. Fischer & A. Opitz (Eds.), *Learning to diagnose with simulations* (pp. 123–141). Springer International Publishing. [https://doi.org/10.1007/978-3-030-89147-3\\_10](https://doi.org/10.1007/978-3-030-89147-3_10)
- Radkowsch, A., Sailer, M., Schmidmaier, R., Fischer, M. R., & Fischer, F. (2021).

- Learning to diagnose collaboratively – Effects of adaptive collaboration scripts in agent-based medical simulations. *Learning and Instruction*, 75, 101487.  
<https://doi.org/10.1016/j.learninstruc.2021.101487>
- Radkowsch, A., Vogel, F., & Fischer, F. (2020). Good for learning, bad for motivation? A meta-analysis on the effects of computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*, 15(1), 5–47.  
<https://doi.org/10.1007/s11412-020-09316-4>
- Radović, S., Firssova, O., Hummel, H. G. K., & Vermeulen, M. (2021). Improving academic performance: Strengthening the relation between theory and practice through prompted reflection. *Active Learning in Higher Education*, 24(2), 139–154.  
<https://doi.org/10.1177/14697874211014411>
- Radović, S., Firssova, O., Hummel, H. G. K., & Vermeulen, M. (2023). The case of socially constructed knowledge through online collaborative reflection. *Studies in Continuing Education*, 45(2), 168–187. <https://doi.org/10.1080/0158037X.2022.2029389>
- Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, 38(1), 1–37. <https://doi.org/10.1111/cogs.12086>
- Renner, B., Prilla, M., Cress, U., & Kimmerle, J. (2016). Effects of prompting in reflective learning tools: Findings from experimental field, lab, and online studies. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00820>
- Richters, C., Stadler, M., Brandl, L., Schmidmaier, R., Fischer, M. R., & Fischer, F. (2023). Reflection on collaborative action: Fostering collaborative diagnostic reasoning in an agent-based medical simulation. In C. Damsa, M. Borge, E. Koh, & M. Worsley (Eds.), *Proceedings of the 16th International Conference on Computer Supported Collaborative Learning - CSCL 2023* (pp. 209–212). International Society of the Learning Sciences. <https://doi.org/10.22318/csl2023.596913>
- Richters, C., Stadler, M., Radkowsch, A., Behrmann, F., Weidenbusch, M., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2024). Fostering collaboration in simulations: How advanced learners benefit from collaboration scripts and reflection. *Learning and Instruction*, 93, 101912. <https://doi.org/10.1016/j.learninstruc.2024.101912>
- Richters, C., Stadler, M., Radkowsch, A., Behrmann, F., Weidenbusch, M., Fischer, M.R., Schmidmaier, R., & Fischer, F. (2022). Making the rich even richer? Interaction of structured reflection with prior knowledge in collaborative medical simulations. In A. Weinberger, W. Chen, D. Hernández-Leo, & B. Che (Eds.), *Proceedings of the 15<sup>th</sup> International Conference on Computer Supported Collaborative Learning – CSCL*

- 2022 (pp. 155-162). International Society of the Learning Sciences.  
<https://repository.isls.org/handle/1/8270>
- Richters, C., Stadler, M., Radkowsch, A., Schmidmaier, R., Fischer, M. R., & Fischer, F. (2023). Who is on the right track? Behavior-based prediction of diagnostic success in a collaborative diagnostic reasoning simulation. *Large-Scale Assessments in Education*, 11(1), 3. <https://doi.org/10.1186/s40536-023-00151-1>
- Rikers, R. M. J. P., Schmidt, H. G., & Boshuizen, H. P. A. (2000). Knowledge encapsulation and the intermediate effect. *Contemporary Educational Psychology*, 25(2), 150–166. <https://doi.org/10.1006/ceps.1998.1000>
- Roosevelt, F. D. (2008). Zone of proximal development. In *Encyclopedia of educational psychology* (pp. 1017–1022). SAGE Publications.
- Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer supported collaborative learning* (pp. 69–97). Springer Berlin Heidelberg.  
[https://doi.org/10.1007/978-3-642-85098-1\\_5](https://doi.org/10.1007/978-3-642-85098-1_5)
- Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of human-to-agent approach. *International Journal of Artificial Intelligence in Education*, 25(3), 380–406. <https://doi.org/10.1007/s40593-015-0042-3>
- Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *Journal of the Learning Sciences*, 14(2), 201–241. [https://doi.org/10.1207/s15327809jls1402\\_2](https://doi.org/10.1207/s15327809jls1402_2)
- Rummel, N., Spada, H., & Hauser, S. (2009). Learning to collaborate while being scripted or by observing a model. *International Journal of Computer-Supported Collaborative Learning*, 4(1), 69–92. <https://doi.org/10.1007/s11412-008-9054-4>
- Ryan, M. (2013). The pedagogical balancing act: Teaching reflection in higher education. *Teaching in Higher Education*, 18(2), 144–155.  
<https://doi.org/10.1080/13562517.2012.694104>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction*, 83, 101620. <https://doi.org/10.1016/j.learninstruc.2022.101620>



- Saleh, S. E. (2019). Critical thinking as a 21st century skill: Conceptions, implementation and challenges in the EFL classroom. *European Journal of Foreign Language Teaching, 4*(1), 1–16. <https://doi.org/10.5281/ZENODO.2542838>
- Sandars, J. (2009). The use of reflection in medical education: AMEE Guide No. 44. *Medical Teacher, 31*(8), 685–695. <https://doi.org/10.1080/01421590903050374>
- Schank, R. C. (1999). *Dynamic memory revisited* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511527920>
- Schellenbach-Zell, J., Molitor, A.-L., Kindlinger, M., Trempler, K., & Hartmann, U. (2023). Wie gelingt die Anregung von Reflexion über pädagogische Situationen unter Nutzung bildungswissenschaftlicher Wissensbestände? Die Bedeutung von Prompts und Feedback [How to stimulate reflection on pedagogical situations using educational science knowledge? The significance of prompts and feedback]. *Zeitschrift für Erziehungswissenschaft, 26*(5), 1189–1211. <https://doi.org/10.1007/s11618-023-01189-1>
- 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., & 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>
- Schmutz, J. B., Lei, Z., & Eppich, W. J. (2021). Reflection on the fly: Development of the team reflection behavioral observation (TuRBO) system for acute care teams. *Academic Medicine, 96*(9), 1337–1345. <https://doi.org/10.1097/ACM.0000000000004105>
- Schmutz, J. B., Lei, Z., Eppich, W. J., & Manser, T. (2018). Reflection in the heat of the moment: The role of in-action team reflexivity in health care emergency teams. *Journal of Organizational Behavior, 39*(6), 749–765. <https://doi.org/10.1002/job.2299>
- Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.
- Shafran, R., Bennett, S., & McKenzie Smith, M. (2017). Interventions to support integrated psychological care and holistic health outcomes in paediatrics. *Healthcare, 5*(3), 44. <https://doi.org/10.3390/healthcare5030044>

