Leveraging Process Data for Assessing and Supporting Collaborative Problem-Solving

Collaborative Diagnostic Reasoning in Agent-Based Simulations

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
an der Ludwig-Maximilians-Universität München
vorgelegt von

Laura Brandl
aus Starnberg

2025

Referent: Prof. Dr. Frank Fischer
Korreferent: Prof. Dr. Matthias Stadler
Tag der mündlichen Prüfung: 12.02.2025

Acknowledgements

There is quite a number of people that deserve to be acknowledged for their contributions and support to the completion of this thesis. First of all, I would like to express my sincere gratitude to Frank Fischer for supervising my thesis and guiding my research over the past years. You have given me the space to develop my research and personal skills, but have always been there with just the right amount of guidance. Thank you also for always inspiring me to think further and for taking my discussions to the next level. Matthias Stadler, thank you for being a great mentor and always having an open ear for my concerns and needs. I am impressed how you always find a quick solution to almost any problem I have and how great you are at finding great places to eat during conferences. I am also grateful for the opportunity to work with Ralf Schmidmaier and Martin Fischer. Thank you for being expert advisors and sharing your expertise in both medicine and medical education, and thank you, Ralf, for being my third examiner.

Furthermore, I would like to express my deepest gratitude to the DFG research group COSIMA, where I started my scientific journey in 2018 as a research assistant and was able to continue my academic work until now. In this context, I would like to thank Anika Radkowitsch, you were one of the first to suggest that I should do a master's degree and continue working in academia, thank you for your support from the very beginning. Next, it was a pleasure to work with you, Constanze Richters, and to continue our work on the TP6 together for most of the time. Special thanks also to Olga Chernikova and Nicole Heitzmann for their great support in the whole research group.

A big thank you also goes to the great team at the Chair of Educational Psychology at the LMU, where I have had the pleasure of working for the past three years. First and foremost, thanks to Simone Steiger and Alexander Kacina, who were always there to help with any bureaucratic or technical issues. But also, to all the others who made the lunch breaks so joyful and fruitful. I'm deliberately not mentioning anyone in particular here, so as not to miss anyone! But a special thanks goes to Michael Sailer, even though you are no longer at the chair - thank you for your support since my undergraduate thesis, but also for the many funny and insightful conversations we have had since then.

Next, I would like to thank all my fellow PhD students from the DTP - especially Katharina Bach, Franziska Eckerskorn, Lixiang Gao, Mevsim Küçükakyüz, Melissa Özsoy, Marina Pfeifer, Meral Roeben, David Sachs, Johanna Vejvoda, and Miriam Wünsch - for the great retreats, discussions and support over the past three years. Special thanks to Meral for being the queen of sending the most appropriate memes whenever a little smile was needed. Also, a big

thank you to Dragan Gašević and the whole CoLAM team for making my stay so enlightening and enjoyable.

Finally, I would like to turn to important support outside the academic world and thank all my friends for their support, which has been characterized not only by academic but also by personal challenges. No matter what happened, I could always count on you. Special thanks to Sina, thank you for reminding me of what is important in life when I lost my focus and for always having a spare minute when I needed it. I would also like to thank my parents for supporting me all my life and always encouraging me that only the sky is the limit and Werner for always being there when I needed you. Last but not least, I wouldn't be where I am without you, Christoph. I know you lost your patience with me sometimes, but thank you for never losing faith in me. Thank you for being by my side all these years, supporting me emotionally and traveling the world with me.

Extended Summary

Collaborative problem-solving has emerged as a critical skill in the 21st century, as it is essential for addressing complex and multifaceted challenges inherent in modern work environments (Graesser et al., 2018). In this thesis collaborative problem-solving skills are defined as the capacity of an individual to effectively engage in a process, where two or more agents with different knowledge bases attempt to solve complex tasks. The process of collaborative problem-solving involves active interaction with the problem, decision-making under uncertainty, and the integration of knowledge and skills to create a shared problem representation needed to reach a solution. Medicine is domain where it is of critical importance to reduce diagnostic errors and thus ensure high quality patient care. Collaborative diagnostic reasoning, a form of collaborative problem-solving in knowledge-rich domains like medical diagnosing, describes the critical role of collaboration when solving diagnostic problems in order to achieve accurate, well-reasoned and efficient diagnoses. Building upon research on collaborative problem-solving and diagnostic reasoning the collaborative diagnostic reasoning model (CDR-M; Radkowitsch et al., 2022) proposes a joint perspective in solving diagnostic problems in a collaborative effort. While this thesis focuses primarily on medical contexts, the insights and methods developed are expected to be applicable across disciplines.

To support the development of expertise in collaborative problem-solving and collaborative diagnostic reasoning, it is important to provide authentic situations allowing for knowledge application and schema acquisition. Through repeated exposure to diagnostic problems and experience with cases, knowledge gets encapsulated and a data-base of already seen cases is created or updated. This leads to prototypical abstract case representations enabling greater accuracy and efficiency when solving diagnostic problems (Boshuizen et al., 2020). The educational implications are straightforward: For the restructuring and reorganization of biomedical knowledge, the early exposure to patient cases is considered essential. However, the opportunity to engage in real-life problem-solving is limited and relevant situation to learn may arise less often or are too critical to be approached by novices.

One way to overcome this issue and also facilitate the assessment of collaborative problem-solving skills, is the use of technology-based assessments and simulation-based learning environments. Simulation-based learning environments offer authentic situations for learners to practice collaborative diagnostic reasoning without the risks associated with real patient cases (Chernikova et al., 2020). The use of computerized agents as collaboration partners allows to create a standardized and controlled setting that is hard to establish in collaborations among humans. However, although the use of simulation-based learning environments and the

integration of technology-based assessments presents opportunities it also entails challenges in assessing and supporting collaborative problem-solving skills. The development of technology-based interactive tasks and simulation-based learning using computerized tasks enables a closer approximation of real-world scenarios. These tasks allow for monitoring the process through observable problem-solving behaviors, which are stored as computer-generated log-file data and can be accessed to provide valuable additional insights. Hence, process data can not only be used to examine what has been achieved, but also how it was achieved, and to make inferences about the cognitive processes involved in collaborative problem-solving. These inferences are implications for assessing performance differences, developing predictive models, and providing personalized support (Ulitzsch et al., 2023). However, there are also a number of challenges associated with its use: Starting with ethical considerations before and during data collection, through to the complexities of analyzing the data and the need for theory in interpreting the results.

The goal of this thesis is to improve the use of process data for assessing and supporting collaborative problem-solving, specifically in the context of collaborative diagnostic reasoning in medical education. To do so, this thesis compromises three papers having different foci on the usage of process data. The first paper takes a meta-perspective and elaborates recent developments in leveraging process data through technology-based assessments for creating new knowledge, improving learning and instruction, and providing actionable advice to policy stakeholders. Building on these considerations, two empirical studies illustrate how process data can be used for theoretical advancements and to improve instruction. The second paper and first empirical study validates the CDR-M using process data. The third paper and second empirical study then demonstrates how the combination of process data and theory can be used to predict outcomes that can inform instruction in simulation-based learning of collaborative diagnostic reasoning.

The first paper, a theoretical paper, analyzes the impact of process data from interactive tasks in large-scale assessments. The paper highlights necessary changes that need to be undertaken at the scientific level in how we analyze process data to foster sustainable changes at the practical and policy levels. Firstly, linking process data to educational theory is crucial for enhancing the generalizability of our findings and hence facilitate theoretical advancements. Secondly, the design of assessment should be aligned with instructional design to inform learning and instruction.

Paper 2 employs process data to empirically test and refine the CDR-M and thus demonstrates how process data can be harnessed to generate new insights and advance theoretical

frameworks in education. By analyzing data from three studies in a simulation-based environment the aim of the study was to better understand the collaborative diagnostic reasoning and the processes involved using a structural equation model including indirect effects. Results identified various stable relations between individual characteristics and collaborative diagnostic activities, and between collaborative diagnostic activities and diagnostic outcome, highlighting the multidimensional nature of collaborative diagnostic reasoning. In summary, the second paper found that for successful collaborative problem-solving in knowledge-rich domains, knowledge about the domain of the collaboration partner and collaborative diagnostic activities play a crucial role in addition to content knowledge, which is traditionally in the focus of expertise research.

The third paper focuses on enhancing simulation-based learning by predicting diagnostic accuracy in collaborative diagnostic reasoning using process data. This study developed a random forest classification model based on theoretically derived process indicators to predict success in a simulated learning environment. Results showed a satisfactory prediction rate for collaborative diagnostic reasoning performance, indicated by diagnostic accuracy. The model predicted accurate and inaccurate diagnoses and was therefore suitable for making statements about the performance by only using process data of collaborative diagnostic reasoning. Hence, Paper 3 showed that using prediction models enables researchers to provide practical solutions such as identifying learners at risk to show inadequate performance in need of adaptive instructional support.

In a nutshell, in terms of theoretical advancements, the papers presented indicate support for four assumptions proposed in the CDR-M, as well as adding two new assumptions to the CDR-M. Firstly, unique contribution of collaborative diagnostic activities to collaborative diagnostic reasoning and secondly, the need to investigate complex non-linear interactions between collaborative diagnostic activities. With respect to supporting the development of collaborative diagnostic reasoning skills, practical implications are to focus on collaboration knowledge and collaborative diagnostic activities and turn the measurement of processes like collaborative diagnostic activities into a design factor. In addition, a strategy for providing adaptive instructional support is proposed. Lastly, the findings in this thesis also reveal several insights into how the usage of process data analyses can be enhanced when assessing and supporting collaborative problem-solving skills. Most importantly, by leveraging theory-based frameworks to describe collaborative problem-solving processes, we can create a common ground for assessing and enhancing collaborative problem-solving skills across different domains and thus further improve the use of process data analyses.

Overall, findings of the three papers illustrate how process data can be used to advance theoretical models, as shown by the CDR-M, to support learning of collaborative diagnostic reasoning skills and, thus, ultimately enhance the usage of process data of collaborative problem solving. In conclusion, this thesis highlights the need of leveraging theory-based frameworks to describe collaborative problem-solving processes. This will lead to more proficient collaborators in the future, not only in the medical domain.

Zusammenfassung

Kollaboratives Problemlösen hat sich im 21. Jahrhundert als eine entscheidende Fähigkeit für die Bewältigung von komplexen und vielschichtigen Herausforderungen in modernen Arbeitsumgebungen herausgestellt (Graesser et al., 2018). In der vorliegenden Arbeit werden kollaborative Problemlösefähigkeit als die Fähigkeit einer Person definiert, sich effektiv an einem Prozess zu beteiligen, bei dem zwei oder mehr Agenten mit unterschiedlichen Wissensständen versuchen, komplexe Aufgaben zu lösen. Der Prozess des kollaborativen Problemlösens beinhaltet die aktive Interaktion mit dem Problem, Entscheidungsfindung unter Unsicherheit und die Integration von Wissen und Fähigkeiten, um eine geteilte Problemrepräsentation zu schaffen, die wiederum für die Lösung erforderlich ist.

Ein Bereich, in dem es von entscheidender Bedeutung ist, diagnostische Fehler zu reduzieren und somit eine hochwertige Patientenversorgung zu gewährleisten, ist die Medizin. Kollaboratives diagnostisches Denken ist eine Form des kollaborativen Problemlösens im Kontext von Aufgaben, die einen hohen Wissensstand erfordern, wie es der Fall bei der medizinischen Diagnose ist. Es beschreibt die kritische Rolle der Zusammenarbeit beim Lösen diagnostischer Probleme, um genaue, gut begründete und effiziente Diagnosen zu erreichen. Aufbauend auf der Forschung zu kollaborativem Problemlösen und diagnostischen Denken, schlägt das Modell zum kollaborativen diagnostischen Denken (CDR-M; Radkowitsch et al., 2022) eine gemeinsame Perspektive beim Lösen diagnostischer Probleme in kollaborativer Zusammenarbeit vor. Obwohl sich diese Arbeit primär auf medizinische Kontexte konzentriert, wird davon ausgegangen, dass die gewonnenen Erkenntnisse und Methoden disziplinübergreifend gültig sind.

Um die Entwicklung von Expertise in kollaborativem Problemlösen und kollaborativem diagnostischen Denken zu unterstützen, ist es wichtig, authentische Situationen bereitzustellen, die Wissensanwendung und den Schemata-Erwerb ermöglichen. Durch wiederholte Auseinandersetzung mit diagnostischen Problemen und Erfahrung mit Fällen wird Wissen verkapselt und eine Datenbank bereits gesehener Fälle erstellt oder aktualisiert. Dies führt zu prototypischen abstrakten Fallrepräsentationen, die eine größere Genauigkeit und Effizienz beim Lösen diagnostischer Probleme ermöglichen (Boshuizen et al., 2020). Die pädagogischen Implikationen, die man daraus ziehen kann, sind eindeutig: Für die Restrukturierung und Reorganisation von biomedizinischem Wissen ist es essentiell früh mit Patientenfällen konfrontiert zu sein. Allerdings sind Gelegenheiten, bei denen man, sich an realen Patientenfällen beteiligen kann, begrenzt und relevante Lernsituationen sind oft zu kritisch, dass es unverantwortlich wäre, Anfänger damit zu betrauen.

Eine Möglichkeit, diesem Problem zu begegnen und zudem die Bewertung von kollaborativen Problemlösefähigkeiten zu erleichtern, bietet die Nutzung von technologiegestützten Assessments und simulationsbasierten Lernumgebungen. Simulationsbasierte Lernumgebungen bieten Lernenden authentische Situationen, um kollaboratives diagnostisches Denken zu üben, ohne die mit realen Patientenfällen verbundenen Risiken zu fürchten (Chernikova et al., 2020). Der Einsatz von computergestützten Agenten als Kollaborationspartner:innen ermöglicht es, ein standardisiertes und kontrolliertes Setting zu schaffen, das in der menschlichen Zusammenarbeit schwer umzusetzen ist. Allerdings bringt die Nutzung von simulationsbasierten Lernumgebungen und die Integration von technologiegestützten Assessments nicht nur Chancen mit sich, sondern auch Herausforderungen bei der Bewertung und Unterstützung von kollaborativen Problemlösefähigkeiten.

Die Entwicklung von technologiegestützten interaktiven Aufgaben und simulationsbasierten Lernumgebungen unter Verwendung computergestützter Aufgaben ermöglicht eine zunehmende Annäherung an reale Szenarien. Diese computergestützten Aufgaben ermöglichen die Beobachtung des Problemlöseprozesses, repräsentiert durch beobachtbares Problemlöseverhalten. Diese Daten werden als computergenerierte Logfiles gespeichert und können so zusätzliche wertvolle Einblicke liefern. Prozessdaten können daher nicht nur verwendet werden, um zu untersuchen, *welches* Ergebnis erreicht wurde, sondern auch, *wie* dieses Ergebnis erreicht wurde. Dies erlaubt Rückschlüsse auf die kognitiven Prozesse, die beim kollaborativen Problemlösen ablaufen. Diese Rückschlüsse haben Implikationen für die Bewertung von Leistungsunterschieden, die Entwicklung von prädiktiven Modellen und die Bereitstellung personalisierter Unterstützung (Ulitzsch et al., 2023). Allerdings gibt es auch eine Reihe von Herausforderungen bei der Verwendung von Prozessdaten: Beginnend mit ethischen Überlegungen vor und während der Datenerhebung, bis hin zu den Komplexitäten bei der Analyse der Daten und der Notwendigkeit von Theorien bei der Interpretation der Ergebnisse.