- Sherbino, J., Dore, K. L., Wood, T. J., Young, M. E., Gaissmaier, W., Kreuger, S., & Norman, G. R. (2012). The relationship between response time and diagnostic accuracy: *Academic Medicine*, *87*(6), 785–791.  
<https://doi.org/10.1097/ACM.0b013e318253acbd>
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. <https://doi.org/10.3102/0013189X015002004>
- Siebeck, M., Schwald, B., Frey, C., Röding, S., Stegmann, K., & Fischer, F. (2011). Teaching the rectal examination with simulations: Effects on knowledge acquisition and inhibition: Learning with simulation. *Medical Education*, *45*(10), 1025–1031.  
<https://doi.org/10.1111/j.1365-2923.2011.04005.x>
- Simmons, B. (2010). Clinical reasoning: Concept analysis. *Journal of Advanced Nursing*, *66*(5), 1151–1158. <https://doi.org/10.1111/j.1365-2648.2010.05262.x>
- Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2021). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist*, *57*(1), 31–54. <https://doi.org/10.1080/00461520.2021.1939700>
- Snow, R. E. (1978). Aptitude-treatment interactions in educational research. In L. A. Pervin & M. Lewis (Eds.), *Perspectives in interactional psychology* (pp. 237–262). Springer.  
[https://doi.org/10.1007/978-1-4613-3997-7\\_10](https://doi.org/10.1007/978-1-4613-3997-7_10)
- Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology*, *59*(2), 205–216. <https://doi.org/10.1037/0022-006X.59.2.205>
- Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence*, *53*, 92–101.  
<https://doi.org/10.1016/j.intell.2015.09.005>
- Stadler, M., Brandl, L., & Greiff, S. (2023). 20 years of interactive tasks in large-scale assessments: Process data as a way towards sustainable change? *Journal of Computer Assisted Learning*, *39*(6), 1852–1859. <https://doi.org/10.1111/jcal.12847>
- 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>
- Stadler, M., Radkowsch, A., Schmidmaier, R., Fischer, M., & Fischer, F. (2020). Take your time: Invariance of time-on-task in problem solving tasks across expertise levels. *Psychological Test and Assessment Modeling*, *62*(4), 517–525.  
<https://psycnet.apa.org/record/2021-30909-006>

- Stark, R., & Krause, U.-M. (2009). Effects of reflection prompts on learning outcomes and learning behaviour in statistics education. *Learning Environments Research*, 12(3), 209–223. <https://doi.org/10.1007/s10984-009-9063-x>
- Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6), 1467–1478. <https://doi.org/10.1037/0022-3514.48.6.1467>
- Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, 106(2), 331–347. <https://doi.org/10.1037/a0034752>
- Stegmann, K., Mu, J., Baum, V., & Fischer, F. (2011). The myth of over-scripting: Can novices be supported too much? In Spada, H., Stahl, G., Miyake, N., & Law, N. (Eds.), *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings. Volume I — Long Papers* (pp.406–413). International Society of the Learning Sciences. <https://repository.isls.org//handle/1/2476>
- Strauß, S., Tunnigkeit, I., Eberle, J., Vom Bover, L., Avdullahu, A., Schmittchen, M., & Rummel, N. (2023). Differential effects of a script and a group awareness tool on the acquisition of collaboration skills. In Damşa, C., Borge, M., Koh, E., & Worsley, M. (Eds.), *Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning - CSCL 2023* (pp. 75-82). International Society of the Learning Sciences. <https://doi.org/10.22318/csl2023.110098>
- Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143, 103672. <https://doi.org/10.1016/j.compedu.2019.103672>
- Suthers, D. D. (2000). Initial evidence for representational guidance of learning discourse. *In Proceedings of International Conference on Computers in Education*, November 21-24, 2000, Taipei, Taiwan.
- Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 19–30). Cambridge University Press. <https://doi.org/10.1017/CBO9780511816819.003>
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory* (1st ed.). Springer.
- Tabak, I. (2004). Synergy: A complement to emerging patterns of distributed scaffolding. *Journal of the Learning Sciences*, 13(3), 305–335. [https://doi.org/10.1207/s15327809jls1303\\_3](https://doi.org/10.1207/s15327809jls1303_3)

- Tabak, I., & Kyza, E. A. (2018). Research on scaffolding in the learning sciences. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (1st ed., pp. 191–200). Routledge.  
<https://doi.org/10.4324/9781315617572-19>
- Tay, S. W., Ryan, P., & Ryan, C. A. (2016). Systems 1 and 2 thinking processes and cognitive reflection testing in medical students. *Canadian Medical Education Journal*, 7(2), e97–e103.
- Tetzlaff, L., Edelsbrunner, P., Schmitterer, A., Hartmann, U., & Brod, G. (2023). Modeling interactions between multivariate learner characteristics and interventions: A person-centered approach. *Educational Psychology Review*, 35(4), 112.  
<https://doi.org/10.1007/s10648-023-09830-5>
- Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review*, 33(3), 863–882.  
<https://doi.org/10.1007/s10648-020-09570-w>
- Tsang, H., Park, S. W., Chen, L. L., & Law, N. (2019). Assessing collaborative problem solving: What and how? In Lund, K., Niccolai, G. P., Lavoué, E., Hmelo-Silver, C., Gweon, G., & Baker, M. (Eds.), *Proceedings of the 13th International Conference on Computer-Supported Collaborative Learning - CSCL 2019* (pp. 416-423). International Society of the Learning Sciences.  
<https://repository.isls.org/handle/1/4434>
- Tschan, F., Semmer, N. K., Gurtner, A., Bizzari, L., Spsychiger, M., Breuer, M., & Marsch, S. U. (2009). Explicit reasoning, confirmation bias, and illusory transactive memory: A simulation study of group medical decision making. *Small Group Research*, 40(3), 271–300. <https://doi.org/10.1177/1046496409332928>
- Ulitzsch, E., Ulitzsch, V., He, Q., & Lüdtke, O. (2022). A machine learning-based procedure for leveraging clickstream data to investigate early predictability of failure on interactive tasks. *Behavior Research Methods*, 55(3), 1392–1412.  
<https://doi.org/10.3758/s13428-022-01844-1>
- Van Den Boom, G., Paas, F., & van Merriënboer, J. J. G. (2007). Effects of elicited reflections combined with tutor or peer feedback on self-regulated learning and learning outcomes. *Learning and Instruction*, 17(5), 532–548.  
<https://doi.org/10.1016/j.learninstruc.2007.09.003>
- Van Laar, E., Van Deursen, A. J. A. M., Van Dijk, J. A. G. M., & De Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review.

- Computers in Human Behavior*, 72, 577–588.  
<https://doi.org/10.1016/j.chb.2017.03.010>
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring Systems. *Educational Psychologist*, 46(4), 197–221.  
<https://doi.org/10.1080/00461520.2011.611369>
- Vermunt, J. D. (2023). Understanding, measuring and improving simulation-based learning in higher education: Student and teacher learning perspectives. *Learning and Instruction*, 86, 101773. <https://doi.org/10.1016/j.learninstruc.2023.101773>
- Vogel, F., Wecker, C., Kollar, I., & Fischer, F. (2017). Socio-cognitive scaffolding with computer-supported collaboration scripts: A meta-analysis. *Educational Psychology Review*, 29(3), 477–511. <https://doi.org/10.1007/s10648-016-9361-7>
- Vogel, F., Weinberger, A., Hong, D., Wang, T., Glazewski, K., Hmelo-Silver, C. E., Uttamchandani, S., Mott, B., Lester, J., Oshima, J., Oshima, R., Yamashita, S., Lu, J., Brandl, L., Richters, C., Stadler, M., Fischer, F., Radkowsch, A., Schmidmaier, R., ... Noroozi, O. (2023). Transactivity and knowledge co-construction in collaborative problem solving. In Damşa, C., Borge, M., Koh, E., & Worsley, M. (Eds.), *Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning - CSCL 2023* (pp. 337-346). International Society of the Learning Sciences. <https://doi.org/10.22318/cscl2023.646214>
- Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes* (M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. <https://doi.org/10.2307/j.ctvjf9vz4>
- Walberg, H. J., & Tsai, S.-L. (1983). Matthew effects in education. *American Educational Research Journal*, 20(3), 359. <https://doi.org/10.3102/00028312020003359>
- Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). Springer. [https://doi.org/10.1007/978-1-4612-4634-3\\_9](https://doi.org/10.1007/978-1-4612-4634-3_9)
- Wetzels, S. A. J., Kester, L., Van Merriënboer, J. J. G., & Broers, N. J. (2011). The influence of prior knowledge on the retrieval-directed function of note taking in prior knowledge activation: Notes and prior knowledge activation. *British Journal of Educational Psychology*, 81(2), 274–291. <https://doi.org/10.1348/000709910X517425>
- Wise, A. F., & Schwarz, B. B. (2017). Visions of CSCL: Eight provocations for the future of the field. *International Journal of Computer-Supported Collaborative Learning*, 12(4), 423–467. <https://doi.org/10.1007/s11412-017-9267-5>