Ziel dieser Arbeit ist es, die Nutzung von Prozessdaten zur Bewertung und Unterstützung des kollaborativen Problemlösens zu verbessern, insbesondere im Kontext des kollaborativen diagnostischen Denkens in der medizinischen Ausbildung. Dazu umfasst diese Arbeit drei Artikel mit unterschiedlichen Schwerpunkten auf der Nutzung von Prozessdaten. Der erste Artikel nimmt eine Meta-Perspektive ein und erläutert jüngste Entwicklungen bei der Nutzung von Prozessdaten durch technologiegestützte Assessments zur Schaffung neuen Wissens, zur Verbesserung von Lehren und Lernen und zur Bereitstellung umsetzbarer Ratschläge für politische Entscheidungsträger. Aufbauend auf diesen Überlegungen illustrieren zwei empirische Studien, wie Prozessdaten für theoretische Fortschritte und zu verbesserter Instruktion genutzt werden

können. Im zweiten Artikel wird eine empirische Studie zur Validierung des CDR-M vorgestellt. Der dritte Artikel und die dort berichtete zweite empirische Studie zeigen dann, wie die Kombination von Prozessdaten und Theorie genutzt werden kann, um Lernenden-Ergebnisse vorherzusagen, welche genutzt werden können um in simulationsbasierten Lernumgebungen des kollaborativen diagnostischen Denken instruktionale Anpassungen vorzunehmen.

Im ersten theoretischen Artikel wird die Nutzung von Prozessdaten aus interaktiven Aufgaben in large-scale Assessments analysiert. Der Artikel hebt hervor, welche Änderungen hinsichtlich der Art und Weise, wie Prozessdaten analysiert werden auf der wissenschaftlichen Ebene unternommen werden müssen, um nachhaltige Veränderungen auf praktischer und politischer Ebene zu fördern. Zum einen ist die Verknüpfung von Prozessdaten und Bildungstheorien entscheidend, um die Generalisierbarkeit unserer Ergebnisse zu verbessern und somit theoretische Fortschritte zu erleichtern. Zum anderen sollte die Gestaltung von Assessments mit der instruktionalen Gestaltung abgestimmt sein, um Lehren und Lernen zu verbessern.

Der zweite Artikel testet und verfeinert das CDR-M empirisch mithilfe von Prozessdaten und zeigt somit, wie Prozessdaten genutzt werden können, um neue Erkenntnisse zu generieren und theoretische Modelle weiter zu entwickeln. Ziel der Studie war es die Daten aus drei simulationsbasierten Studien zu analysieren um das kollaborative diagnostische Denken und die beteiligten Prozesse besser zu verstehen, indem ein Strukturgleichungsmodell mit indirekten Effekten verwendet wurde. Die Ergebnisse identifizierten verschiedene stabile Beziehungen zwischen individuellen Merkmalen und kollaborativen diagnostischen Aktivitäten sowie zwischen kollaborativen diagnostischen Aktivitäten und diagnostischen Ergebnissen, was die multidimensionale Natur des kollaborativen diagnostischen Denkens hervorhebt. Zusammenfassend zeigte der zweite Artikel, dass für erfolgreiches kollaboratives Problemlösen in wissensreichen Aufgaben neben dem Fachwissen, das traditionell im Fokus der Expertiseforschung steht, Wissen über den Bereich der Kollaborationspartner:innen und kollaborative diagnostische Aktivitäten eine entscheidende Rolle spielen.

Der dritte Artikel konzentriert sich auf die Verbesserung des simulationsbasierten Lernens durch die Vorhersage der diagnostischen Genauigkeit im kollaborativen diagnostischen Denken unter Verwendung von Prozessdaten. Diese Studie entwickelte ein Random-Forest-Klassifikationsmodell basierend auf theoretisch abgeleiteten Prozessindikatoren, um den Erfolg in einer simulierten Lernumgebung vorherzusagen. Die Ergebnisse zeigten, dass diagnostische Genauigkeit, als Indikator für Erfolg im kollaborativen diagnostischen Denken, zufriedenstellend mithilfe von Prozessdaten vorhergesagt werden kann. Das Modell sagte sowohl genaue als auch ungenaue Diagnosen vorher und war daher dafür geeignet, Aussagen über die Leistung

ausschließlich unter Verwendung von Prozessdaten des kollaborativen diagnostischen Denken zu treffen. Daher zeigte der dritte Artikel, dass die Verwendung von Vorhersagemodellen es ermöglicht, praktische Lösungen bereitzustellen, wie z. B. die Identifizierung von Lernenden, die wahrscheinlich unzureichende Leistungen zeigen werden und daher instruktionale Unterstützung benötigen.

Zusammenfassend lässt sich sagen, dass die vorgestellten Artikel in Bezug auf den theoretischen Fortschritt Hinweise auf die Gültigkeit von vier im CDR-M vorgeschlagenen Annahmen sowie von zwei neuen Annahmen für das CDR-M liefern. Als neue Annahmen sollte erstens der einzigartige Beitrag kollaborativer diagnostischer Aktivitäten zum kollaborativen diagnostischen Denken und zweitens die Notwendigkeit, komplexe nicht-lineare Interaktionen zwischen kollaborativen diagnostischen Aktivitäten zu untersuchen berücksichtig werden. Hinsichtlich der Unterstützung der Entwicklung von Fähigkeiten im kollaborativen diagnostischen Denken bestehen praktische Implikationen darin, sich auf Kooperationswissen und kollaborative diagnostische Aktivitäten zu konzentrieren und die Messung von Prozessen wie kollaborativen diagnostischen Aktivitäten in einen Gestaltungsfaktor zu verwandeln. Darüber hinaus wird eine Strategie zur Bereitstellung adaptiver instruktionaler Unterstützung vorgeschlagen. Schließlich geben die Ergebnisse dieser Arbeit auch Einblicke in wie die Nutzung von Prozessdatenanalysen bei der Bewertung und Unterstützung von kollaborativen Problemlösefähigkeiten verbessert werden kann. Am relevantesten ist jedoch, dass wir durch die Verwendung theoriegeleiteter Modelle zur Beschreibung von kollaborativen Problemlöseprozessen eine gemeinsame Sprache für die Bewertung und Verbesserung von kollaborativen Problemlösefähigkeiten in verschiedenen Bereichen schaffen und somit die Nutzung von Prozessdatenanalysen weiter verbessern können.

Insgesamt veranschaulichen die Ergebnisse der drei Artikel, wie Prozessdaten verwendet werden können, um theoretische Modelle, wie das CDR-M voranzutreiben und somit das Lernen von Fähigkeiten im kollaborativen diagnostischen Denken zu unterstützen und somit letztendlich die Nutzung von Prozessdaten des kollaborativen Problemlösens zu verbessern. Abschließend ist anzumerken, dass diese Arbeit die Notwendigkeit der Nutzung eines theoriegestützten Modells zur Beschreibung kollaborativer Problemlöseprozesse hervorhebt. Dies wird nicht nur im medizinischen Bereich in Zukunft zu einer besseren Zusammenarbeit führen.

Table of Contents

GENERAL INTRODUCTION

1

Laura Brandl

1.1 Aim and Structure of the Thesis

Diagnostic errors are estimated to be the third leading cause of death in the US (Makary & Daniel, 2016). Furthermore, a study from the Netherlands found that almost all reported cases of such serious adverse events were associated with at least one human factor, such as errors in coordination or communication between healthcare teams (Hooftman et al., 2024). It is therefore crucial to improve collaborative diagnostic reasoning skills in medical contexts to ensure high-quality patient care. The concept of collaborative diagnostic reasoning emphasizes the pivotal role of collaboration in the process of solving diagnostic problems and achieving accurate, well-reasoned and efficient diagnoses (Radkowitsch et al., 2022). This thesis primarily focuses on medical contexts. However, it can be expected that the insights and methods developed will be applicable across disciplines, given that collaborative diagnostic reasoning, or more broadly collaborative problem-solving, are critical skills in a variety of professional domains due to the increasing complexity of the problems that professionals are required to solve (Fiore et al., 2018). Consequently, collaborative problem-solving skills have been identified as a pivotal 21st-century skill, fundamental for navigating complex challenges and integral to numerous aspects of modern work, particularly in fields that necessitate the integration of diverse perspectives and expertise (Graesser et al., 2018).

The utilization of technology-based assessments and simulation-based learning environments presents a promising basis for assessing and supporting the development of collaborative diagnostic reasoning skills. This is because such environments offer the potential to collect detailed process data, which can provide insights into the underlying cognitive processes and the complexities of the collaborative problem-solving process, which are not depicted in outcome or self-report measures (OECD, 2010). The aim of this thesis is to investigate the potential of process data derived from interactive collaborative problem-solving tasks, particularly within the context of collaborative diagnostic reasoning in agent-based simulations, to enhance both assessment and support in a way that is both sustainable and meaningful. This thesis demonstrates how process data can be practically applied to gain deeper insights, develop more robust theories and thereby support learning and instruction.

The remainder of this thesis is structured in three main parts, the first of which is dedicated to the theoretical underpinning. First and foremost, collaborative problem-solving is defined, along with an explanation of how expertise is developed in this area and how simulations contribute to the assessment and support of these skills. The subsequent section will focus on collaborative problem-solving in medical contexts, defining collaborative diagnostic reasoning and outlining how expertise development in this domain differs from that of general problem-

solving expertise. This is followed by an introduction to agent-based simulations as a means of facilitating the acquisition of collaborative diagnostic reasoning skills. Subsequently, the concept of process data analyses is introduced, along with an overview of the benefits and challenges associated with its implementation for the assessment and support of collaborative problem-solving skills. The first part concludes with a description of the general aim of the thesis, as well as brief outlines of the included papers and their research questions.

The second part of the thesis presents three papers conducted to achieve the stated goals. The first paper is a theoretical paper that takes a meta-perspective on the sustainable utilization of process data in large-scale assessments. Although this paper focuses on large-scale assessments, it is assumed that the recommendations regarding the use of process data are also applicable to other contexts, such as simulation-based learning. The second paper presents the findings of an empirical study which aims to investigate the extent to which process data can facilitate the creation of new knowledge, particularly in the context of validating theoretical models in educational research. In order to validate the collaborative diagnostic reasoning model (CDR-M), a multi-study structural equation model is analyzed. The third paper investigates the potential of process data to inform learning and instruction by predicting learners' needs for additional support. In particular, it investigates whether process data can be used to identify learners who may benefit from adaptive instructional interventions during collaborative problem-solving tasks in medical education. The third and final part of the thesis presents a synthesis of the findings in light of the initial theoretical assumptions. In conclusion, the thesis discusses the implications for research and practice, with focusing on leveraging process data for the assessment and support of collaborative problem-solving in the context of collaborative diagnostic reasoning within agent-based simulations.

1.2 Collaborative Problem-Solving

The ability to collaborate with others is a central skill in the 21st century, spanning a range of contexts, including computer-supported collaborative learning and collaborative problem-solving in professional practice (Fiore et al., 2018; Griffin & Care, 2015; OECD, 2017a). The focus of this thesis will be on the topic of collaborative problem-solving. This is due to the fact that many of the key problems faced by modern societies are of a highly complex nature, and therefore require the input and collaboration of multiple individuals rather than single individual in order to be solved (Graesser et al., 2022). After defining the construct, the chapter elaborates on developing, assessing, and supporting expertise in these skills.

1.2.1 Collaborative Problem-Solving Skills

The integration of multiple perspectives and sources of knowledge and expertise through collaboration has been demonstrated to enhance the quality of solutions (Graesser et al., 2018). While collaboration offers certain advantages, such as the sharing of knowledge, the combination of specialist skills, and the distribution of work; it also presents challenges in the form of miscommunication, coordination issues, and potential conflicts in goal alignment (Funke et al., 2018). When problems are solved collaboratively, the cognitive activities that are required for individual problem-solving are extended by collaborative activities[1] that are needed to achieve the desired outcome. The construct of collaborative problem-solving comprises several components, which highlight different aspects related to the collaborative problem-solving process. Accordingly, the relevant components and their conceptualization within the field of collaborative problem-solving will be outlined, followed by a definition that incorporates a synthesis of these elements.

Irrespective of whether a solution is reached individually or collaboratively, a problem is encountered when the desired goal state differs from the actual current state and there is no routine method of solution available (Mayer & Wittrock, 2006). The early research on problem-solving concentrated on relatively simple, *knowledge-lean*, tasks that did not require a lot of knowledge, such as the Tower of Hanoi or other puzzle-like tasks. These tasks are distinguished by the provision of all necessary information within the task instructions, thereby necessitating minimal prior knowledge and relying primarily on general cognitive skills and reasoning abilities. In knowledge-lean tasks, the problem space is typically well-defined, with clear initial states, operators, and goal states provided within the task instructions (van Lehn, 1989). The underlying assumption was that the cognitive processes used to solve these knowledge-lean problems were generalizable to more complex problems, suggesting that problem-solving skills were domain-general (Newell et al., 1959). Consequently, these problems are useful for the evaluation of general cognitive abilities, as they are not dependent on specific content knowledge.

In contrast, *knowledge-rich tasks* that require a high level of domain-specific knowledge are relevant whenever it comes to learning. Examples of domains that are particularly knowledge-rich include engineering, physics, medical diagnosing, and other specialized fields. In these domains, good problem solvers possess content knowledge that is well-organized, coherent, and chunked. This organization of knowledge enables the efficient representation of problems and the selection of appropriate strategies for solving knowledge-rich problems (Sugrue, 1995).

[1] Also referred to as social activities

The presence or absence of solution-relevant knowledge is a critical factor in determining whether a situation is perceived as a problem. To provide an example, a simple arithmetic question may be easily solved by the majority of adults, yet it could be unsolvable for a preschooler due to the lack of relevant knowledge. This illustrates that the problem-solving process is inherently linked to the knowledge base of the individual attempting to solve the problem, and that the perception of a problem can vary considerably due to the prior knowledge and experience of the individual (Funke et al., 2018). Consequently, there has been a shift in focus towards studying problem-solving within specific domains (Mayer & Wittrock, 1996). The problem-solving process that occurs in the context of knowledge-rich tasks can be described as follows: Firstly, a mental representation of the problem is established; secondly, relevant schemas or scripts (see 1.2.2) are activated; and thirdly, this knowledge is applied in order to derive a solution (Greiff et al., 2016).

There are several different categories of problem-solving tasks, such as knowledge-lean, well-defined, complex, interactive, ill-defined, open-ended, knowledge-rich and much more with no obvious boundary between the different labels or the constructs they represent (Funke et al., 2018). This thesis will focus on complex problems, which are defined as dynamic systems that individuals must manage in conditions of uncertainty (Dörner, 1975), in knowledge-rich tasks. These problems typically comprise a number of interconnected elements that are capable of changing autonomously over time. Complex problems frequently lack transparency, necessitating the retrieval and management of information. Furthermore, complex problems may entail polytelic goals, which are competing or conflicting objectives that must be balanced (Dörner, 1975; Funke et al., 2018). The complexity of a problem is frequently attributed to the structure of the external problem representation. The perceived complexity of a problem is subject to variation, depending on the level of expertise of the individuals engaged in problem-solving. For example, a problem may be perceived as less complex by experts than by novices. This distinction is crucial, as it acknowledges that the complexity and difficulty of a problem are not inherent properties but are also dependent on the expertise of the problem solver (A. Fischer et al., 2011).