- 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>
- Woods, N. N. (2007). Science is fundamental: The role of biomedical knowledge in clinical reasoning: Clinical expertise. *Medical Education*, *41*(12), 1173–1177. <https://doi.org/10.1111/j.1365-2923.2007.02911.x>
- Yudkowsky, R., Park, Y. S., Hyderi, A., & Bordage, G. (2015). Characteristics and implications of diagnostic justification scores based on the new patient note format of the USMLE Step 2 CS exam: *Academic Medicine*, *90*, S56–S62. <https://doi.org/10.1097/ACM.0000000000000900>
- Zambrano, J., Kirschner, F., Sweller, J., & Kirschner, P. A. (2019). Effects of prior knowledge on collaborative and individual learning. *Learning and Instruction*, *63*, 101214. <https://doi.org/10.1016/j.learninstruc.2019.05.011>
- Zhang, Z., Bekker, T., Markopoulos, P., & Skovbjerg, H. M. (2023). Supporting and understanding students' collaborative reflection-in-action during design-based learning. *International Journal of Technology and Design Education*, *34*(1), 307–343. <https://doi.org/10.1007/s10798-023-09814-0>
- Ziegler, E., Edelsbrunner, P. A., & Stern, E. (2021). The benefit of combining teacher-direction with contrasted presentation of algebra principles. *European Journal of Psychology of Education*, *36*(1), 187–218. <https://doi.org/10.1007/s10212-020-00468-3>
- Ziv, A., Wolpe, P. R., Small, S. D., & Glick, S. (2003). Simulation-based medical education: An ethical imperative. *Academic Medicine*, *78*(8), 783–788. <https://doi.org/10.1097/00001888-200308000-00006>
- Zottmann, J. M., Dieckmann, P., Taraszow, T., Rall, M., & Fischer, F. (2018). Just watching is not enough: Fostering simulation-based learning with collaboration scripts. *GMS Journal for Medical Education*; *35*(3). <https://doi.org/10.3205/ZMA001181>



## **8 Appendices**

---

As all the studies were conducted with German medical students, the original material was in German. For the appendices of the thesis, parts of the material (especially the case material, the knowledge tests, and the reflection interventions) have been translated into English to make them accessible to all readers.

## Appendix A: Case Material

**Table A**

*Overview of the Patient Cases Used in All Studies*

Case	Diagnosis	Usage		
		Study 1	Study 2	Study 3
Marianne Freundorf	Acute pancreatitis	/	/	Patient case 2
Ute Wenninger	Sigmoid diverticulitis	/	/	Patient case 4
Herma Goettlich	Aspiration pneumonia	Pretest case	Pretest case	Patient case 1
Anton Fomin	Acute tuberculosis	Intervention case 2	Intervention case 1	
Mark Binder	Pneumocystis jirovecii Pneumonia (PJP)	Intervention case 3	Intervention case 2	Patient case 3
Maria Schenker	Hospital acquired pneumonia	Test case	Intervention case 3	Patient case 5
Sabine Winkler	Community acquired pneumonia (CAP)	Intervention case 1	Test case	

Each case has three parts. The structure of each case is described in the following example case (patient name: Herma Goettlich). The original material was developed in German and translated into English to make it accessible to all interested readers.

### ***Part I: Health Record***

#### **Introduction**

You have been working for several months at a medium-sized district hospital and are currently assigned to a general internal medicine ward. Today, you are also assisting in the emergency department. Late Monday morning, 78-year-old Herma Goettlich is brought in by the emergency medical services, accompanied by her concerned husband. Mrs. Goettlich is suffering from severe shortness of breath, so her husband answers most of your questions. You have taken blood samples and sent them ‘urgently’ to the laboratory, obtained as much medical history from Mr. Goettlich as possible, and conducted an examination. By the time you finish, the laboratory results are also ready, allowing you to review the patient’s file and consider the next diagnostic steps.

#### **Emergency Medical Services Report**

78-year-old patient with fever since this morning and rapidly worsening shortness of breath. Improvement of symptoms with 2 liters of oxygen; decision made to postpone intubation for now. Dysphagia with a history of stroke. *Medication:* Aspirin protect, ramipril, simvastatin, calcium/D3.

**Medical History**

Mr. Goettlich reports that his wife has been experiencing significant shortness of breath and a worsening fever since this morning. Everything was fine yesterday. They watched *Tatort* together and then went to bed. Normally, she has no lung issues and is generally in excellent internal health. Upon inquiry, Mr. Goettlich mentions that his wife has had swallowing difficulties since her stroke a few months ago and occasionally chokes. This happened last night as well, but he doesn't consider it worse than usual. There are no B symptoms.

***Pre-existing Conditions***

- History of media infarction (middle cerebral artery infarction) in December 2017, resulting in residual right hemiparesis
- Osteoporosis
- Early stage of dementia syndrome
- History of tonsillectomy in 1962

***Medications***

Aspirin protect, ramipril, simvastatin, calcium/D3

***Substance Use History***

Approximately 10 pack-years of smoking, quit 40 years ago. Alcohol consumption is rare.

***Social History***

Retired, formerly worked as a butcher's assistant.

**Physical Examination**

78-year-old patient with decreased general condition and good general appearance (height: 1.75 m, weight: 72 kg, BMI: 23.5 kg/m<sup>2</sup>).

***Vital signs***

Blood pressure 100/60 mmHg, heart rate 100/min regular, temperature 37.9°C, respiratory rate 27/min, oxygen saturation 96% on 2 liters of oxygen. Lymph nodes not enlarged, non-tender. Thyroid gland is unremarkable.

***Cardiovascular system***

No cyanosis. Heart sounds clear, regular, and tachycardic, with no extra sounds or pathological heart murmurs. No jugular venous distention. Moderate bilateral leg edema, slightly more on the right than on the left. Peripheral pulses are palpable bilaterally. Mucous membranes are unremarkable.

***Respiratory system***

Symmetrical chest expansion, no retractions, normal thoracic shape. No vocal fremitus, no stridor. Diaphragmatic excursion equal at 4 cm bilaterally, with no dullness to percussion. Lungs evenly ventilated, with coarse breath sounds throughout, cough with foul-smelling sputum, no pleural rub.

***Abdomen***

Abdominal wall soft, non-tender, no masses, no guarding, bowel sounds normal in all quadrants. Kidneys not tender to palpation, spleen not palpably enlarged, liver 11 cm in the right midclavicular line, smooth surface. No hernias. No visible surgical scars.

***Skin***

Unremarkable skin findings. Extremities warm, no varicose veins. No nail abnormalities.

***Musculoskeletal system***

Normal range of motion in all joints. No joint pain, swelling, or deformities. Spine non-tender to percussion.

***Neurological examination***

Friendly, cooperative, oriented in all aspects, no evidence of formal thought disorder or suicidality. Pupillary light reflex direct and consensual prompt and equal. Known right hemiparesis and facial paresis. No other weakness, no sensory deficit, no pathological reflexes, no drop in manual muscle testing. No signs of meningeal irritation. Vibration sensation intact 8/8 in all four extremities.