In order to solve a problem, one or more individuals must engage in a problem-solving process. This process involves searching for an operation or a series of operations with the aim of transferring the given actual state of the system to a goal state (Dunbar, 1998; Newell & Simon, 1972). It requires a goal-oriented sequence of cognitive activities (Anderson, 1993; Funke et al., 2018). Individual problem-solving skills are defined as “an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of

solution is not immediately obvious" (OECD, 2013, p. 122). Problem-solving requires the application of problem-solving strategies, which can be classified as either domain-general or domain-specific. Domain-general strategies are employed in knowledge-lean task or tasks where domain-specific knowledge is lacking, whereas domain-specific strategies are employed in knowledge-rich tasks. Additionally, an accurate problem representation is essential, either through solely interacting with the problem or through the activation of domain-specific knowledge. Finally, self-regulation is necessary to monitor the execution of the problem-solving process (O'Neil, 1999).

The problem-solving process can be divided into two distinct phases: knowledge acquisition and knowledge application. The process of knowledge acquisition entails the creation of a mental representation of the problem, which encompasses an understanding of the problem's structure and the relevant information necessary for its solution (Klahr & Dunbar, 1988). This phase is of critical importance for the establishment of a clear and accurate problem representation, which subsequently serves as a foundation for decision-making and strategy development. The second phase, knowledge application, involves the implementation of the solution process based on the established problem representation (Novick & Bassok, 2005). The application of knowledge entails the selection and execution of appropriate actions that facilitate the transition from the current state to the desired goal state, based on the problem representation. This process requires not only the retrieval of relevant knowledge but also the application of strategic thinking and problem-solving heuristics. A study by Nicolay et al. (2021) investigated both phases of individual problem-solving in 1151 students in 9th grade working on nine problems. The findings showed that despite the acquisition of all relevant information about the problem during the knowledge acquisition phase, two in five students were unable to fully solve the problem in the subsequent knowledge application phase. This emphasizes the importance of both phases for successful problem-solving.

The Program for International Student Assessment (PISA) 2012 framework identifies four distinct cognitive processes that constitute the individual problem-solving process (OECD, 2013): (1) exploration and understanding of the problem task; (2) creation of a problem representation through integration of acquired information with relevant prior knowledge, leading to specific hypotheses about potential solutions. To reduce the uncertainty of these hypotheses, (3) a plan is created and executed, as well as (4) monitored and reflected on, in order to reach the solution. This is consistent with the hypothesis stated by Klahr and Dunbar (1988), which posits that problem-solving in the domain of scientific discovery involves searching through both a hypothesis generation space and a hypothesis testing space. To put it differently, the

understanding process or knowledge acquisition phase generates the person's internal representation of the problem, whereas the search process or knowledge application phase generates the person's solution (van Lehn, 1989).

In the context of collaborative problem solving, in addition to individual cognitive activities, collaborative activities (e.g. exchanging ideas, negotiating ideas, regulating problem solving and maintaining communication) are crucial to the collaborative problem-solving process. The main goal of collaborative activities is to construct a shared problem representation (Rochelle & Teasley, 1995). Research indicates that collaborative problem-solving performance is enhanced when the initial problem representation of each individual is consistent across collaboration partners (Hesse et al., 2015). A study by Mathieu et al. (2000) involving 26 student dyads working on a flight simulation found that concurrent problem representations between collaboration partners improved the quality of collaborative problem solving, leading to positive outcomes. The construction of a shared problem representation requires the conscious and continuous monitoring and coordination of individual cognitive activities and collaborative activities related to shared knowledge (Hesse et al., 2015; Liu et al., 2016; Rochelle & Teasley, 1995). A variety of models have been developed to describe the processes and required skills by which humans collaborate to solve problems. These models differ primarily in terms of their granularity (see Table 1).

The model proposed by Liu et al. (2016) identifies four key social skills, whereas Hesse et al. (2015) suggest three main skills with several sub-skills. In particular, the ability to recognize the information required by a collaborator to construct a shared problem representation is highlighted, as well as the identification of the specific information that needs to be shared (Rochelle & Teasley, 1995). The OECD (2017a) based their theoretical framework on the work of Hesse et al. (2015), but expressed it in the form of a 4x3 matrix, comprising four cognitive activities and three collaborative activities. Lastly, Sun et al. (2020) synthesize recent models of collaborative problem-solving into three overarching categories: constructing shared knowledge, negotiation/coordination, and maintaining team function. Each category is further divided into two subcategories and associated indicators. To illustrate, an indicator of the sub-facet *establishing common ground* for the facet *constructing shared knowledge* is defined as verifying the understanding of others' ideas through questioning or paraphrasing. Notwithstanding the discrepancies in granularity, these models exhibit considerable overlap: The majority of these models align on several fundamental collaborative activities, although there are some discrepancies in the terminology used to describe them. These collaborative activities include the effective sharing of information with collaboration partners, the elicitation of information from

collaboration partners to expand knowledge, the negotiation of conflicting ideas, and the regulation of collaborative processes by setting goals and monitoring the process. The sharing and eliciting of information, which is often referred to as information pooling, is of particular importance for collaborative information processing (F. Fischer et al., 2002; Hinsz et al., 1997). This sharing and eliciting of information facilitate the construction of a shared problem representation and potential solutions, which is a crucial element of successful collaboration (Rochelle & Teasley, 1995). The ability to negotiate conflicting ideas is of great importance when disagreements arise among collaboration partners (Hesse et al., 2015). Effective negotiation helps prevent groups from dismissing opposing viewpoints or prematurely ending discussions (Patel et al., 2002). It is also important to note that regulation is a crucial element in aligning the goals and strategies of a group in order to achieve those goals (Järvelä & Hadwin, 2013). Although collaborative problem-solving initially assigns a higher value to collaborative activities than to the cognitive activities involved in individual problem-solving (OECD, 2017a), both are essential for success. However, research lacks evidence on which factors are particularly relevant (Graesser et al., 2018).

Table 1

Overview of Three Collaborative Problem-Solving Frameworks

	Liu et al., 2016	**Hesse et al., 2015**	**OECD, 2017a**
individual cognitive activities	conceptual understanding	planning	exploring & understanding
	inquiry skills in science: data collection, data analysis, prediction making, and evidence-based reasoning	executing & monitoring	representing & formulating
		flexibility	planning & executing
		learning	monitoring & reflecting
collaborative activities	sharing ideas	participation: action, interaction, task completion/ perseverance	establishing & maintaining shared understanding
	regulating problem-solving activities	perspective taking: adaptive responsiveness, audience awareness	taking appropriate action to solve the problem
	negotiating ideas	social regulation: negotiation, self-evaluation, transactive memory, responsibility initiative	establishing and maintaining team organization
	maintaining communication		

Summarizing the different aspects of collaborative problem-solving skills, a problem in the process of collaborative problem-solving is understood as a complex system that needs to be transformed into a goal state under conditions of uncertainty requiring knowledge. The knowledge required for these tasks can be either inherent to the task itself (i.e., knowledge-lean tasks) or necessitate the utilization of prior knowledge organized in scripts (i.e., knowledge-rich tasks). Both types of tasks necessitate interaction with the problem to be solved. Consequently, in thesis collaborative problem-solving skills is defined as the capacity of a single individual to engage effectively in a process involving two or more agents, each with a different knowledge base, in order to solve complex tasks (OECD, 2017a). Although being a collaborative process, the skill is individual and therefore can be assessed and developed at the individual level, rather than at the group level. It entails active engagement with the problem at hand, decision-making in the presence of uncertainty, and the integration of knowledge and skills to construct a shared problem representation that is essential for reaching a solution. Further, collaborative problem-solving skills are understood to be a formative construct with varying degrees of its components' generalizability. These components are domain-specific knowledge and domain-general cognitive activities (e.g., creating a problem representation, generating hypotheses and a solution plan, monitoring the process) and collaborative activities (e.g., sharing and eliciting information, negotiating hypotheses, and regulating the process).

1.2.2 Expertise in Collaborative Problem-Solving

Both domain-specific knowledge and domain-general strategies are critical for problem-solving. The balance between them may shift over time with increasing age (Schäfer et al., 2024; cf. Geary et al., 2017), with domain-specific knowledge becoming more relevant for expert performance. Expert performance is understood as “consistently superior performance on a specified set of representative tasks for a domain” (Ericsson & Lehmann, 1996, p. 277). This means, for example, that experts can solve a problem faster and more accurately than novices and that they have better metacognitive abilities (van Lehn, 1989). In addition, experts seem to be able to store and recall more information and to select relevant strategies based on their previous experience with similar problems through the activation of schemas (Chase & Simon, 1973). A schema is conceptualized as a cognitive structure that stores knowledge from experience in a concrete or abstract form (Sweller, 1988). Another difference between experts and novices is the categorization of problems: While novices tend to identify surface features, experts group problems according to their schemas (Sweller, 1988; van Lehn, 1989). One explanation for the differences in expert performance compared to novices is that they can store a greater number of items in their working memory due to their organization of knowledge. This

leads to a reduced cognitive load and thus more capacity to engage in relevant problem-solving activities (A. Fischer et al., 2011). Hence, experts use knowledge when they need to search for the next step in the problem-solving process, which may be cognitively overwhelming for novices (Funke & Frensch, 2007; Mayer, 1992; Sweller, 1988). Vicente and Wang (1998) found that there are at least 51 studies in at least 19 different domains demonstrating the superior memory performance of experts. In summary, expertise is the development of cognitive structures necessary for effective problem-solving. Building on this, cognitive load theory offers crucial foundations for understanding the information processing demands inherent in (collaborative) problem-solving.

Cognitive load theory describes cognitive structures that include a virtually limitless long-term memory and a limited working memory (Atkinson & Shiffrin, 1968). The long-term memory acts as a storage for accumulated knowledge, while the working memory deals with the processing of information, either before it is encoded in the long-term memory or when it is retrieved for usage. The contents of working memory correspond to our conscious thoughts, whereas the vast contents of long-term memory are typically beyond our immediate awareness. Cognitive load theory is primarily concerned with how this extensive knowledge can be effectively acquired, given that the capacity of working memory is limited in both duration and the amount of new information it can hold. However, these limitations do not apply to information that is already well established in long-term memory (Paas et al., 2010).

Building on these assumptions, cognitive load is broadly understood as the amount of mental effort required by a task including intrinsic, extraneous, and germane cognitive load (Sweller et al., 2011). Intrinsic cognitive load is determined by the inherent complexity of the task (structure and interactivity) as a result of the individual's prior knowledge. Thus, a lack of prior knowledge can lead to cognitive overload, and expertise (a) helps to reduce intrinsic load given a certain interactivity between the elements of the task and (b) is assumed to moderate the usefulness of certain strategies and the effect of problem characteristics (*expertise reversal effect*, see Kalyuga, 2007). In addition, germane cognitive load is the mental effort required to cope with intrinsic load, whereas extraneous cognitive load is caused by poor instructional design that complicates the learning process (Paas et al., 2004).

According to cognitive load theory, learning is described as the acquisition of cognitive schemata that enable the categorization of the problem, the selection of the correct strategies to apply and the regulation of problem-solving. The construction of such schemata is cognitively demanding. Consequently, the processing of the task itself will compete with the construction of cognitive schemata if the task is too demanding. In summary, cognitive load theory addresses

the challenges that individuals face when engaging in complex cognitive tasks, which often involve the simultaneous management of numerous interactive elements. Thus, an individual's ability to perform in a particular domain depends on the amount of relevant knowledge stored in their long-term memory. Given the importance of this knowledge, it is essential to consider schemata—the structured form in which information is organized and stored (Sweller et al., 2011).

Expert performance depends on the acquisition of specific schemas stored in long-term memory. Schema theory became increasingly important in the 1980s because it seemed that domain-specific knowledge, organized into schemas, distinguishes experts from novices in problem-solving performance (Sweller et al., 2011). A "schema is defined as a structure which allows problem solvers to recognize a problem state as belonging to a particular category of problem states that normally require particular moves", allowing individuals to chunk information effectively, making it easier to retrieve and apply relevant knowledge during problem-solving tasks (Sweller, 1988, p. 259). Put differently, a problem schema consists of information about the class of problems to which the schema applies and information about their solutions (van Lehn, 1989). The development of schemata is crucial for overcoming the limitations of working memory, thus reducing cognitive load and errors (Anderson, 1985). According to the ACT* theory (Anderson, 1983), the key factor for expert performance is the ability to encode declarative (factual knowledge) and procedural knowledge (cognitive skills), which is basically reflected in the amount of experience. The transition from declarative to procedural knowledge, also known as knowledge compilation, is essential for the development of expertise in knowledge-rich domains (Anderson, 1985). In the initial stages of problem-solving in a new domain, individuals rely on declarative knowledge, which consists of isolated facts and information without an understanding of their application. During knowledge compilation, this declarative knowledge is first transformed into procedural knowledge (proceduralization), which involves knowing how to perform specific tasks. This procedural knowledge is subsequently compiled into larger networks of procedural knowledge (composition) through a gradual and laborious process known as knowledge compilation (Anderson, 1985). Knowledge compilation enables the creation of problem schemata that guide the selection, adaptation, and execution of solution procedures (Van Lehn, 1989). This transformation allows individuals to apply knowledge more efficiently, leading to improved problem-solving skills and distinguishing experts from novices. In summary, the construction of schemas in knowledge-rich domains is a critical mechanism for reducing cognitive load and errors. This process highlights the

importance of extensive experience and structured knowledge in the development of expertise, enabling more efficient and accurate problem-solving.

As individuals repeatedly encounter similar problems, they store these experiences in long-term memory, transforming declarative knowledge into procedural knowledge through proceduralization. As a result, experts can handle complex problems with greater ease and accuracy than novices, who lack such schemas and are prone to cognitive overload. They can quickly recognize familiar problems, retrieve appropriate schemas, and adapt them to specific situations. In contrast, novices must search for solutions without the benefit of pre-existing schemas, leading to trial-and-error approaches or *weak methods* (Perkins & Salomon, 1989; van Lehn, 1989). There is a so-called "power-generality tradeoff": The more general the method (i.e., means-end analysis), the weaker the method (Perkins & Salomon, 1989). In contrast, expert problem-solving consists of three steps: Selecting a schema, adapting (instantiating) it to the problem, and executing its solution procedure (van Lehn, 1989). Once an initial schema is triggered, it guides the interpretation of the rest of the problem. However, when more than one schema is applicable to the given problem, even experts must search for the appropriate one to reduce uncertainty in decision making. By enabling the recognition and application of relevant problem schemas, experience allows individuals to solve problems more accurately and efficiently, underscoring the importance of extensive learning and practice in developing expertise.

This interplay between knowledge and strategy application highlights that domain-specific and domain-general problem-solving are not distinct categories but rather two ends of a continuum (Greiff et al., 2014). In knowledge-lean tasks or in the absence of structured domain knowledge, domain-general strategies play a critical role. Conversely, when tackling knowledge-rich tasks with well-structured domain knowledge, domain-specific strategies become essential. Thus, domain-general problem-solving strategies can be seen as a tool needed to solve problems, but it takes domain-specific knowledge gained through experience to learn when and how to apply these strategies, leading to domain-specific strategies (Perkins & Salomon, 1989). Someone who is a very skilled problem solver in one domain may not be able to transfer their problem-solving skills and strategies to another domain in which they lack expertise. Both domain-specific and domain-general strategies are developed through experience, leading to an increase in the quality of the problem-solving process as individuals develop the ability to recognize and apply relevant problem schemas with less conscious processing (Sweller et al., 2011).

In conclusion, by examining complex real-world problems, researchers have gained deeper insights into the cognitive processes underlying expertise, highlighting the importance of

domain-specific knowledge that is connected, integrated, coherent, and chunked through extensive experience in effective problem-solving (Anderson, 1993; Funke et al., 2018; Sugrue, 1995). Initially, problem solvers in knowledge-rich tasks experience high cognitive load due to limited working memory capacity, resulting in frequent errors. Over time, with increasing experience and exposure to domain-specific problems, individuals construct schema - organized knowledge structures - that increase working memory capacity and reduce cognitive load (Anderson, 1985). Because expert performance is a product of knowledge (van Lehn, 1989), problem-solving expertise is a domain-specific skill that, unlike general intelligence, can be learned and supported (Funke et al., 2018).

1.2.3 Simulations for Assessing & Supporting Collaborative Problem-Solving Skills

In order to support the learning of collaborative problem-solving skills, we need to be able to assess them. Educational assessment is a systematic method of collecting information or artifacts about a learner and learning processes in order to make inferences about the individual's skills (E. L. Baker et al., 2016).There are three main purposes: assessment to support learning (formative assessment), assessment of individual student performance (summative assessment), and assessment to evaluate programs (evaluative assessment; Pellegrino et al., 2001). Many traditional educational assessments use multiple-choice and constructed-response items (Lee et al., 2019). However, this is not suitable for assessing collaborative problem-solving skills as such items require responsiveness to the test taker's input. This is necessary because, despite being an individual skill, an assessment of collaborative problem-solving skills would hardly be valid if there was no interaction between the test-takers and the collaboration partners (Stadler, Herborn, et al., 2020). However, this leads to the limitation of measuring collaborative problem-solving as an individual skill, since the difficulty of the task lies not only in the nature of the problem but also in the collaboration partner, making standardized assessments that control for the effect of collaboration challenging (Herborn et al., 2020). Furthermore, it is important to mention that, according to the definition of the OECD (2017a), problem-solving skills focus on the attempt and not only on the outcome of the process.

To address these demands, technological advancements have enabled a shift from traditional paper-pencil assessments to technology-based assessments[2]. These approaches, including simulated and interactive tasks, provide a more dynamic and accurate approach of assessing collaborative problem-solving skills and other 21st-century competencies (Care et al., 2012). By reducing reliance on paper-pencil tasks, these innovations better capture the nuances of problem-solving processes, aligning with modern educational and assessment needs (OECD, 2010).

[2] These assessments are also referred to as technology-enhanced assessments or computer-based assessments.

Technology-based assessments make it possible to implement tasks that are responsive to test-takers' input to allow for appropriately complex and realistic tasks and provide new sources of evidence to assess test-takers' skills, such as their interactions with the virtual environment (Lee et al., 2019). By leveraging features like multimedia, simulations, interactive tasks, and virtual reality, these assessments offer innovative ways to evaluate skills more dynamically and comprehensively (Goldhammer et al., 2020). In addition to enabling the operationalization of previously unattainable skills, the use of technology-based assessments allows for the continuous measurement of the problem-solving process (i.e., process data), rather than just discrete states of problem-solving performance represented by answers to a task (i.e., product data; Thille et al., 2014). Thus, it is possible to measure underlying processes beyond the outcome of a task, which can be interpreted in terms of the cognitive and collaborative activities that occur during task completion, and to move from *if* a problem was solved to *how* it was solved (Goldhammer et al., 2013, Greiff et al., 2015; see 1.4.).

All current assessments of collaborative problem-solving skills are technology-based and can be described by primarily two approaches (Li et al., 2024; for a review of assessments of collaborative problem-solving skills see Chai et al., 2024): human-to-human collaboration (e.g. ATC21S) and human-to-agent collaboration (e.g. PISA 2015). While human-to-human collaboration tasks involve a more authentic representation of natural collaboration, they lack controllability, and the group composition could affect the validity of the individual assessment, as the weakest collaboration partner determines the capabilities in the collaborative problem-solving process (Herborn et al., 2020; OECD, 2017a; Swiecki et al., 2020). However, because collaborative problem-solving is understood as an individual skill, human-to-agent collaboration tasks ensure the independence of students' behavior during the assessment (Herborn et al., 2020). This comes, in turn, with the limitation of a priori limited collaboration options and the risk of test takers pretending to know what the desired response or outcome is, rather than what they would do under the natural conditions (Graesser et al., 2017; Herborn et al., 2020; Oliveri et al., 2017). Nevertheless, there are benefits to using computerized agents as collaboration partners, allowing the creation of a standardized and controlled environment that is difficult to achieve with human-to-human collaboration (Rosen, 2015). Computerized agents allow for greater control over the collaboration process without deviating significantly from human-to-human interaction (Graesser et al., 2018; Graesser et al., 2017; Herborn et al., 2020). In less controlled settings, it is difficult to ensure that a particular process is taking place during collaborative problem-solving. For example, in a human-to-human collaboration, it is possible that although we intend to measure a specific activity, it is not taking place. For example, Rosen

(2014) explored this with respect to the comparability of conflict opportunities in a human-to-human and a human-to-agent assessment of collaborative problem-solving. One-hundred-thirty-six 14-year-old students from the United States, Singapore, and Israel worked in a human-to-agent setting, while 43 participated in a human-to-human setting. Both conditions worked on the identical collaborative problem-solving task, and students knew whether their collaboration partner was a computer agent or a classmate. The results indicated that while collaboration in the human-to-agent setting strongly promotes opportunities for conflict situations (25.3%), these situations are rare in the human-to-human setting (6.1%). However, in order to measure high levels of collaboration (OECD, 2013), it is critical that students have opportunities to engage in conflict-related behaviors (e.g., negotiating conflicting ideas). When using agents in technology-based interactive tasks, it is possible to ensure that all necessary activities take place during collaborative problem-solving (Rosen, 2015).

A prominent example of the use of human-to-agent collaboration is PISA, arguably the most comprehensive educational assessment program in the world, which in 2015 moved to technology-based assessment and human-to-agent collaboration, using conversational agents as collaboration partners. This allowed for the development of a standardized assessment environment, as agents can generate their responses from the same pre-programmed set of responses for each test-taker to assess collaborative problem-solving skills (Davier et al., 2019; OECD, 2017a). For instance, one of the tasks was to collaborate with two agents while taking part in a competition to answer questions about the fictional country of Xandar, this task can be considered as a knowledge-lean task (see OECD (2017b) for a detailed task description). Additional analyses of 483 German students within the PISA population found that self-rated collaboration, teacher-rated collaboration, peer collaboration, and reasoning were moderately related to performance on the PISA 2015 collaborative problem-solving tasks, even after controlling for individual differences in reading achievement, making the human-to-agent collaboration approach a valid assessment task (Stadler, Herborn, et al., 2020).

Results from PISA 2015 showed that only 8% of students worldwide performed at the highest level of proficiency, while 29% of students performed at the lowest level in PISA 2015 collaborative problem-solving tasks (OECD, 2017b). That is, while only 8% were able to "balance the collaboration and problem-solving aspects of a presented task, identify efficient pathways to a solution, and take actions to solve the given problem" (OECD, 2017b, p. 74), nearly one third “tend to focus on their individual role within the group" (OECD, 2017b, p. 74) and required support from their collaboration partners to solve even simple problems. Furthermore, the results showed that collaborative problem-solving performance is positively related to

performance in the other domains assessed, but the relation is weaker than that observed between performance in these other domains (OECD, 2017b), despite the fact that collaborative problem-solving was assessed in knowledge-lean tasks (OECD, 2017a). Thus, it is crucial to support the development of collaborative problem-solving skills.

To develop expertise in complex skills such as collaborative problem-solving, it is important to provide authentic situations that allow for knowledge application and schema acquisition (Kolodner, 1992). However, opportunities to engage in real-world problem-solving are limited, and relevant learning situations may occur infrequently or be too critical for novices to approach (Chernikova et al., 2020; Mislevy et al., 2017). Furthermore, it is not just a matter of having a lot of experience with collaborative problem-solving, but also of engaging in deliberate practice to reduce the risk of being cognitively overwhelmed and having the cognitive resources available for schema construction (Corbalan et al., 2006). This means that (1) some aspect of the process is focused on a well-defined problem with (2) immediate feedback on performance and (3) the opportunity to gradually improve by repeatedly performing the same or similar activities of the process (Ericsson, 2004). One way to address both aspects, the accessibility of relevant authentic problem situations and the consideration of cognitive resources, is to use simulation-based tasks as approximations of practice (Grossman et al., 2009).

Simulations are "a model or representation of reality (object, system, or situation) with certain parameters that can be controlled or manipulated" (Chernikova et al., 2022, p. 5), with an emphasis on interacting with authentic objects (Cook et al., 2013). The use of technology-based simulations spans tasks and domains as diverse as pilot training in flight simulators (L. Wong et al., 2012), decision making in business simulations (Siewiorek & Gegenfurtner, 2010), or medical diagnosis using simulated patients (Cook et al., 2010). Simulation-based learning is thought to produce more transferable skills than traditional learning because task similarity is a critical prerequisite for transfer (Cannon-Bowers & Bowers, 2010; Mayer & Wittrock, 2006). In addition, a meta-analysis by Chernikova et al. (2020) pooled the results of 145 empirical studies and found that simulations are among the most effective means of facilitating the learning of complex skills across domains compared to no intervention. The effect size is still very large when simulation-based learning is compared with different types of instruction. In addition, simulation-based learning can be particularly effective when additional adaptive instructional support is provided (Leutner, 1993). Adaptivity of instructional support is understood as the provision of support that is tailored to the specific needs of individuals (Plass & Pawar, 2020). Much research has been conducted in the area of simulation-based learning in medical education (Cook, 2014; Cook et al., 2013; Hegland et al., 2017).

1.3 Collaborative Problem-Solving in Medicine: Collaborative Diagnostic Reasoning

Diagnostic reasoning, i.e. accurately diagnosing a patient's illness is one of the physician's most important tasks, and often requires collaboration between physicians from different specialties. While problem-solving is generally understood as transferring the current state of a system to a goal state (Newell & Simon, 1972; see 1.2.1), diagnostic reasoning refers to identifying the causes of the current, mostly undesired, state not only in medical diagnosing but also engineering and teacher education (Abele, 2018). Despite this difference in the goal of the process, diagnostic reasoning is considered a form of problem-solving (Heitzmann et al., 2019).

The medical literature (e.g., Bowen, 2006; Patel et al., 2002) describes the ideal diagnostic process as consisting of three steps: After data collection, in which elements of the patient's history, physical examination, and other information are gathered, an initial representation of the problem is created and compared to an illness script (see 1.3.2), which is tested and leads to the exclusion of alternative hypotheses (Charlin et al., 2012; Tschan et al., 2009). Therefore, it is crucial that physicians have sufficient prior medical knowledge to use effective reasoning strategies to solve diagnostic problems (Cutrer et al., 2013). A central goal of diagnostic reasoning is to reach an accurate diagnosis, referred to as *diagnostic accuracy* (Chernikova et al., 2022; Simmons, 2010). In addition to achieving an accurate diagnosis, it is critical to adequately justify that diagnosis with evidence (e.g., key clinical findings), referred to as *diagnostic justification* (Daniel et al., 2019; Yudkowsky et al., 2015). Diagnostic justification makes the reasoning behind the decision transparent and understandable to others (Bauer et al., 2022). *Diagnostic efficiency* is related to the time and effort required to reach the accurate diagnosis, given that diagnosticians in practice are usually under time pressure (Braun et al., 2017).

Like the collaborative problem-solving process, the collaborative diagnostic reasoning process requires interaction with an agent (human or computerized) to find a solution to the diagnostic problem. This chapter first introduces the necessary components for effectively performing such processes, based on the collaborative diagnostic reasoning model (CDR-M). It then elaborates on how expertise in this domain is achieved, drawing on considerations of expertise development in collaborative problem-solving. Finally, the chapter focuses on how agent-based simulations support the development of collaborative diagnostic reasoning skills.

1.3.1 Collaborative Diagnostic Reasoning Skills

Diagnostic reasoning, whether performed individually or collaboratively, is the "goal-oriented collection and interpretation of case-specific or problem-specific information to reduce uncertainty in order to make […] [professional] decisions" (Heitzmann et al., 2019, p. 4). However, most medical problems are too complex (i.e., increased intrinsic load; see 1.2.2) to be

solved individually and require the interaction of multiple disciplines (Kiesewetter et al., 2017; Patel et al., 2002). Therefore, diagnosticians need to engage in collaborative diagnostic reasoning, which is defined as solving a problem, such as diagnosing a patient, "by generating and evaluating evidences and hypotheses that can be shared with, elicited from, or negotiated among" collaboration partners based on their conceptual and strategic knowledge (Radkowitsch et al., 2020, p. 2). This makes it a context-dependent and domain-specific skill consisting of individual and collaborative activities (Simmons, 2010).

Starting with the individual activities, the scientific discovery as dual search model (SDDS; Klahr & Dunbar, 1988) describes individual reasoning as a coordinated search through hypothetical evidence and hypotheses spaces. The SDDS assumes that successful reasoning depends not only on performing high-quality cognitive activities within these spaces, but also on being able to coordinate between them by using a hierarchy of cognitive activities. These activities include specifying hypotheses, deriving predictions from hypotheses, and testing and evaluating hypotheses in the light of existing evidence (Klahr & Dunbar, 1988). On a more abstract level, reasoning processes have been further described by so-called dual-process theories (Croskerry, 2009), in which reasoning can occur through a fast, unconscious retrieval process (System 1) or a more analytical, slow, deliberate, and conscious logical process (System 2). In diagnostic reasoning, this means that if the problem representation is familiar and matches already known problems, System 1 processes will quickly and effortlessly lead to the diagnosis and nothing further may be required; if this is not the case, effortful System 2 processes will take place (Croskerry, 2009). Although such models may provide some insights into how easily and accurately diagnosticians make a diagnosis (see 1.3.3), they are less useful for explaining the processes of diagnostic reasoning. Thus, a non-hierarchical conceptualization of eight epistemic activities, including (a) identifying a problem, (b) asking questions, (c) generating hypotheses, (d) constructing artifacts, (e) generalizing evidence, (f) evaluating evidence, (g) drawing conclusions, and (h) communicating process and results, seems promising (F. Fischer et al., 2014). Diagnostic reasoning may not always require all eight epistemic activities, and no generally valid order is assumed for these eight activities, but rather depends on the diagnostic problem and the situation in which the problem is presented, as well as the expertise of the diagnostician. Thus, it is not only the order, but also the quality of these activities that determines diagnostic success (Heitzmann et al., 2019).