**Laboratory**

<b>Parameter</b>	<b>Value</b>	<b>Reference range (women)</b>
<i>Blood Count</i>		
Erythrocytes	$3.8 \times 10^6 / \mu\text{l}$	$3.5 - 5 \times 10^6 / \mu\text{l}$
Hemoglobin (Hb)	13.6 g/dl	12 - 15 g/dl
MCH	28 pg	27 - 34 pg
MCV	84 fl	81 - 100 fl

MCHC	33 g/dl	32 - 36 g/dl
Hematocrit (Hkt)	38%	33 - 43 %
Leukocytes	13.6 x 10 <sup>3</sup> / $\mu$ l	4 - 11 x 10 <sup>3</sup> / $\mu$ l
Platelets	182,000 / $\mu$ l	150,000 - 400,000 / $\mu$ l
Reticulocytes	1%	0.5 - 2 %
<i>Differential Blood Count</i>		
Neutrophilic Granulocytes	78%	45 - 78 %
Stab Cells	4%	0 - 4 %
Segmented Cells	74%	45 - 74 %
Eosinophilic Granulocytes	1%	0 - 7 %
Basophilic Granulocytes	1%	0 - 2 %
Lymphocytes	16%	16 - 45 %
Monocytes	4%	4 - 10 %
<i>Coagulation</i>		
Quick	100%	70 - 120%
INR	1	1
PTT	38 sec.	28 - 40 sec.
<i>Serum</i>		
Sodium	142 mmol/l	136 - 148 mmol/l
Potassium	4.7 mmol/l	3.6 - 5.2 mmol/l
Calcium (total)	2.3 mmol/l	2.1 - 2.6 mmol/l
Creatinine	0.9 mg/dl	< 0.9 mg/dl
eGFR	>60 ml/min/1.73 m <sup>2</sup>	>60 ml/min/1.73 m <sup>2</sup>
Urea	>60 ml/min/1.73 m <sup>2</sup>	>60 ml/min/1.73 m <sup>2</sup>
Alkaline Phosphatase	21 mg/dl	10 - 50 mg/dl
Bilirubin (total)	45 U/l	40 - 190 U/l
Bilirubin (direct)	1 mg/dl	< 1.1 mg/dl
CHE	0.6 mg/dl	< 0.6 mg/dl
GOT (AST)	4.6 kU/l	2.5 - 7.4 kU/l
GPT (ALT)	13 U/l	< 15 U/l
$\gamma$ -GT	8 U/l	< 17 U/l
$\alpha$ -Amylase	14 U/l	< 18 U/l
Lipase	22 U/l	10 - 53 U/l
Blood Sugar	89 U/l	< 190 U/l
HbA1c	89 mg/dl	55 - 100 mg/dl
CK	5.40%	4 - 6 %
CK-MB	34 U/l	< 80 U/l
CRP	4 U/l	< 10 U/l
Ferritin	53 mg/l	< 6 mg/l
TSH basal	83 $\mu$ g/l	15 - 250 $\mu$ g/l
Erythrocyte Sedimentation Rate	1.8 $\mu$ U/ml	0.2 - 3.1 $\mu$ U/ml
<i>Urine-Stick</i>		
pH	5	5-7
Protein	-	-
Bilirubin	-	-
Urobilinogen	-	-
Nitrite	-	-
Glucose	-	-
Acetone	-	-
Blood	-	-



*Part II: Request Form for Interaction with the Agent-Based Radiologist*

**Figure A1**

*Screenshot from the Request Form for Radiological Examinations in the Simulation*

**Patient**  
 Vorname: Herma  
 Nachname: Göttlich  
 Geburtsdatum: 27.01.1940

**Examination**

<b>Method</b>	<b>Body part</b>	<b>Contrast agent</b>
<input checked="" type="radio"/> CT <input type="radio"/> Röntgen <input type="radio"/> MRT <input type="radio"/> Ultraschall	<input type="radio"/> Schädel <input type="radio"/> Thorax <input checked="" type="radio"/> Abdomen <input type="radio"/> Schulter r. <input type="radio"/> Oberarm r.	<input checked="" type="checkbox"/> ja <input type="checkbox"/> nein <input type="radio"/> Ellenbogengelenk r. <input type="radio"/> Herz <input type="radio"/> Wirbelsäule <input type="radio"/> Gefäße <input type="radio"/> Hals

**Fragestellung und Angaben an den Radiologen**

**Previous findings**

- + Anamnese
- + Vorerkrankungen
- + Rettungsdienstprotokoll
- + Labor
- + Körperliche Untersuchung allgemein
- + Vitalparameter
- + Lymphknoten
- + kardiovaskulär
- + respiratorisch
- + Abdomen
- + Haut
- + Bewegungsapparat
- + neurologisch

**Suspected diagnosis** - Verdacht auf:

pnj  
 Pneumonie/Lungenentzündung, Aspirationspneumonie  
 Pneumonie/Lungenentzündung, atypisch  
 Pneumonie/Lungenentzündung, bakteriell  
 Pneumonie/Lungenentzündung, begleitend bei systemischem Wurmbefall  
 Pneumonie/Lungenentzündung, CAP  
 Pneumonie/Lungenentzündung, HAP

Long menu: 249 diagnoses available

Submit

Evidence Elicitation (EE)  
 Evidence Sharing (ES)  
 Hypothesis Sharing (HS)

**Part III: Case Solution****Figure A2**

*Screenshot of the Simulation When Entering the Final Diagnosis.*

**Wählen Sie bitte Ihre abschließende Diagnose aus.**  
 Bitte formulieren Sie die Diagnose so spezifisch wie möglich (z.B. "Restriktive Lungenerkrankung bei Skoliose" statt "Restriktion").

**Final diagnosis**

Bitte geben Sie Ihre Antwort in das Textfeld ein und selektieren Sie dann einen Begriff.

Long menu: 249 diagnoses available

- Abdomen
- Abdomenleeraufnahme
- Abgeschlagenheit
- Absolute Arrhythmie bei Vorhofflimmern
- Abszess
- Abszess, Weichteilabszess
- Abszess, Weichteilabszess mit Knochenbeteiligung
- Abszess, Weichteilabszess ohne Knochenbeteiligung

 Drawing Conclusions (DC)


**Figure A3**

*Screenshot of the Simulation When Justifying the Final Diagnosis.*

Please justify your diagnosis.

**Unbewertete Freitextantwort**

Bitte geben Sie Ihre Antwort in das Textfeld ein (max. 4000 Zeichen).

Free-text field for justification  Drawing Conclusions (DC)

## Appendix B: Coding Schemes

### *B1: Coding Manual for Diagnostic Outcomes*

In principle, the same coding schemes for the diagnostic outcomes (diagnostic accuracy and diagnostic justification in Studies 1 and 2; diagnostic accuracy in Study 3) were used for all patient cases. However, during the course of the studies (chronological order: Study 3, Study 1, Study 2), the coding schemes were revised and further improved in collaboration with the medical experts among the project members. Therefore, the coding schemes for the same cases differ slightly between the studies. The final version of the most recently revised coding scheme is provided below for the previously used example case.