In collaborative diagnostic reasoning, these individual diagnostic activities are extended by collaborative activities (see 1.2.1). When collaboration partners have roughly equally distributed knowledge, engaging in all proposed collaborative activities has shown to be beneficial

for team performance (Andrews-Todd & Forsyth, 2020), but when collaboration partners are dependent on each other's knowledge, research has emphasized the importance of sharing and eliciting information (e.g., Tschan et al., 2009). This is in line with the transactive memory theory emphasizing that when information is distributed among collaboration partners, it is important to know how the information is distributed among the collaboration partners (Wegner, 1987). Building on these considerations, transactivity is understood as the extent to which collaborators operate on the reasoning of their collaboration partners (Teasley, 1997). Recently, two key aspects of transactivity have been highlighted for collaborative learning that are also relevant for collaborative problem-solving: *Novelty* allows contributions to be enriched with new ideas, laying the groundwork for the collaborative construction of a shared problem representation, and *reference* keeps contributions connected (Vogel et al., 2023). In order to share and elicit novel but referenced information, it is crucial to know what collaboration partners know (or do not know). This is consistent with research on group awareness tools, which emphasizes the need for knowledge and information awareness, i.e., knowledge of others' knowledge and information (Engelmann & Hesse, 2010). A study by Noroozi et al. (2013) investigated the effect of providing participants with an external script that made them aware of the knowledge distribution in the group, in terms of building a transactive memory system and problem-solving performance. Sixty participants from two disciplinary backgrounds worked in pairs to promote sustainable behavior among farmers. Fifteen pairs received a transactive memory script, while the others did not. Results showed that the presence of the script, which makes participants aware of the knowledge distribution, facilitated the construction of a transactive memory system, improved the quality of problem-solving plans, and transferred knowledge from the group to the individual, but not vice versa. Another study found that when information is not shared within the team, this can lead to inaccurate diagnoses (Larson et al., 1998). Thus, the ability to effectively pool information (i.e. elicit and share information) is crucial for successful collaborative diagnostic reasoning. In particular, in interdisciplinary collaboration, the pooling of information seems to be the most relevant for collaborative activities. Elicitation involves requesting information from a collaboration partner in order to access additional knowledge resources (Weinberger & Fischer, 2006). Sharing involves identifying the information needed by the collaborator to build a shared problem representation (Rochelle & Teasley, 1995).

Building on research on collaborative problem-solving and diagnostic reasoning, the CDR-M (Radkowitsch et al., 2022) proposes a joint perspective in solving diagnostic problems (Abele, 2018) in a collaborative effort (see Figure 1). The CDR-M is based on the SDDS model

(Klahr & Dunbar, 1988) and its extension by van Joolingen and Jong (1997). In the extended SDDS model (van Joolingen & Jong, 1997), which focuses on learning in knowledge-rich domains, a learner hypothesis space has been added that contains all hypotheses that can be searched for without additional knowledge. The CDR-M builds on these considerations and describes the relations between individual characteristics, diagnostic processes, and diagnostic outcomes. As in the SDDS model, collaborative diagnostic reasoning involves activities within an evidence and hypothesis space; however, unlike the SDDS, in the CDR-M these spaces are understood as cognitive storages of information. This is more in line with the extended dual search space model of scientific discovery learning (van Joolingen & Jong, 1997).

Figure 1

Collaborative Diagnostic Reasoning Model

Note. This figure is used from Radkowitsch et al. (2022, p. 120)

In summary, for successful collaborative diagnostic reasoning it is essential to coordinate between evidence (data) and hypothesis (theory) by engaging in individual and collaborative activities. More specifically, the CDR-M distinguishes between *collaborative diagnostic activities*, namely eliciting, sharing, negotiating, and coordinating evidence and hypotheses, and *individual diagnostic activities*, namely generating and evaluating evidence and hypotheses and drawing conclusions (Radkowitsch et al., 2022). For evidence and hypotheses to become part of a shared diagnostic space, and thus to construct and maintain a shared understanding of the problem (Rochelle & Teasley, 1995), diagnosticians need to enact the proposed individual and collaborative diagnostic activities with high quality (Radkowitsch et al., 2022).

The collaborative diagnostic reasoning process is influenced by four factors, namely professional knowledge, collaboration knowledge, general cognitive skills, and general social skills

(Radkowitsch et al., 2022). *Professional knowledge*[3] refers to knowledge about strategies, concepts and procedures in a specific domain (see 1.3.2). A distinction can be made between conceptual, strategic and conditional knowledge. Conceptual knowledge refers to declarative knowledge about pathophysiological relations underlying a disease, also known as biomedical knowledge (Boshuizen & Schmidt, 1992). Strategic knowledge entails knowledge about problem-solving (Schmidmaier et al., 2013), and conditional knowledge describes knowledge about when to successfully apply the other two parts (Stark et al., 2011). Previous research has shown that it is not the mere existence of knowledge that is important for accurately diagnosing a patient, but rather the goal-directed application (Kiesewetter et al., 2016; Kiesewetter et al., 2020). To be able to apply this knowledge in a goal-directed way, *collaboration knowledge* is required, i.e. a combination of meta-knowledge about the collaboration partner (e.g. goals, typical requirements, Engelmann & Hesse, 2010) and internal collaboration scripts (Kollar et al., 2006). Especially when professional or collaboration knowledge is low, general cognitive and general social skills are relevant (Radkowitsch et al., 2022). *General cognitive* skills refer to domain-general problem-solving skills (see 1.2.1), which are especially relevant if domain-specific schemata based on professional knowledge are missing to guide the collaborative diagnostic reasoning process (see 1.2.2). *General social skills* mainly influence the collaborative aspect of collaborative diagnostic reasoning and less the individual diagnostic reasoning aspect (Graesser et al., 2018). Social skills are considered particularly important when collaboration knowledge is low (F. Fischer et al., 2013) and are understood as the ability to share and negotiate ideas, to coordinate, and to take the perspective of collaboration partners (Radkowitsch et al., 2022; see also Liu et al., 2016, and Hesse et al., 2015). In addition, the importance of a collaboration partner in the collaborative diagnostic reasoning process may be diminished if the diagnostician has a great deal of prior knowledge and is thus able to solve the diagnostic problem solely through individual diagnostic reasoning. In this case, the evidence that can be generated through collaboration may be less relevant to the diagnostic outcome than when the diagnostician has less knowledge. However, most clinical problems are too complex (i.e., require too much specialized knowledge to be known by one physician) to be solved individually and require collaborative diagnostic reasoning (Kiesewetter et al., 2017; Patel et al., 2002). In practice, this often functions insufficiently. For example, previous research has shown that inadequate information sharing has been identified as a major cause of errors in radiology (Brady, 2017) and emergency medicine (Tschan et al., 2009). One explanation is that groups often fail to successfully pool the information held by different members (Stasser & Stewart, 1992). Thus,

[3] Also known as domain knowledge (Hetmanek et al., 2018) or content knowledge (Förtsch et al., 2018)

it is crucial to develop domain-specific schemata based on professional and collaboration knowledge to effectively guide the collaborative diagnostic reasoning process.

1.3.2 Expertise in Collaborative Diagnostic Reasoning

To reach a diagnosis, diagnosticians often map similarities and differences among their current and previous patients or prototypical examples with the aim to engage in System 1 processes, which demand less cognitive resources than effortful System 2 processes, because like problem-solving, diagnostic reasoning is dependent on working memory capacity (Croskerry, 2009; Dumas et al., 2018; Hruska et al., 2016; Sweller et al., 2011). Thus, diagnostic problem-solving, whether performed individually or collaboratively, is a function of the domain-specific prior knowledge an individual possesses and, more specifically, the quality and organization of that knowledge (Patel et al., 1994). While the same considerations for expert performance presented in 1.2.2 apply to the medical domain, the key to expert performance in diagnostic reasoning is seen in the formation of so-called *illness scripts* (Bowen, 2006).

Illness scripts serve as cognitive representations of an illness, encompassing typical symptoms and findings derived from these encapsulated biomedical and clinical knowledge structures (Schmidt & Rikers, 2007). Such scripts consist of problem representations constructed from previously solved problems, which include enabling conditions (i.e., patient and contextual factors), fault (i.e., the underlying pathophysiological processes), and consequences (i.e., symptoms when the fault occurs; Custers, 2015). There are several important factors entailed in illness scripts (Charlin et al., 2007): (1) the knowledge stored in an illness script is not exclusive, so it can be stored in multiple illness scripts; (2) when one illness script is activated, it can lead to the activation of other illness scripts, for example, for an illness that is often confused; (3) illness scripts have slots that correspond to attributes associated with the specific illness they describe, with expectations about values that may or may not be found in a patient case for each attribute. For each slot, the attribute value with the highest probability of occurrence is the default value; and finally, (4) when confronted with a patient case, illness scripts are instantiated with information from that specific patient case, and these instantiated illness scripts are used to update existing illness scripts. Thus, instantiated illness scripts are problem representations constructed from information related to the diagnostic problem, but guided by a generic illness script (Bellezza & Bower, 1981; Graesser et al., 1980). In many problem situations (e.g., routine problems), illness scripts are activated and instantiated automatically, without conscious awareness (System 1). Only when more than one illness script is activated simultaneously for a problem situation, or when some of the default values in the activated script contradict the information in the patient case that requires conscious reasoning (System 2;

Charlin et al., 2007). Because of this unconscious process of script activation, research has found that expert performance is impaired when information is not presented in the usual order (Coughlin & Patel, 1987). Summarizing these aspects, illness scripts "contain […] encapsulated pathophysiological knowledge of the disease and its consequences, in addition to clinical knowledge of the constraints under which a disease occurs" with default values guiding the diagnostic process (Schmidt & Boshuizen, 1993a, p. 214).

Illness scripts are developed due to *knowledge encapsulation* through experience with diagnostic problems (Boshuizen et al., 1995; Schmidt & Boshuizen, 1993a). A process similar to knowledge compilation, transforming declarative knowledge into procedural knowledge (Anderson, 1985; see 1.2.2), called encapsulation, is crucial: Through repeated exposure to diagnostic problems and thus the application of conceptual biomedical knowledge, these structures become procedural networks organized in illness scripts (Schmidt & Boshuizen, 1993a). Compiled knowledge is automatically and effortlessly activated by relevant cues in a case because repeated activation in response to the same cues has caused its compilation (Anderson, 1983). In the absence of illness scripts, novices must engage in System 2 reasoning processes that are considered deliberate, slow, and error-prone (Rikers et al., 2000). However, with repeated use, pathophysiological knowledge is encapsulated in simplified models that are more efficient but have the same explanatory power. With increasing expertise, diagnosticians use System 1 reasoning processes through activated illness scripts, but they still have access to their declarative knowledge and use it when necessary, for example during patient communication or when diagnosing particularly difficult patient cases (Charlin et al., 2007; Patel et al., 1990). The process of knowledge encapsulation is based on empirical findings from studies investigating how experts, intermediates, and novices recall clinical cases, i.e., what their problem representation looks like (Kintsch & Greeno, 1985). These experiments typically follow the same four-step approach (Patel & Groen, 1986): (1) reading a clinical case, (2) recalling the case, (3) explaining the signs and symptoms, and (4) providing a diagnosis. Boshuizen and Schmidt (1992) replicated Patel and Groen's (1986) study with 20 participants (six novices, four lower and five higher intermediates, and five experts) in an online setting using think aloud protocols. The results show that most biomedical knowledge was recalled by lower intermediates. This phenomenon has become known as the *intermediate effect* (Schmidt & Boshuizen, 1993b), which describes an inverted U-shaped relation instead of a monotonically increasing function with increasing expertise (Patel et al., 2005). Building on these findings, Schmidt et al. (1988) manipulated the amount of time available to process the clinical case. They studied 120 participants (24 each of novices, lower and upper intermediate, and experts) who were given either

the traditional three minutes and 30 seconds (Patel & Groen, 1986), one minute and 15 seconds, or only 30 seconds to study the patient case (step a). Instead of think aloud protocols, they used written text for steps one to three (without time constraints). The results showed that when processing time is restricted, the intermediate effect disappeared in both recall and pathophysiology. Instead, a weak positive linear relation with expertise level was found. Thus, it appears that experts and novices use different knowledge when representing a clinical case due to knowledge encapsulation (Schmidt & Boshuizen, 1993a). While novices recall the fewest aspects of a case, intermediates recall the most if given enough time, but only experts benefit from their superior knowledge structures (illness scripts) by being able to recall relevant information under time pressure, resulting in accurate diagnoses. In summary, knowledge encapsulation is the result of extensive practice and confrontation with actual patients, leading to illness scripts, which are schemas that restructure biomedical knowledge. This restructuring eventually leads to abbreviations in lines of reasoning (cf. Koedinger & Anderson, 1990).

In order to develop expertise in (collaborative) diagnostic reasoning, it is therefore crucial to develop illness scripts by encapsulating knowledge through gaining experience (Boshuizen et al., 1995). The theory of knowledge restructuring through case processing (Boshuizen et al., 2020) posits that repeated exposure to complex cases is fundamental to the development of illness scripts. As professionals encounter and process a variety of cases, they undergo significant cognitive adaptations that allow them to refine their knowledge structures, integrating both theoretical understanding and practical experience. This restructuring enables professionals not only to perform routine tasks with increasing efficiency, but also to adapt to novel situations that may fall outside the scope of their initial training. The importance of case-based experience lies in its ability to foster the development of illness scripts that are critical to expert performance. Furthermore, the theory of knowledge restructuring through case processing emphasizes that expertise is not simply the accumulation of knowledge, but the ongoing restructuring of that knowledge in response to new and diverse cases. This process is supported by empirical evidence from several domains, including medicine, counseling, business management, and law, where case processing is integral to professional practice (Boshuizen et al., 2020). The ability to draw on previous case experiences allows professionals to identify patterns, anticipate outcomes, and adjust their approaches as needed, which is essential for maintaining high levels of performance in dynamic and complex professional environments.

Building on the theory of knowledge restructuring through case processing and case-based learning (Kolodner, 1992), Radkowitsch et al. (2023, September) recently proposed a conceptual model for the development of diagnostic reasoning competence. The model assumes that

through experience with cases, learners develop a database of previously seen cases in their long-term memory. When confronted with new cases, cues within these cases activate illness scripts and relevant cases from this database, both of which guide the subsequent diagnostic process. If the activated case helped to diagnose the patient's case (i.e., solved the problem at hand), this leads to a greater likelihood that this case from the database will be activated in the future when confronted with similar cases.

In summary, experience with cases (1) encapsulates knowledge into illness scripts and (2) creates and updates a database of already seen cases, resulting in prototypical abstract case representations. Both lead to greater accuracy and efficiency in solving diagnostic problems, both individually and collaboratively. The pedagogical implications of the presented theoretical and empirical evidence are straightforward: Early exposure to patient cases is considered essential for the restructuring and reorganization of biomedical knowledge (Eva, 2005; Lubarsky et al., 2015). Therefore, medical students should be exposed to a large number of patient cases with different diseases to develop expertise in (collaborative) diagnostic reasoning.

1.3.3 Agent-Based Simulations to Support Collaborative Diagnostic Reasoning Skills

Collaborative diagnostic reasoning skills are essential professional skills in the medical field. They must be developed through training that involves exposure to patient cases, allowing learners to construct illness scripts. Consequently, simulation-based learning is regarded as an effective approach in medical education, providing learners with authentic patient cases and preparing them for real patient contact (Issenberg et al., 2005). However, in many simulation-based courses, only a limited number of students interact with the simulation, while the majority observe the ongoing process (Zottmann et al., 2018). Theories such as social learning theory, vicarious learning, and cognitive apprenticeship suggest that learning can occur through mere observation of others (Bandura, 1977; Bandura, 2008; Collins et al., 1991). However, models like the ICAP framework (Chi & Wylie, 2014) argue that acquiring complex skills, such as collaborative diagnostic reasoning, requires active engagement with the task, which cannot be achieved through observation alone (cf. Stegmann et al., 2012). This is consistent with the transition from paper-and-pencil assessment to technology-based assessment to facilitate interactive tasks and collaborative problem-solving (see 1.2.4; Graesser et al., 2018; Herborn et al., 2020). In order to overcome the limitations of assessing collaborative problem-solving skills on an individual basis and without the dependency of the collaboration partner, agent-based collaboration partners are incorporated into the assessment process. Adopting this approach from an assessment context to a learning context has the potential to facilitate the acquisition of collaborative diagnostic reasoning skills for a greater number of learners. This is achieved

by enabling them to interact in relevant simulated situations, rather than merely observing them, thereby enhancing their learning experience. With respect to collaborative diagnostic reasoning skills, empirical studies have demonstrated that physicians frequently exhibit deficiencies in information sharing skills, which represent a critical subskill of collaborative diagnostic reasoning skills (Kiesewetter et al., 2017; Tschan et al., 2009).