**Table B1**

*Coding Scheme for Diagnostic Accuracy and Justification for an Example Case*

<b>Main Diagnosis</b>	<b>Synonyms</b>	<b>Points</b>
Aspiration pneumonia	Aspiration pneumonia	1
Pneumonia	Bacterial pneumonia, community-acquired pneumonia	0.5
CAP Pneumonia		0.5
Atypical pneumonia	Lobar pneumonia	0.5

<b>Justification</b>	<b>Synonyms</b>	<b>Points</b>
Dyspnea	Shortness of breath, difficulty breathing, short of breath	1
Tachypnea	Increased respiratory rate (RR); RR 27/min	1
Fever	Fever since this morning, sweating, specific temperature measurements, subfebrile temperature	1
Decreased SpO <sub>2</sub>	SpO <sub>2</sub> 92%, hypoxia, reduced oxygen saturation, 2l O <sub>2</sub>	1
Cough with foul-smelling sputum	Foul-smelling sputum, cough with purulent sputum, productive cough, excluded: cough alone	1
Dysphagia	Swallowing disorder, history of stroke with hemiparesis	1
Coarse crackles	Rales, mainly on the right	1
Elevated inflammatory markers	Leukocytosis, elevated leukocytes, leukocytes: 13.6x10 <sup>3</sup> /μl; Elevated CRP, CRP: 53 mg/l; elevated ESR, ESR: 10/23 mm; Infection markers, inflammatory markers, signs of infection (with reference to laboratory)	1
Chest X-ray/CT: Consolidations or infiltrations	CT thorax findings: consolidations in both right and left lower lobes; chest X-ray: Reticular consolidations in the right lower lobe; chest X-ray findings: striped consolidations in the lower lobe; Increased markings/shadows/infiltrate; Excluded: correlation, lower lobe abnormalities, lower lobe involvement	1
Maximum Points		9

*Note.* The original coding scheme was developed and applied in German and translated into English for transparency.

## ***B2: Metrics and Sample Solutions for Collaborative Diagnostic Activities***

### **Performance in Evidence Sharing**

In Study 1, the performance in evidence sharing was measured using a sensitivity score that indicated how much of the total relevant evidence for a case was shared by the participant. In Study 2, the performance was measured using a precision score that indicated how much of the evidence that the participants shared with the radiologist was actually relevant to the radiologist. Below is a list of all available evidence from the example case presented earlier (Herma Goettlich) in the original German language. The evidence relevant to the radiologist is in bold. Depending on the diagnoses shared by the participant, some of the relevant evidence should be shared, and some should not. If the participant did not share any diagnoses, all relevant evidence in bold had to be shared to receive a sensitivity score of 1.

- Atemnot
- Schneller Beginn
- Beginn heute morgen
- Z.n. Nikotinabusus 10 py
- Medikation mit ASS protect
- Medikation mit Ramipril
- Medikation mit Simvastatin
- Medikation mit Calcium/D3
- Überwiegend im Rollstuhl mobilisiert
- Gewichtsverlust 8 kg
- Starkes Schwitzen
- Körperliche Unruhe
- **Keine bekannten Allergien, auch nicht auf Medikamente oder Kontrastmittel**
- Z.n. Mediainfarkt vor 6 Wochen
- Residuale Hemiparese rechts
- Osteoporose
- Beginnendes dementielles Syndrom
- Z.n. Tonsillektomie 1962
- Z.n. Tiefer Beinvenenthrombose rechts 2005
- Fieber seit heute morgen
- Akut einsetzende Luftnot

- Dysphagie
- Z.n. Stroke
- **EKG unauffällig**
- **Trop T Schnelltest unauffällig**
- pO<sub>2</sub> initial 92 %
- Leukozyten 13,6 x 10<sup>3</sup>/μl
- CRP 53 mg/dl
- Blutsenkung 10/23
- **TSH 1,8 μU/ml**
- **eGFR > 60 ml/min/1,73 m<sup>2</sup> KOF**
- 78-jährige Patientin
- reduzierter AZ
- **guter EZ**
- **BMI 23,5 kg/m<sup>2</sup>**
- **RR 105/60 mmHg**
- Puls 102/min.
- **Puls regelmäßig**
- Temp. 37,9°C
- AF 27/min
- pO<sub>2</sub> 96 % unter 2 l O<sub>2</sub>
- **Keine vergrößerten Lymphknoten tastbar**
- **Keine Zyanose**
- **Herztöne rein**
- **Herztöne regelmäßig**
- Herztöne tachykard
- **keine Extratöne oder pathologische Herztöne**
- **Keine Jugularvenenstauung**
- Mäßige Unterschenkelödeme
- Unterschenkelödeme rechts > links Seitendifferenz 2 cm
- **Periphere Pulse seitengleich tastbar**
- **Schleimhäute unauffällig**
- **Symmetrische Thoraxexkursion**
- **keine Einziehungen am Thorax**
- **normale Thoraxform**

- **Kein Stimmfremitus**
- **kein Stridor**
- **Gleichstand der Zwerchfelle**
- **Zwerchfelle bilateral 4 cm atemverschieblich**
- **kein Hinweis auf Pleuraerguss**
- **Lunge ubiquitär belüftet**
- Lunge mit grobblasigen RGs rechts
- Husten mit Auswurf
- Auswurf übelriechend
- **Kein Pleurareiben**
- **Bauchdecke weich**
- **Abdomen nicht druckschmerzhaft**
- **Abdomen ohne Resistenzen**
- **Abdomen ohne Abwehrspannung**
- **Darmgeräusche regelrecht in allen Quadranten**
- **Nieren nicht klopfschmerzhaft**
- **Milz nicht vergrößert tastbar**
- **Leber 11 cm in der rechten MCL**
- **Leber mit glatter Oberfläche**
- **Keine Hernien**
- **Keine sichtbaren Operationsnarben**
- **Unauffälliger Hautbefund**
- **Extremitäten warm**
- **Keine Varikosis**
- **Keine Nagelveränderungen**
- **Normale Beweglichkeit der Gelenke**
- **Keine Gelenkschmerzen**
- **Keine Gelenkschwellungen**
- **Keine Gelenkdeformitäten**
- **Wirbelsäule nicht klopfschmerzhaft**
- **Meyer-Homanns-Payr-Zeichen negativ**
- **Freundlich zugewandt**
- **Agitiert**
- **In allen Qualitäten orientiert**



- **kein Hinweis auf formale Denkstörungen**
- **Kein Hinweis auf Suizidalität**
- **Pupillenlichtreaktion direkt und indirekt prompt und seitengleich**
- Bekannte Hemiparese rechts
- **Kein Meningismus**
- **Vibrationsempfinden 8/8 an allen vier Extremitäten**

#### **Performance in Hypothesis Sharing**

In both Studies 1 and 2, hypothesis sharing was measured using a precision score indicating how many of the hypotheses (diagnoses) that the participants shared with the radiologist were actually relevant to the case. All relevant diagnoses for the example case are listed in the following in the original German language.

- Alveolitis
- Alveolitis, exogen allergisch (EAA)
- Autoimmunes Geschehen
- Bronchitis
- Bronchitis, bakteriell akut
- Bronchitis, viral akut
- COPD
- COPD, akut exazerbiert
- COPD, chronisch
- Degeneratives Geschehen
- Entzündliches Geschehen
- Grippaler Infekt
- Herzinsuffizienz
- Herzinsuffizienz, akut bei Myokardinfarkt/Herzinfarkt
- Herzinsuffizienz, akut bei Myokarditis
- Herzinsuffizienz, chronisch, akut dekompensiert
- Infekt
- Infekt, bakteriell
- Infekt, viral
- Influenza/Grippe
- Ischämie, Lungenarterienembolie, Lungenembolie
- Mykobakteriose, atypisch

- 
- Pneumonie/Lungenentzündung
  - Pneumonie/Lungenentzündung, Aspirationspneumonie
  - Pneumonie/Lungenentzündung, atypisch
  - Pneumonie/Lungenentzündung, bakteriell
  - Pneumonie/Lungenentzündung, begleitend bei systemischem Wurmbefall
  - Pneumonie/Lungenentzündung, CAP
  - Pneumonie/Lungenentzündung, Pilzpneumonie
  - Pneumonie/Lungenentzündung, Pneumocystis jirovecii Pneumonie (PCP)
  - Pneumonie/Lungenentzündung, viral
  - Pneumothorax
  - Pneumothorax, spontan
  - Pneumothorax, traumatisch
  - Rheumatisches Fieber
  - Sarkoidose
  - Sepsis/Blutvergiftung
  - Thrombose, tiefe Beinvenenthrombose (TVT)

## Appendix C: Knowledge Tests

### *CI: Content Knowledge*

Prior content knowledge was assessed in Studies 1 and 2 by conceptual (Boshuizen & Schmidt, 1992) and strategic knowledge (Stark et al., 2011) of radiology and internal medicine, respectively.

#### **Conceptual Knowledge**

Conceptual knowledge was measured with single-choice items focusing on pathophysiology, disease triggers, and radiologic interpretation. Below is an example item from internal medicine:

*Which of the following statements about pneumonia is most likely true?*

- 1) Mycoplasmas are strictly intracellular pneumonia pathogens.
- 2) In elderly, multimorbid patients, pneumonia usually begins more abruptly with a high fever.
- 3) Legionella is the most common cause of bronchopneumonia.
- 4) Respiratory rate measurement is an important parameter for assessing the severity of the disease and for quality assurance.
- 5) The typical pathogen of community-acquired pneumonia is Haemophilus influenza.

#### **Strategic Knowledge**

Strategic knowledge was measured by text-based cases using the key feature approach (M. R. Fischer et al., 2005). Key feature cases capture clinical knowledge and skills in multiple steps. The following is an example item from internal medicine: It is Tuesday afternoon in the general practitioner's office where you work as a resident physician. 72-year-old Dieter Klemenz comes in to see you. He complains of a severe cough he has been experiencing for several days. The cough is painful and uncontrollable and has even led to vomiting. Previously, he had a minor infection with an elevated temperature of around 38°C (100.4°F), rhinitis, and what he describes as a "normal cough."

*What is your most likely suspected diagnosis?*

- 1) Bronchitis
- 2) COLD (chronic obstructive lung disease)
- 3) Common cold
- 4) Pertussis (correct)**
- 5) Dry pleurisy

- 6) Pneumonia
- 7) Tuberculosis
- 8) Typhoid fever

*Please assume that the patient has influenza. What diagnostic test will you order to confirm the diagnosis?*

- 1) Blood gas analysis
- 2) Blood cultures
- 3) Complete blood count (CBC)
- 4) IgM in serum
- 5) Basic blood count
- 6) CRP in serum
- 7) Nasopharyngeal swab (correct)**
- 8) Pulse oximetry

*Influenza was confirmed. What is the most important measure now?*

- 1) Bronchoscopy
- 2) Thoracic CT for risk stratification
- 3) Symptomatic measures (correct)**
- 4) Hospital admission
- 5) Checking the vaccination record
- 6) Non-disclosure report to the health department
- 7) Isolation
- 8) Oral antibiotic therapy, e.g., with amoxicillin + clavulanic acid

### *References*

- 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](https://doi.org/10.1016/0364-0213(92)90022-M)
- Fischer, M. R., Kopp, V., Holzer, M., Ruderich, F., & Jünger, J. (2005). A modified electronic key feature examination for undergraduate medical students: Validation threats and opportunities. *Medical Teacher*, *27*(5), 450–455. <https://doi.org/10.1080/01421590500078471>
- 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>

## **C2: Collaboration Knowledge**

Prior collaboration knowledge is based on meta-cognitive knowledge, which is information about the collaborators' knowledge, roles, and tasks that is critical for successful collaboration (Engelmann & Hesse, 2011). Collaboration knowledge was measured in Studies 1 and 2 with seven text-based patient cases with the leading symptoms of ascites, joint pain, impaired vigilance, B symptoms (fever, night sweats, and weight loss), back pain, dyspnea, and weakness, which combined required a radiological examination in the next step of the diagnostic workup. An example case follows.

### **Introduction**

28-year-old Ulf Schäfer was found lying in front of a ladder. He had a contusion on his left forehead and abrasions on the left side of his body. Mr. Schäfer appears absent, does not respond appropriately to speech, and has vomited multiple times since being admitted to the emergency room. Only in response to a painful stimulus does he open his eyes and deliberately ward it off. Anisocoria is observed, with the left pupil reduced and the right pupil slim. The patient breathes shallowly, with a respiratory rate of 20/min, pulse 90/min, and blood pressure 100/65 mmHg. Lungs are ventilated on all sides, abdomen is soft, and extremities are unremarkable upon inspection.

**Patient:** Ulf Schäfer

**Date of birth:** November 3, 1991

**Examination:** Emergency CCT

*From the information provided below, please select the details that you would communicate to a radiologist for the above-mentioned examination.*

Item	Category	Correctness	
1	<b>Condition after fall from ladder</b>	Cause	1
2	<b>Impaired vigilance</b>	Additional information	1
3	<b>Multiple episodes of vomiting</b>	Additional information	1
4	Reduced left eye aperture, right eye slim	Additional information	0
5	Shallow breathing	Additional information	0
6	<b>Contusion on the left forehead</b>	Physical examination	1
7	Abrasions on the left side	Physical examination	0
8	Respiratory rate 20/min	Physical examination	0
9	Pulse 90/min	Physical examination	0
10	Blood pressure 100/65 mmHg	Physical examination	0
11	Extremities inspection unremarkable	Physical examination	0
12	<b>Anisocoria</b>	Physical examination	1

*References*

- Engelmann, T., & Hesse, F. W. (2011). Fostering sharing of unshared knowledge by having access to the collaborators' meta-knowledge structures. *Computers in Human Behavior*, 27(6), 2078–2087. <https://doi.org/10.1016/j.chb.2011.06.002>



## Appendix D: Reflection Guidance

### *D1: Reflection on Individual Activities (Study 1)*

The questions for reflection on the individual activities were adopted from Mamede et al. (2014) and were implemented in the diagnostic process as follows:

#### **Instructions for the Participants**

Before we continue with the diagnostic process, we would like to ask you to take a few moments to reflect on your previously suspected diagnoses. Please answer the questions below in the free text box.

**Step 1:** Please state your most likely current suspected diagnosis.

**Step 2:** What symptoms and findings support your current suspected diagnosis?

**Step 3:** What symptoms and findings contradict your current suspected diagnosis?

**Step 4:** What other symptoms and findings would you have expected in this case if this suspected diagnosis were correct, and which were missing?

**Step 5:** Please provide an alternative suspected diagnosis.

**Step 6:** What is your most likely suspected diagnosis? List your diagnoses in descending order, starting with the most likely.

#### *References*

Mamede, S., van Gog, T., Sampaio, A. M., de Faria, R. M. D., Maria, J. P., & 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>

### *D2: Reflection on Collaborative Activities (Study 2)*

The reflection guidance on collaborative activities based on the *script theory of guidance* (F. Fischer et al., 2013) can be found below. All participants received the same introduction to the reflection phase. Afterwards, the participants in the low-structured conditions received questions at the scene level, and the participants in the high-structured conditions received questions at the scriptlet level.

#### **Introduction**

You have just collaborated with your colleague, Dr. Schmidt, from the Radiology Department. You may now be wondering how successful the collaboration was. Reflecting on how well you worked with your colleague is crucial in the collaborative diagnostic process. It

helps you to better understand your own activities and improve them. Before moving on to the case solution, we ask you to take a moment to reflect on your collaboration with Dr. Schmidt.

Below are questions regarding:

- 1) Choice of radiological examinations
- 2) Sharing information from medical records with radiology
- 3) Sharing diagnoses with radiology

Please answer all the questions in writing. Feel free to use bullet points in your answers.

Please make your best effort to answer the questions.