Accordingly, the agent-based CoSiMed simulation was developed to foster these subskills and is used as a training and assessment instrument for collaborative diagnostic reasoning. The simulation was developed and validated by Radkowitsch et al. (2020). By conducting interviews with seven practitioners from the fields of internal medicine and radiology, the researchers were able to identify a specific scenario that highlighted the ongoing challenges associated with the sharing and elicitation of information with and from collaboration partners, as previously documented by Tschan et al. (2009): Requesting a radiologic examination during collaborative diagnostic reasoning of a patient case (see Appendix 8.1). This is because it necessitates the sharing and elicitation of evidence and hypotheses that inform the diagnostic process (Davies et al., 2018). The CoSiMed simulation was developed through a joint effort involving medical experts, psychologists, and software engineers. As the majority of actual collaborations between internal medicine and radiology are presumed to be document-based, and as simulations are intended to represent actual practice, the CoSiMed simulation is also document-based. This implies that the required information is provided in written or video format, whereas in interaction-based simulations, it can only be accessed through active interaction with the patient. The advantage of document-based simulations is that learners have more opportunities to reflect on their processes due to the absence of time pressure (Heitzmann et al., 2019).

In the CoSiMed simulation, learners are instructed to assume the role of an internist and are required to complete three steps (Radkowitsch et al., 2020): (1) reading the health record of the patient case, (2) collaborating with the agent-based radiologist, and (3) reaching a solution. In the initial phase, learners are required to engage in evidence generation, with the objective of developing an initial individual problem representation. By reading through the different parts (e.g., medical history, physical examination, and laboratory values) of the health record ideally the entailed cues activate one or more illness scripts which then lead to initial suspected diagnoses, thereby structuring the following collaboration with the agent-based radiologist. During the collaboration, learners must enact the collaborative diagnostic activities of evidence elicitation and sharing, as well as hypotheses sharing with high quality, in order to obtain results from the agent-based radiologist. The learners are thus required to complete a radiological request form, selecting from a total of 42 different combinations of radiological methods and

body parts (evidence elicitation), sharing patient information from the health record (evidence sharing), and suspected diagnoses from 249 possible diagnoses (hypotheses sharing) that are considered relevant for the agent-based radiologist. The aforementioned request form can be considered the shared problem representation in this agent-based collaboration. Only if this request form is filled out sufficiently, learners receive the information they asked for from the agent-based radiologist, otherwise they get the opportunity to revise the request form (three times in assessment situations and up to ten times in learning situations). The final step in the process is for learners to draw conclusions based on the previous collaboration with the agent-based radiologist. This entails indicating the final diagnosis and providing a justification for it.

The CoSiMed simulation, comprising fictitious but authentic patient cases, was validated by investigating the differences between three levels of expertise (low (n = 45), advanced (n = 28), and high (n = 25) prior knowledge level) with respect to the participants' diagnostic accuracy, diagnostic efficiency, their information sharing skills, as well as their intrinsic cognitive load (see 1.2.3) and perceived authenticity (Radkowitsch et al., 2020). The results showed that, on average, practitioners with high levels of prior knowledge perceived the CoSiMed simulation as authentic and had significantly higher diagnostic accuracy, diagnostic efficiency, and information sharing skills than the low prior knowledge group. However, there was no significant difference between the high and advanced prior knowledge groups, except for intrinsic cognitive load. This finding is consistent with previous research on medical expertise conducted in the absence of time pressure (see 1.3.2). Moreover, as anticipated, the high prior knowledge group exhibited significantly lower intrinsic cognitive load, followed by the advanced and the low prior knowledge groups (Radkowitsch et al., 2020). In conclusion, the CoSiMed simulation can be considered a valid instrument for assessing and supporting the development of collaborative diagnostic reasoning skills.

Previous research has employed the CoSiMed simulation to examine the effects of different types of instructional support, such as collaboration scripts and reflection phases, on learning. The importance of such instructional support is underscored by evidence that simulation-based learning of problem-solving—such as collaborative diagnostic reasoning—is most effective when learners receive guidance (Leutner, 1993). Without instructional support, unguided problem-solving can place excessive demands on learners' working memory capacity, potentially impairing their ability to learn effectively (P. A. Kirschner et al., 2006). Radkowitsch et al. (2021) investigated whether learners benefit more from the provision of an adaptive collaboration script than from a static collaboration script or no support at all (see 1.4.2). The study, which involved 160 intermediate medical students randomly assigned to one of three

conditions, revealed that the performance of evidence sharing was facilitated by an adaptive collaboration script, while the performance of evidence elicitation was also facilitated by the static collaboration script, with respect to collaborative diagnostic activities. Moreover, the researchers discovered that the adaptive collaboration script enhanced the learners' perception of competence. Building on these results, Richters et al. (2022) investigated the effect of adaptive collaboration scripts and reflection on individual diagnostic activities, specifically with regard to the quality of evidence sharing, hypotheses sharing, and diagnostic accuracy. Additionally, the role of prior knowledge, including both professional and collaboration knowledge, was examined. The researchers employed a 2x2 design with 151 intermediate medical students. The findings indicated that adaptive collaboration scripts are more beneficial for learners with low prior knowledge, whereas reflection on individual diagnostic activities enhances outcomes for those with high prior knowledge. To gain deeper insights into the role of reflection in facilitating collaborative diagnostic reasoning, Richters, Stadler, Brandl, et al. (2023) investigated the effects of low and highly structured reflection phases on collaborative diagnostic activities regarding the quality of those collaborative diagnostic activities and diagnostic outcomes, with a particular focus on learners' collaboration knowledge. The data set comprised 195 intermediate learners engaged in the CoSiMed simulation, with an equal distribution across the three experimental conditions (low-structured, high-structured, and no reflection support). Results showed a moderating role of prior knowledge in the effectiveness of structured reflection: Learners with low collaboration knowledge benefit from low-structured reflection while both forms of reflections are not beneficial for learners with high collaboration knowledge. The findings indicate that different forms of instructional support facilitate the learning of collaborative diagnostic reasoning using an agent-based simulation.

It is notable that the studies made use of data collected as a result of learners interacting with the simulation, such as how they filled out the request form. Such data are referred to as *process data*, which are stored immediately in log-files without the need of extra measurement (Goldhammer et al., 2017). The following chapter will examine the potential of process data to enhance the assessment and support of collaborative problem-solving skills.

1.4 Improving Assessment & Support using Process Data

The cognitive (and collaborative) activities that play a role in (collaborative) problem-solving have long been a subject of interest to researchers. Initially, however, it was only possible to infer the outcome, for example, through think-aloud protocols in expertise research (see 1.3.2; Ericsson & Simon, 1980). The advent of technology-based interactive tasks and simulation-based learning using computerized tasks has enabled a closer approximation to reality

through the monitoring of the process, as observable problem-solving behaviors, which are stored in computer-generated log-file data and can be accessed to provide additional information (Bunderson et al., 1988; Goldhammer et al., 2020; Goldhammer et al., 2017). A significant benefit of utilizing process data is that it can be gathered without disrupting the natural flow of the task, thus avoiding any additional measurements that might increase the cognitive load on the participants (Matcha et al., 2019). Consequently, research can now examine the sequences of thinking and action that underpin the problem-solving performance, facilitating an analysis of the problem-solving process without the necessity of additional measurement (Csapó & Funke, 2017; He & Davier, 2015). For example, Stadler, Hofer, and Greiff (2020) employed process data from 1,491 9th graders working on five technology-based individual problem-solving tasks to demonstrate that participants exhibited significant differences in both the time required to solve the problem and the number of interactions performed, despite exhibiting similar performance outcomes. This reinforces the notion that, with regard to problem-solving skills, the solution itself is not the only relevant factor; the process by which the solution was reached is also of importance (Greiff et al., 2013). This is in accordance with the OECD (2017) definition highlighting that collaborative problem-solving is an attempt to solve a problem not only the solution to a problem. Therefore, different behaviors can actually represent differences in skills beyond product data (Stadler, Hofer, & Greiff, 2020). Building on the aforementioned line of reasoning, this chapter introduces process data analyses and presents a selection of relevant studies that employ it in the context of collaborative problem-solving. It then highlights the advantages of using process data as a valuable source of evidence for assessing and supporting collaborative problem-solving skills. Finally, the chapter concludes with an in-depth examination of the key challenges associated with utilizing process data.

1.4.1 Process Data Analyses

Process data allows "a potentially fluid window into the minds" of individuals during the problem-solving process (Rupp et al., 2012, p. 73). Thus, process data allows researcher to give answers to the question "what particular [collaborative problem-solving] behaviors give rise to successful problem-solving outcomes?" (Sun et al., 2022, p. 1), which is needed to understand, assess, and support collaborative problem-solving skills.

Traditionally, educational assessment has made a distinction between two types of data: product data and process data for performance measures. In this context, *product data* refers to the solution to a given task, while *process data* refers to the methodology employed to achieve this solution (Levy, 2020; Zumbo et al., 2023). Process data can be understood as "as any data automatically collected about test-takers' response process" (Anghel et al., 2024, p. 2).

However, as Ercikan et al. (2020) noted, process data are merely traces of cognitive processes. Therefore, it is crucial to have a robust theoretical foundation to ensure the accurate matching of these traces to relevant psychological constructs (Knight & Buckingham Shum, 2017; J. Wong et al., 2019). In light of these considerations, a distinction must be drawn between log-file data and process data (Provasnik, 2021). Log-file data refers to the information stored during interactions with technology-based tasks, whether for assessment of or support for collaborative problem-solving. Process data, on the other hand, represent the psychological constructs to which the information in log-file data is matched. Thus, while log-file data are understood to exist as a byproduct of interactive and simulation-based tasks, process data must be extracted from log-file data in the presence of relevant theory (Goldhammer et al., 2020). Nevertheless, this clear differentiation between log-file and process data is not consistently observed in the literature. This is despite the fact that log-file data is high-dimensional and heterogeneous information that requires careful consideration to be transformed into meaningful process data (Anghel et al., 2024; Lindner & Greiff, 2023). Goldhammer et al. (2021) proposed an approach to move from seeing log-file data as by product of technology-based tasks but incorporate process data into the design of such tasks. They also provided an approach how to transform log-file data into process data: (1) information stored in log-file data is labelled with *low-level features*, that is meaningful actions or states within the context of the specific task. (2) These low-level features are aggregated to form *high-level features* (Mislevy, 2019), which represent meaningful *process indicators* of psychological constructs. Therefore, while low-level features can only be interpreted in light of the concrete task, high-level features, especially when informed by theory, allow for more generalizable results from process data analyses (Tomasevic et al., 2020).

An example of the utilization of low-level features can be observed in the study conducted by Ma et al. (2023), which employed data from 9,841 students in China who completed the Xandar task from PISA 2015. The researchers identified four distinct profiles of collaborative problem solvers based on their time on task, number of actions, and collaborative problem-solving skill levels. The four profiles were identified as disengaged, struggling, adaptive, and excellent. The study found that the disengaged profile was characterized by minimal time and actions on task, resulting in poor collaborative problem-solving skills. While the struggling profile was characterized by more time and action on task, but also resulted in poor collaborative problem-solving skills. Conversely, the excellent profile showed the highest performance in collaborative problem-solving skills with efficient use of time and actions. The adaptive profile still has relatively high collaborative problem-solving skill performance, but is

characterized by the greatest number of actions. The findings suggest that collaborative problem-solving performance can vary considerably depending on individual behaviours, with efficiency and skill level being pivotal factors in determining success. However, as the researchers employed low-level features, it is challenging to generalize the findings to other contexts.

In contrast, the study by Andrews-Todd et al. (2023) examined the manifestation of collaborative problem-solving skills across different tasks. The researchers analyzed the interactions of 100 students aged 12-15, who were randomly assigned into pairs. The study comprised two separate tasks: The T-Shirt Math Task, which focused on linear functions and argumentation, and the Physics Playground, an educational game on Newtonian physics. The researchers employed an ontology-based competency model to code the collaboration skills exhibited during the tasks, thereby deriving high-level features. The video recordings were analyzed by trained raters who identified nine distinct collaborative problem-solving skills. The study revealed that specific skills, such as sharing information and negotiating, were frequently observed across both tasks, indicating their importance in collaborative problem-solving regardless of task characteristics. However, the prevalence of other skills varied depending on the task, indicating that the effectiveness of specific collaborative problem-solving skills may be task-dependent.

The analyses of collaborative problem-solving through the lens of process data can be approached in three distinct ways (Ulitzsch et al., 2023): theory-based, exploratory, and predictive. Most of the studies analyzing process data of collaborative problem-solving utilize process data to explain performance differences or to gain deeper insights into the process.

Theory-driven approaches are employed with the objective of enhancing comprehension of the construct and supporting the refinement of existing theories. Consequently, they seek to identify particular strategies that have been derived from the theory (Ulitzsch et al., 2023). Nevertheless, purely theory-driven approaches are rare, particularly in the context of collaborative problem-solving. One illustrative example of such strategies in the context of individual problem-solving is the application of the strategy of varying one thing at a time (VOTAT). For instance, Greiff et al. (2015) employed log-file data from 16,219 students who participated in PISA 2012 to investigate whether the implementation of the VOTAT strategy in a problem-solving task was associated with their performance in that task. Consequently, the researchers used log-file data to derive a dichotomous variable indicating whether or not VOTAT was applied. The results indicated a strong correlation between the application of VOTAT and item performance. Additionally, the difference in performance between students who applied VOTAT and those who did not was statistically significant. These findings support the hypothesis that the VOTAT strategy is a significant predictor of success in problem-solving tasks.

Exploratory approaches, like theory driven approaches, seek to enhance understanding of the information embedded within process data. However, while theory-driven approaches examine particular strategies with the objective of validating theories, exploratory approaches ideally use theory to construct process indicators and use them as features in prediction models and sequence mining with the goal to uncover key behavioral patterns that distinguish success from failure (Ulitzsch et al., 2023). Therefore, while Greiff et al. (2015) constructed a binary variable indicating the presence or absence of the VOTAT strategy, aggregating hundreds of clicks into a single binary variable in exploratory approaches, the focus is on the complete problem-solving process. One illustrative example of an exploratory approach combined with theory is the study conducted by Richters, Stadler, Radkowitsch, et al. (2023), who employed n-grams (Damashek, 1995) of collaborative diagnostic activities to predict diagnostic accuracy (see 1.3.1). The aforementioned process indicators have been constructed from log-file data and represent theory-based features. The coding of each click in the simulation as a diagnostic activity was followed by the transformation of these data points into bigrams, which facilitated more effective interpretation. This approach allowed for the examination of specific aspects of the diagnostic process, such as the time spent on a single activity or the frequency of transitioning from one activity to another. Using data from 73 students working on the CoSiMed simulation (see 1.3.3) they could show that a random forest (Breiman, 2001) prediction model is capable to predict diagnostic success using bigrams of diagnostic activities after approximately two thirds of the median time working on the task. Moreover, the researchers found that diagnosticians who spent more time with individual diagnostic activities were more likely to be successful, while those who spent more time with collaborative diagnostic activities were more likely to be unsuccessful (Richters, Stadler, Radkowitsch, et al., 2023).