### *References*

Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. *Educational Psychologist*, 48(1), 56–66. <https://doi.org/10.1080/00461520.2012.748005>

**Table D2**

## Overview of the Questions for Low- and High-Structured Reflection on Collaborative Activities

<b>Collaboration script component</b>		<b>Low-structured reflection</b>	<b>High-structured reflection</b>
Collaborative activities (Scenes)	Scriptlets	Scene-level questions	Scriptlet-level questions
<b>Evidence Elicitation</b> Choice of radiological examination	<b>Identify missing evidence and request</b> it from the collaboration partner	<i>Did the tests you requested help you make a diagnosis?</i>	<i>Did you obtain the information you needed from the radiological tests you requested?</i>
<i>The following questions are intended to help you think about the radiological exams you have requested.</i>	<b>Evaluate requested evidence</b>		<i>Did the radiological information help support or refute your suspected diagnosis?</i>
		<i>What could you improve about the test request in the future?</i>	
<b>Evidence Sharing</b> Sharing information from medical records with radiology	Identify and share <b>evidence relevant</b> to the collaboration partner Identify and share evidence <b>critical</b> to the collaboration partner	<i>Has sharing information from the medical record with radiology been helpful in your diagnostic process?</i>	<i>Have you provided enough important information from your collaboration partner (radiologist)?</i>
<i>The following questions are intended to help you think about the patient information you have shared with the radiologist.</i>	Distinguish <b>irrelevant evidence</b> from <b>relevant evidence</b>		<i>Have you provided your collaboration partner (radiologist) with all the information from the medical record they need to conduct high-risk examinations?</i>
		<i>What could you improve about information sharing in the future?</i>	<i>When sharing information from the medical record, have you differentiated what is important to your collaboration partner (radiologist) and what is less important?</i>
<b>Hypothesis Sharing</b> Sharing diagnoses with radiology	<b>Targeted sharing of suspicion and exclusion hypotheses</b>	<i>Has sharing the hypotheses with radiology been useful in your diagnostic process?</i>	<i>Based on the medical record, you have generated suspicion or exclusion hypotheses and communicated some or all of them to your collaboration partner (radiologist)? What considerations led to these choices?</i>
<i>The following questions are intended to help you think about the diagnoses you have shared with the radiologist.</i>	<b>Consider</b> collaborator's <b>contributions</b> to suspected <b>hypotheses</b>		<i>Did you allow your collaboration partner (radiologist) to participate in validating or eliminating the hypotheses you shared with them?</i>
		<i>What could you improve about hypothesis sharing in the future?</i>	

*Note.* The first column contains the scenes defined according to the script theory of guidance (Fischer et al., 2013), namely, the collaborative activities. All learners were given information about which scene they should reflect on (explanation in italics in column 1). The second

column contains the scriptlets theoretically assigned to the scenes (i.e., the sub-activities necessary to successfully complete the corresponding collaborative activity). Learners in the low-structured condition received reflection questions at the scene level (column 3), and learners in the high-structured condition received reflection questions at the scriptlet level (column 4).

## Appendix E: Collaboration Script (Study 1)

The collaboration script contained three prompts and was adopted from Radkowsch et al. (2021). The first prompt was static and presented at the beginning of the interaction with the agent-based radiologist. The second prompt was adaptive and presented whenever learners did not sufficiently justify their radiological request with patient information. The third prompt was adaptive and presented whenever the learners requested a radiologic test that did not match the stated suspected diagnosis. The contents of the prompts are described in detail below.

### *Prompt 1 (static)*

Hello, this is your radiologist on duty again. For us to work well together, I would like to remind you of what is most important to me about your request. **I am particularly concerned about the following aspects of your request. First, is the requested examination sufficiently justified?** This means weighing the benefits of the test against the harm it may cause the patient, for which I am liable. The information you give me about the patient will serve as the basis for this weighting. **I will “translate” your request into a work order:**

- What should I look for?
- Where should I look?

The following information will help me:

- main symptoms
- course of symptoms
- suspected diagnoses
- key laboratory and physical findings
- information about the patient that is important for performing the examination

Certain information about the patient will make it easier for me to perform and interpret the images, so please remember: **Which of your details are particularly valuable from a radiological point of view?** Please try to include as much relevant information as possible in your request, and remember that this information is important for our collaboration on all patient cases. If you have specific questions about the radiological findings, you can learn more about the radiological signs under “Request more information about findings.”

I look forward to our productive collaboration!

***Prompt 2 (adaptive): Example of an insufficiently justified contrast agent CT***

You have requested a CT scan with a contrast agent. **Please be aware** that this examination involves various risks:

- a. Because of the radiation risk (oncogenic, teratogenic), the patient's age is important to me. For female patients of childbearing age, I need to know whether they are **pregnant** or **not**. If an examination with little or no radiation provides equally meaningful results, I would prefer it.
- b. **Allergic reactions**, including anaphylaxis, may occur as a reaction to **iodinated contrast media**. Therefore, it is important for me to know if the patient has had any problems with this in a previous examination and to be informed about allergic conditions.
- c. Iodinated contrast media may cause **problems with kidney function** and **thyroid function**. Thyroid dysfunction may result in a thyrotoxic crisis. Contrast media may cause renal failure. I am interested in the patient's current organ functioning, especially the **eGFR** and **TSH levels**.

If this test is required, please remember to provide me with all this information.

***Prompt 3 (adaptive): Example of stating pneumonia/lung inflammation as a suspected diagnosis while requesting contrast CT***

You did not provide a sufficient reason for the examination, or you selected the wrong examination. Would you like to revise your request?

- Please remember to check that you have given me the most valuable information from a radiological point of view. **I explained what information is relevant in the welcome message.**

Please include as much relevant information as possible in your request.

- You proposed “pneumonia/lung inflammation, aspiration pneumonia” as the suspected diagnosis.

**1) First question: As a radiologist, the first question I ask myself is whether there is a decrease in transparency, and if so, in what pattern.** Air provides a good contrast to tissue parts in the lung (inflammatory changes, fluid-filled) for the first assessment. In terms of step-by-step diagnostics and radiation protection, one begins with an X-ray because it is readily available and therefore can be well-evaluated over time. However, due to the overlap in the summation image, a detailed assessment cannot be made.

- 2) **Next:** If pneumonia is still suspected and the assessment is unclear, the next step is detailed imaging of the air-filled parts of the lung compared with the fluid-filled parenchyma. CT is suitable for this, and no contrast media are needed.
- 3) **Other questions, optional:** Pleural effusion? Especially in the case of superficially accessible effusions, e.g., in the recessus costodiaphragmaticus, sonography is very sensitive in showing even small amounts of fluid in the pleural cavity.

Do you want to revise your request?

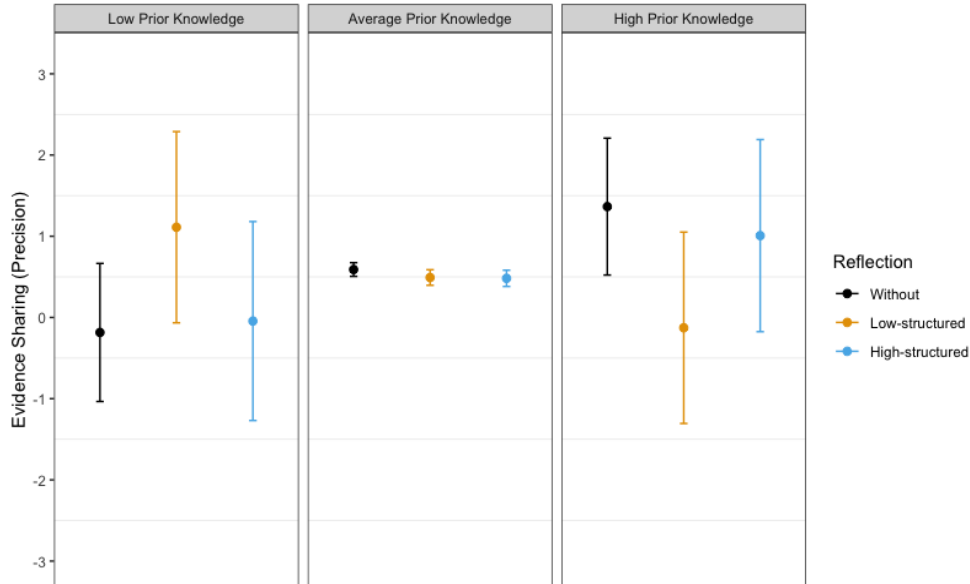
### *References*

- Radkowsch, A., Sailer, M., Schmidmaier, R., Fischer, M. R., & Fischer, F. (2021). Learning to diagnose collaboratively – Effects of adaptive collaboration scripts in agent-based medical simulations. *Learning and Instruction, 75*, 101487.  
<https://doi.org/10.1016/j.learninstruc.2021.101487>



**Appendix F: Additional Graphs for the Inferential Statistics in Study 2****Figure F1**

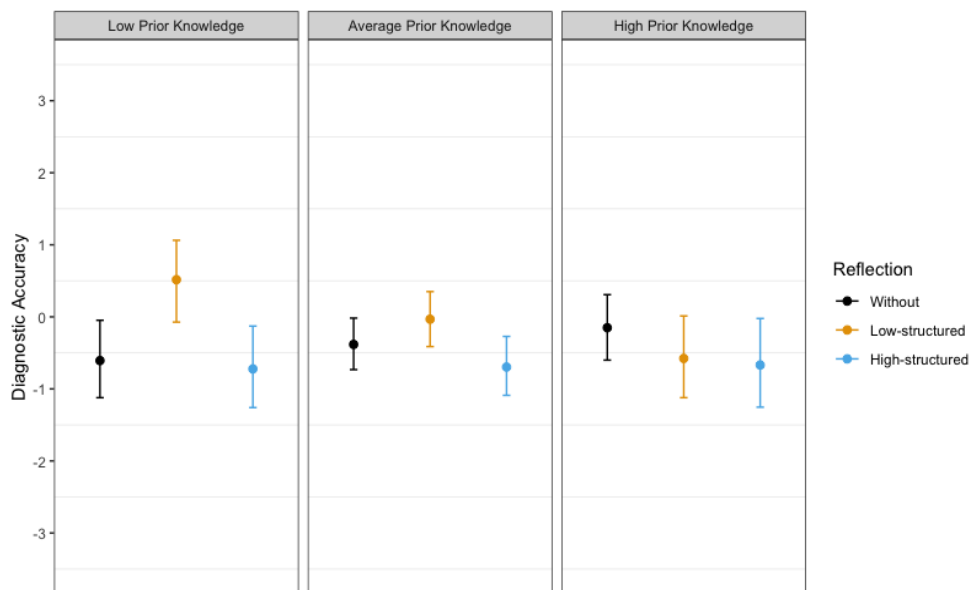
*Effects of Low- and High-Structured Guidance for Reflection on Collaborative Activities on the Performance in Evidence Sharing*



*Note.* Estimated means per group are shown as dots, accompanied by confidence intervals (represented by vertical lines). Prior knowledge refers to prior collaboration knowledge. The score was z-standardized.

**Figure F2**

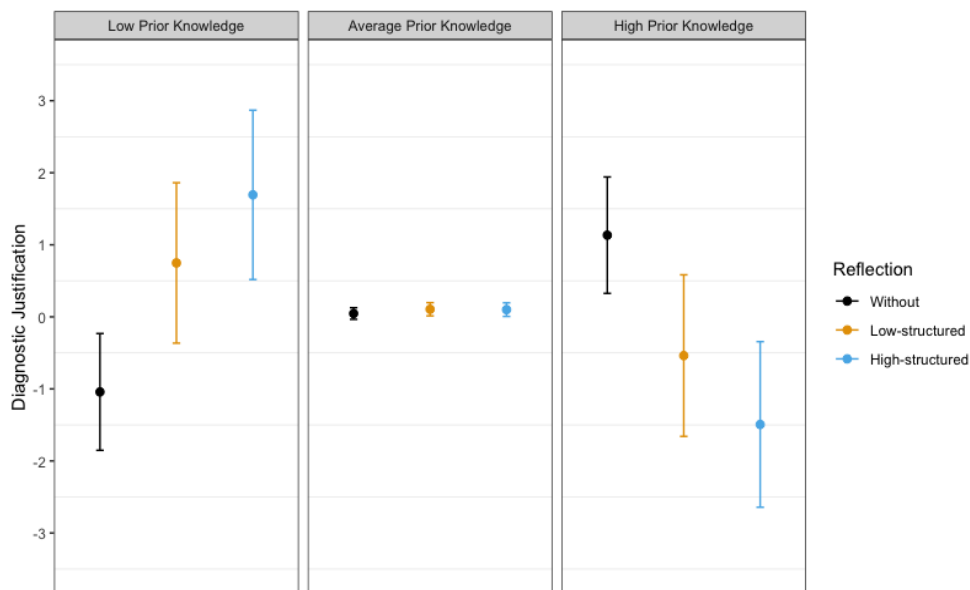
*Effects of Low- and High-Structured Guidance for Reflection on Collaborative Activities on Diagnostic Accuracy*



*Note.* Estimated means per group are shown as dots, accompanied by confidence intervals (represented by vertical lines). Prior knowledge refers to prior collaboration knowledge. The original score (probabilities of making an accurate diagnosis) was z-standardized using the inverse of the cumulative distribution function (CDF) of the standard normal distribution, allowing for better comparison with performance in other subskills.

**Figure F3**

*Effects of Low- and High-Structured Guidance for Reflection on Collaborative Activities on the Quality of Diagnostic Justification*



*Note.* Estimated means per group are shown as dots, accompanied by confidence intervals (represented by vertical lines). Prior knowledge refers to prior collaboration knowledge. Diagnostic justification was z-standardized.

**Eidesstattliche Versicherung**  
Statement of Scientific Integrity

Richters, Constanze Catharina

---

Name, Vorname

*Last name, first name*

Ich versichere, dass ich die an der Fakultät für Psychologie und Pädagogik der Ludwig-Maximilians-Universität München zur Dissertation eingereichte Arbeit mit dem Titel:

*I assert that the thesis I submitted to the Faculty of Psychology and Pedagogy of the Ludwig-Maximilians-Universität München under the title:*

**Learning Collaborative Reasoning:  
Foundations of Adaptive Reflection Support in Agent-based Simulations**

selbst verfasst, alle Teile eigenständig formuliert und keine fremden Textteile übernommen habe, die nicht als solche gekennzeichnet sind. Kein Abschnitt der Doktorarbeit wurde von einer anderen Person formuliert, und bei der Abfassung wurden keine anderen als die in der Abhandlung aufgeführten Hilfsmittel benutzt.

*is written by myself, I have formulated all parts independently and I have not taken any texts components of others without indicating them. No formulation has been made by someone else and I have not used any sources other than indicated in the thesis.*

Ich erkläre, das ich habe an keiner anderen Stelle einen Antrag auf Zulassung zur Promotion gestellt oder bereits einen Dokortitel auf der Grundlage des vorgelegten Studienabschlusses erworben und mich auch nicht einer Doktorprüfung erfolglos unterzogen.

*I assert I have not applied anywhere else for a doctoral degree nor have I obtained a doctor title on the basis of my present studies or failed a doctoral examination.*

München, 01. 04. 2024

---

Ort, Datum

*Place, Date*

Constanze Richters

---

Unterschrift Doktorandin/Doktorand

*Signature of the doctoral candidate*