The use of predictive approaches in this research area is a comparatively recent phenomenon and primarily focused on improving predictive accuracy (Ulitzsch et al., 2023). In contrast to theory and exploratory approaches, predictive approaches alter the perspective. While theory and explanatory approaches are concerned with understanding the underlying processes, predictive approaches are focused on predicting future outcomes. Accordingly, the primary objective is to achieve the highest possible level of predictive accuracy, which can be attained by selecting the ratio of bias and variance that minimizes the occurrence of error. In order to achieve this, it is essential to leverage large data sets and metrics for evaluating the prediction, rather than the representation of internal structure. Furthermore, it is crucial to be open to allowing for bias and nonlinearity in pursuit of superior prediction accuracy (Molnar et al., 2020; Yarkoni & Westfall, 2017). This allowance can result in the development of highly complex

prediction models. While these models may be accurate, their internal mechanisms may lack transparency, leading to less interpretable models, also known as black boxes (Molnar et al., 2018; Yarkoni & Westfall, 2017). Such models have been applied in the field of learning analytics with the objective of identifying students who are at risk of failing and therefore require additional instructional support (Leitner et al., 2017). One illustrative example is the study conducted by Costa et al. (2017), which employed a predictive model to ascertain the likelihood of a student failing a university course. The model was trained on data encompassing socio-demographic characteristics (e.g., age, gender, income) and log-file data (e.g., access frequency to the learning platform, participation in the discussion forum, and the amount of received and viewed files). The results showed that the model could identify students at risk of failing after 10% of the course had been completed with at least 50% accuracy. However, while this offers a promising approach to prevent students from failing their course, it is important to note that, in contrast to the studies presented for the theory-driven and exploratory approaches, Costa et al. (2017) employed low-level and context-dependent features in the absence of a theoretical framework, which limits the generalizability of their findings.

In summary, process data analyses facilitate a more profound comprehension of collaborative problem-solving behaviors, as well as the construction of predictive models that can anticipate future outcomes. By understanding the causes of performance differences through the analysis of behavioral patterns (Eichmann et al., 2020), educators can develop predictive models to tailor interventions, thereby enhancing personalized learning experiences (Tetzlaff et al., 2021).

1.4.2 Benefits of Process Data Analyses

As discussed in the preceding section, technology-based, interactive and simulation-based tasks, which facilitate collaborative problem-solving skills, offer a promising approach to analyzing process data. While most theoretical and explanatory-based approaches are concerned with developing a deeper understanding of the collaborative problem-solving process, predictive approaches aim to advance adaptive learning support. Taken together, the use of process data allows for enhancements of personalized learning experiences in the development of collaborative problem-solving skills.

The incorporation of interactive and simulation-based tasks in technology-based assessments enables the analysis of test-taking behaviors, thereby providing additional information beyond performance outcomes (Greiff et al., 2016; He & Davier, 2015). For example, Han et al. (2023) were able to identify different collaboration strategies and highlight the importance of establishing and maintaining a shared problem representation. Their findings suggested that

a structured approach to agreeing on a team strategy leads to better performance than the trial-and-error approach. The study is based on process data (response times and number of actions taken during the Xandar task) from 2,520 students who participated in the PISA 2015 assessment. Moreover, process data can be leveraged to support data quality control. For instance, it can be used to identify instances of rapid guessing behavior, which may indicate a lack of thoughtful engagement with the task. This can be achieved by setting a task-specific threshold requiring at least a brief period of reading and thinking about the task (S. L. Wise, 2017). While recent research indicates that rapid guessing, and thus a lack of engagement and cognitive processing with multiple choice questions, represents a threat to the validity of individual responses, it has less impact on aggregate scores and country rankings, as seen in large-scale assessments like PISA (Michaelides et al., 2024). There is, however, consensus that it is beneficial to be able to identify atypical behavior using clickstream data, thus enhancing data quality (Tang et al., 2023). This also permits an enhancement in measurement precision (Davier et al., 2019), the validation of test score interpretations (Ercikan & Pellegrino, 2017), and the optimization of the test design (van der Linden, 2008).

Furthermore, technology-based assessments facilitate the adaptation of tasks to different domains and learners/test takers, or even the use of similar tasks for assessment and learning settings. This is made possible by the opportunity of immediate analysis of the data, which in turn allows for the provision of feedback and reports to learners and stakeholders for decision-making purposes (Ifenthaler & Greiff, 2021). These developments permit the integration of assessment and learning due to the potential for continuous, feedback-oriented, and multifaceted data collection, thereby facilitating personalized support (Thille et al., 2014).

Support of the learners by using process data can also be achieved by predicting learner performance, thus enabling researchers to identify individuals who are at risk of inadequate performance. This includes, for example, those learners who are unlikely to benefit from engaging in a specific learning activity (Leitner et al., 2017). This enables a shift in the educational paradigm from a one-size-fits-all approach to personalized education, allowing for the systematic adaptation of instruction and learning materials to individual learners (Tetzlaff et al., 2021; Tsai et al., 2020). One potential approach is the implementation of a learner model, which employs assumptions regarding learning prerequisites, learning processes, and anticipated learning outcomes to optimize decisions regarding the adjustment of instructional support (Basu et al., 2017). The use of learner models for personalization is of particular importance in contexts such as simulation-based learning, where research indicates that tasks can overwhelm learners by demanding excessive cognitive resources (Azevedo & Gašević, 2019). Moreover, simulation-

based learning is most effective when additional instructional support, such as scaffolding, is provided in a timely and tailored manner to meet the specific needs of learners (Leutner, 1993; Plass & Pawar, 2020).

Research on the fading effect (Puntambekar & Hubscher, 2005) and the expertise reversal effect (Kalyuga, 2007; see 1.2.2) underscores the significance of accounting for individual differences, particularly in the context of expertise, when developing instructional support. Scaffolding, a well-established form of instructional support, plays a crucial role in this context. As defined by Tabak and Kyza, scaffolding is "support that enables learners to perform an action that would be outside their independent activity" (2018, p. 191). First introduced by Wood et al. (1976), the objective of scaffolding is to provide support for the learner's current activity while simultaneously facilitating future independent performance. Examples of scaffolds include worked-out examples and metacognitive prompts that encourage reflection or provide external collaboration scripts (Kollar et al., 2018).

The concept of scaffolding is linked to Vygotsky's (1978) zone of proximal development, which describes the range of tasks a learner can perform with assistance but not independently. Given that the zone of proximal development is an individual phenomenon that evolves over time, scaffolds that are initially effective may impede learning as the learner's expertise increases (Kalyuga, 2007). The available evidence suggests that adaptive scaffolding, which is designed to adapt to the evolving needs of the learner, can lead to significantly enhanced learning outcomes in comparison to fixed or no scaffolding. This improvement is observed not only in the acquisition of declarative knowledge but also in learning processes (Azevedo et al., 2005). By ensuring that each learner receives tasks that are tailored to their specific needs and that demand an optimal level of cognitive resources, adaptive instructional support has been shown to maximize learning efficiency and effectiveness (Corbalan et al., 2006).

In a study examining the efficacy of adaptive instructional support for collaborative diagnostic reasoning, Radkowitsch et al. (2021) investigated whether learners benefit more from the provision of an adaptive collaboration script compared to a static collaboration script or no support at all. The implementation of a technology-based simulated task with an agent as a collaboration partner enabled the provision of micro-adaptive support (Tetzlaff et al., 2021). This was achieved by analyzing the completed request form in real-time to identify any missing information (see 1.3.3). Based on these missing but relevant information for the agent-based radiologist further information on the needs and goals (collaboration knowledge, see 1.3.1) was provided for the learner in form of an external collaboration script (F. Fischer et al., 2013). The study, which included 160 intermediate medical students randomly assigned to one of three

conditions, demonstrated that the performance of evidence sharing is enhanced by an adaptive collaboration script, while the performance of evidence elicitation is also facilitated by the static collaboration script. Furthermore, the researchers discovered that the adaptive collaboration script enhanced the learners' perceived competence, leading to the conclusion that the provision of adaptive collaboration scripts is an effective method for facilitating the learning of collaborative diagnostic reasoning using an agent-based simulation.

In summary, the integration of process data enables the transition towards a more adaptive and individualized instructional approach. This adaptivity ensures that learners receive the appropriate level of support, thereby optimizing their cognitive engagement and promoting learning.

The majority of process data analyses in the context of collaborative problem-solving employ theory-driven or explanatory approaches to explain performance (Ulitzsch et al., 2023). To illustrate, the study by Sun et al. (2022) examined the interplay between cognitive and social skills in collaborative problem-solving among 303 undergraduate students engaged in the Physics Playground task. Verbal communication occurring during the collaboratively played game was coded according to the framework established by Sun et al. (2020) in order to construct process indicators (see 1.2.1). Although the study did not employ log-file data, it nevertheless serves as an illustration of how process data analyses can facilitate a deeper understanding of collaborative problem-solving. This is achieved through the identification of critical interaction patterns that contribute to the success of the collaborative process. The findings indicated that conversations centered on the construction and negotiation of shared knowledge were associated with more successful outcomes, whereas discussions of inappropriate ideas were associated with less successful performance. The findings underscore the significance of regular turn-taking and active involvement in attaining effective collaboration, underscoring the socio-cognitive nature of collaborative problem-solving (Sun et al., 2022). In another study that employed process data from the PISA 2015 Assessment, De Boeck and Scalise (2019) examined the correlation between students' activity levels and their performance in collaborative problem-solving tasks. The researchers analyzed data from 986 U.S. students and found that students who exhibited higher levels of activity tended to complete tasks more quickly but performed less well overall. Conversely, successful students took more time on tasks, suggesting that taking time to construct a shared problem representation may be beneficial for performance in collaborative contexts. The study underscores the importance of balancing speed and thoroughness in collaborative problem-solving to achieve better outcomes.

In this way, process data can be utilized to examine not only the outcomes achieved, but also the means by which they were achieved, and to draw inferences about the cognitive processes involved in problem-solving (Greiff et al., 2018; Stadler et al., 2019). The aforementioned benefits for assessment, adaptive support, and theoretical advancements render process data a valuable source of evidence when assessing and supporting collaborative problem-solving skills through the use of interactive and simulation-based tasks. Nevertheless, in order to leverage the substantial advantages offered by process data analyses, it is essential to address and overcome a number of challenges.

1.4.3 Challenges of Process Data Analyses

Despite these benefits of using process data analyses for collaborative problem-solving and collaborative diagnostic reasoning in medical education, there are several challenges associated with its use: Beginning with ethical considerations before and during data collection, continuing with the complexities of data analyses, and the need for theory when interpreting the results.

Ethical considerations need to be addressed when process data is collected, as process data has the potential to contain personal and sensitive details about an individual, such as representing effort or failure to solve problems, answer questions, or learning per se (Maddox, 2023). The main ethical considerations that need to be addressed are informed consent, transparency, privacy, responsibility, validity, minimizing adverse effects, and enabling interventions. These aspects are similar to those faced in the field of learning analytics (Cerratto Pargman & McGrath, 2021; Lindner & Greiff, 2023). Learning analytics are broadly understood as a "research area that focuses on the development of methods for analyzing and detecting patterns within data collected from educational settings, and leverages those methods to support the learning experience" (Chatti et al., 2012, p. 319). Although research using data from learning management systems has some unique ethical considerations, most of the ethical concerns related to data collection are also relevant to process data in the context of collaborative problem-solving. Therefore, as the field of learning analytics has recently focused on the ethics of using process data (e.g., Ferguson et al., 2016; Francis et al., 2023; Khalil et al., 2023a; Sclater, 2016), the research and findings presented are also extended to this area of study. In a review of papers on the use of learning analytics in higher education between 2012 and 2018, it was found that more than 80% of the studies did not mention ethical considerations at all. It is not necessarily the case that all of these studies were conducted unethically, but it does point to the need for more reflective reporting on these aspects. However, the review also found an increase after 2017. Therefore, this change may already be taking place (Viberg et al., 2018).

A cornerstone of ethical data collection is to ensure that participants give informed consent and that the processes involved are transparent to the participants. This includes making learners fully aware of practices such as tracking and analyzing their data, which is often done without their explicit knowledge. Transparency includes clarifying the purpose of the data collection and subsequent analyses, the metrics used, who has access to the data, the boundaries of its use, and how the results will be interpreted (Cerratto Pargman & McGrath, 2021). Without such transparency, participants may not fully understand the implications of data collection, which can lead to distrust or ethical breaches. Privacy concerns in process data analyses and learning analytics revolve around the "restriction of access to an individual's personal information" (Francis et al., 2023, p. 104). Ensuring privacy involves addressing issues related to access and de-identification of learner data, which is critical to preventing misuse or unauthorized sharing of sensitive information (Cerratto Pargman & McGrath, 2021). Effective de-identification practices are essential to protecting learner privacy, but they must be balanced with the need for meaningful data analyses. Institutions have a responsibility to ensure that learning analytics are used legally, ethically, and effectively (Cerratto Pargman & McGrath, 2021). This responsibility extends to careful stewardship of data, ensuring that all practices comply with legal standards and ethical norms. In addition, institutions must consider the broader implications of their use of analytics, such as potential biases in data interpretation and the fair treatment of all learners. The validity of data collection and analysis is another critical ethical consideration. This includes the interpretation and location of learner data, the accuracy of the data, the validity of the algorithms, and the metrics used for predictive analytics or interventions based on learner data (Cerratto Pargman & McGrath, 2021). A major concern, as highlighted by Zumbo et al. (2023), is the impact of variability due to neurodiversity or potential disabilities, which can alter process data in unexpected ways. Such variability may lead to invalid inferences for individuals who deviate from normative patterns, particularly in measures such as time-on-task. Ethical data collection and analyses must also prioritize minimizing adverse effects on learners. This includes addressing issues of harm, nonmaleficence, and the risks associated with the management of learner data (Cerratto Pargman & McGrath, 2021). Institutions must be vigilant in ensuring that the use of data does not inadvertently harm learners or reinforce existing inequities. Finally, the use of process data analyses should focus not only on monitoring, but also on enabling timely and effective interventions. Institutions need to consider the circumstances in which they should intervene based on the results, particularly when data suggests that a learner may benefit from additional support (Cerratto Pargman & McGrath, 2021). However, research shows that enabling interventions is one of the least addressed ethical areas, with limited

guidance on how institutions should act when analytics indicate that learners are struggling (Whitelock-Wainwright et al., 2019).

In summary, the ethical considerations in process data collection are extensive and closely aligned with those in learning analytics. Transparency, privacy, and informed consent are the most frequently addressed ethical issues in research to date, while enabling interventions remain underexplored (Cerratto Pargman & McGrath, 2021). As the field continues to evolve, it is critical that research and practice prioritize these ethical considerations to effectively protect and support learners (Drachsler & Greller, 2016; Ferguson et al., 2016). This has become particularly evident in recent years with the emergence of generative artificial intelligence and large language models in education in general (Bond et al., 2024; Yan et al., 2024), but also in medical education in particular (Lucas et al., 2024). As a consequence, various frameworks and guidelines have been proposed regarding ethical and transparency aspects (e.g. Chaudhry et al., 2022; European Parliament, 2023; Simbeck, 2024). This is of crucial importance as a recent review indicated that 92% of generative artificial intelligence tools currently used for supporting learning practices are transparent only to artificial intelligence experts (Yan et al., 2023).

Once data collection is complete, the next challenge is to analyze the process data, which requires dealing with complexity. This starts with the need for various steps to pre-process the raw data, as direct analyses would often be meaningless due to technically necessary but meaningless noise in the raw data (Rupp et al., 2012). Another factor that adds to the complexity is that product data is usually stored in standard formats, such as multiple-choice questions, whereas process data has a different length for each individual (Zhan & Qiao, 2022) and is prone to situational bias (Lindner & Greiff, 2023). Chetverikov and Upravitelev (2016) investigated this in a simple visual search task with 284 participants, who used their personal computers in an online setting to complete the task. The analyses show that CPU score affects the distribution parameters, while RAM and GPU score do not. Thus, especially in small samples, differences in data collection due to technical differences can negatively affect data quality and increase measurement error.

Process data from collaborative problem-solving tend to be more complex in nature than traditional performance data: While there is only one a priori known correct outcome (or at least tasks are defined as such), there may be multiple correct strategies for solving the same problem in process data. Such heterogeneity and complexity thus require advanced statistical models beyond classical regression (Chen et al., 2019; Goldhammer et al., 2017; Lindner & Greiff, 2023). For example, alternative approaches suggest the use of Bayesian networks and a combination of process and product data to measure performance (for a review of methods for

simulation-based assessments see Klerk et al., 2015). This introduces new complexities given the dependency structure of process (and product) indicators within and across tasks, and research has suggested the use of computational psychometrics to account for this (Goldhammer et al., 2021). Given all these complexities, it is suggested to plan the desired analyses before collecting the data and to use process data not only as a by-product, but to actually pay attention to how specific process indicators can be measured and to incorporate this into the design of tasks (Goldhammer et al., 2021; Lindner & Greiff, 2023). This is consistent with the need to have a deep, contextualized understanding of the structure of process data that should be grounded in theory (R. Baker et al., 2020).

After analyzing process data, the next challenge is interpreting the result. In order to inform theories and derive information, such as which strategies are most beneficial in collaborative problem-solving or which instructional supports might be beneficial, researchers have called for a more robust link from process data to learning theories to better understand and facilitate learning (Gašević et al., 2015). It is only when log-file data are linked to theory-based process indicators that reliable and valid conclusions can be drawn (Zumbo et al., 2023). This is in line with the call to use high-level features instead of low-level features, which can only be interpreted in the light of the concrete task, for more generalizable results from process data analyses (Tomasevic et al., 2020; see 1.4.1). Using high-level features in the light of theory allows research to move beyond idiosyncratic results that are only valid for a specific assessment or learning task, and to compare and replicate findings that cannot be done directly with log-file data or only with low-level features (Goldhammer et al., 2021). This would also make it easier to overcome the challenge of task dependency when using process data (He et al., 2021). However, a recent review of the literature on the use of process data in large-scale assessments, which included 232 articles, found that only in one of the six major topics identified (digital writing) most studies did rely on a theoretical model. While all other topics (response time models, response time in general, aberrant test-taking behavior, action sequences, complex problem-solving) rarely mentioned a theoretical foundation. Although this review does not specifically focus on collaborative problem-solving, it illustrates the need to consider the use of theory when using process data (Khalil et al., 2023b).

In order to obtain meaningful and actionable results when interpreting process data, it is critical to connect the data to established theory. Ideally, this is done not only when interpreting the results of process data analyses, but also from the beginning of task design. As mentioned earlier, the use of agent-based simulations and interactive tasks (see 1.2.3) allows for ensuring that all relevant behaviors are triggered appropriately by the learner. Thus, if the researchers

know from theory which behavior is relevant, such as collaborative diagnostic activities in collaborative diagnostic reasoning (see 1.3.1), they can design tasks that ensure that this behavior is shown and stored in log-file data, such as the request form in the CoSiMed simulation (see 1.3.3), allowing for meaningful interpretation of results and actionable conclusions.

In conclusion, while the use of process data may allow researchers to answer the question, "what particular [collaborative problem-solving] behaviors give rise to successful problem-solving outcomes?" (Sun et al., 2022, p. 1), only theory-driven analyses allow to go further and ask *why* these behaviors are successful and *what can be done now* to facilitate the learning of collaborative problem-solving skills (A. F. Wise & Shaffer, 2015).

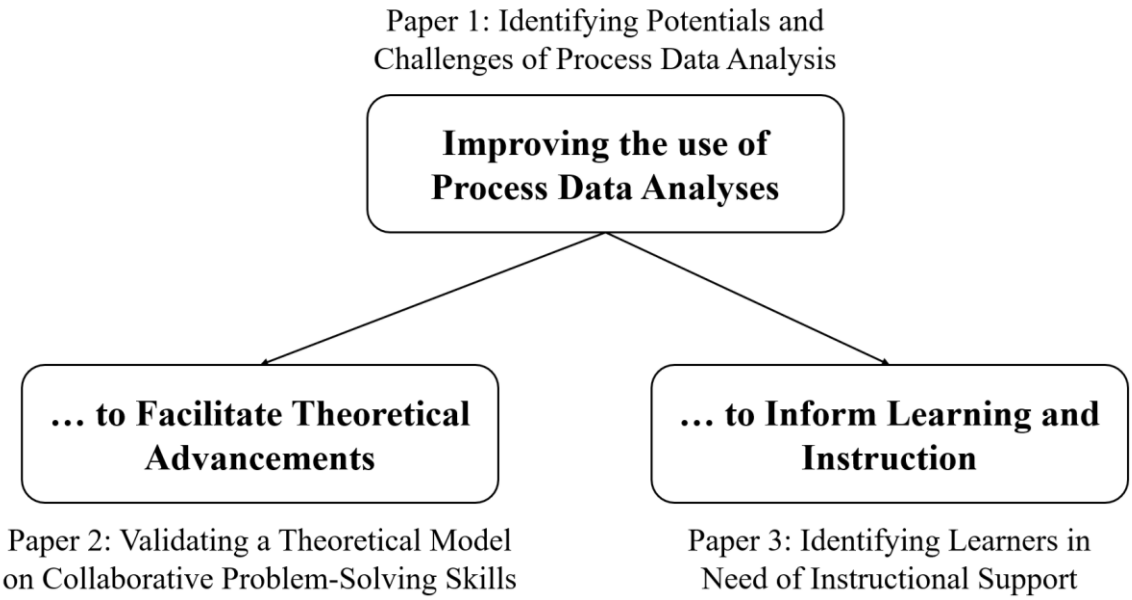
1.5 Research Questions and Outline of the Papers

Collaborative problem-solving involves multiple agents working together to solve complex tasks, with a focus on the process rather than just the outcome (OECD, 2017). Building on research on collaborative problem-solving and diagnostic reasoning, the CDR-M (Radkowitsch et al., 2022) proposes a shared perspective on solving diagnostic problems (Abele, 2018) in a collaborative effort. As individuals gain experience, they develop domain-specific knowledge that allows them to solve diagnostic problems more effectively in a collaborative effort. Simulation-based learning, particularly through technology-based interactive tasks, has been shown to be effective in enhancing these skills by providing authentic situations for knowledge application (Chernikova et al., 2020).

Recent advances in process data analyses enable researchers to collect data unobtrusively, unlike think-aloud protocols, and without requiring additional measurements that could increase cognitive load, making it easier to examine involved cognitive processes (Matcha et al., 2019). Process data derived from log-file data during task interaction can be matched to psychological constructs to gain insight into problem-solving behaviors. Three approaches - theory-based, exploratory, and predictive - are used to analyze process data, with implications for the assessment of performance differences, the development of predictive models, and the provision of personalized support (Ulitzsch et al., 2023). However, ethical guidelines, standardized data collection and analysis methods, and robust theoretical frameworks are needed to fully realize their potential. Moving forward, the focus should shift to hypothesis-driven research that employs validated indicators and responsible data use, balancing innovation with rigorous scientific standards (Lindner & Greiff, 2023). Following this, the overarching goal of this thesis is to improve the use of process data to assess and support collaborative problem-solving in the context of collaborative diagnostic reasoning in agent-based simulations (see Figure 2).

Figure 2

Aims of the Thesis and Corresponding Papers



To this end, this thesis compromises three papers with different focuses on the use of process data. The first paper, a theoretical perspective paper, will take a meta-perspective and review recent developments in the use of process data through technology-based assessment to generate new knowledge, improve learning and instruction, and provide actionable advice to policy stakeholders. Building on these considerations, two empirical studies are presented to illustrate how process data can be used for theoretical advancements and instructional improvement. The first empirical paper validates the CDR-M using process data. The second empirical paper then demonstrates how the combination of process data and theory can be used to predict outcomes that can inform learning and instruction in simulation-based learning of collaborative diagnostic reasoning. By presenting these two empirical contributions, which build on the views presented in the first paper, this thesis aims to shed light on new developments in the assessment and support of collaborative problem-solving skills and how the full potential of process data can be used not only to gain deeper insights and better theories about these skills, but also to use these data sources to support learning and instruction, thus helping to close a research-practice-policy gap.

1.5.1 Research Question and Outline of Paper 1

Over the past two decades, large-scale assessments in education have shifted from traditional paper-and-pencil formats to innovative technology-based assessments. This shift has enabled the collection and analysis of process data, which capture the steps and actions that lead to responses from participants. The first paper, a theoretical perspective paper, focuses on *how process data can bridge a gap between research, practice, and policy.* The paper is theoretically

grounded in the idea that interactive tasks and process data provide a richer understanding of learner behavior than traditional outcome-based assessments (see 1.2.3). By analyzing developments in large-scale assessment over the past two decades, the paper outlines the challenges and opportunities associated with leveraging process data to improve both educational research and policymaking.

The paper aims at identifying the potentials and challenges of using process data in educational settings, especially in large-scale assessments. Therefore, the paper adopts a meta-perspective, analyzing the impact of interactive tasks in large-scale assessments and emphasizing the need to move beyond task-specific findings by linking process data to theoretical constructs, which can enhance the generalizability of research findings.

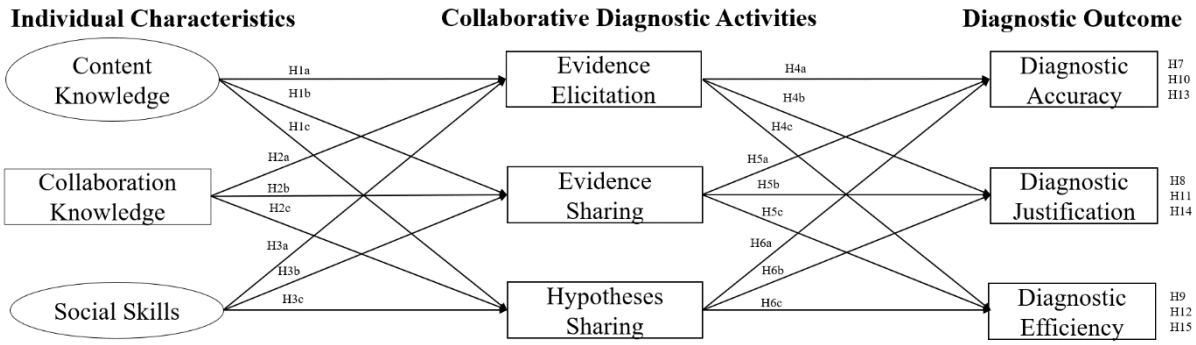
According to the CRediT statement, my contribution to the first paper was along the conceptualizing as well as writing parts of the original draft and reviewing and editing the entire paper. As this was not an empirical paper, most other aspects of the statement do not apply here. I contributed important theoretical ideas to the paper and was also involved in revisions during peer review and proofreading.

1.5.2 Research Question and Outline of Paper 2

The second paper aims at identifying how process data can facilitate theoretical advancements, particularly in the context of validating theoretical models in educational research, by addressing the research question of *the extent to which the relations in the CDR-M are applicable across studies.* Therefore, it tests the CDR-M (see 1.3.1), which posits that effective collaborative problem-solving in knowledge-rich domains, such as medical diagnosing, requires an interplay of individual characteristics, collaborative diagnostic activities enacted with high-quality, and successful diagnostic outcomes. To this end, I have derived a model from the postulated relations of the CDR-M (see Figure 3).

Figure 3

Hypotheses of Paper 2



I expect individual characteristics to be positively related to collaborative diagnostic activities (hypotheses 1-3), and collaborative diagnostic activities to be positively related to diagnostic outcome (hypotheses 4-6). In addition, I expect that the relations between individual characteristics and diagnostic outcomes to be partially mediated by collaborative diagnostic activities (Hypotheses 7-15).

A structural equation model is used to examine the relations between individual characteristics (content knowledge, collaboration knowledge, and social skills; see Appendix 8.2), collaborative diagnostic activities (evidence elicitation, evidence sharing, and hypotheses sharing; see Appendix 8.3), and diagnostic outcomes (accuracy, justification, and efficiency; see Appendix 8.4). Therefore, data from three studies involving 504 intermediate medical students working on the CoSiMed simulation (see 1.3.3) are analyzed. The use of agent-based simulations ensures controlled collaboration settings that simulate real-world diagnostic tasks. The study uses process data to empirically test and refine the CDR-M, demonstrating how process data can be used to generate new insights and advance theoretical frameworks in education.

According to the CRediT statement, my contribution to the second paper was along its conceptualization, methodology, validation, formal analysis, investigation, data curation, and writing the original draft. I was responsible for all major steps in the publication of this paper, from generating the research idea, performing the analyses, writing the paper, to revising during peer review and proofreading.

1.5.3 Research Question and Outline of Paper 3

The third paper explores how process data can inform learning and instruction by predicting learners' need for additional support. Specifically, it addresses the research question of *the extent to which theoretically derived process indicators are suitable for predicting learners' diagnostic accuracy in the context of simulation-based learning of collaborative diagnostic reasoning.* Thus, the third paper focuses on improving simulation-based learning by predicting diagnostic accuracy in collaborative diagnostic reasoning using process data. The study is theoretically grounded in the CDR-M (see 1.3.1), which integrates individual diagnostic processes and collaborative activities as described by Radkowitsch et al. (2020). Key collaborative diagnostic activities such as evidence elicitation, evidence sharing, and hypotheses sharing are identified as critical for accurate diagnostic outcomes.

Methodologically, the study uses a random forest classification model to predict diagnostic accuracy using process indicators derived from the CDR-M. It analyzes log-file data from five patient cases in the CoSiMed simulation (see 1.3.3) depicting the collaboration between a learner in the role of an internist interaction with an agent-based radiologist. The performance

of the model is evaluated in terms of classification accuracy, sensitivity, and specificity, with the goal of developing a reliable predictive tool for adaptive learning interventions.

According to the CRediT statement, my contribution to the paper was along its conceptualization, methodology, validation, formal analysis, investigation, data curation, and writing the original draft. I was responsible for all major steps in the publication of this paper, from conceiving the research idea, performing the analyses, writing the paper, to revising during peer review and proofreading.

PAPER 1

20 Years of Interactive Tasks in Large-Scale Assessments: Process Data as a way Towards Sustainable Change?

2

Matthias Stadler * Laura Brandl * Samuel Greiff

Reference: 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

Received: 10 March 2022 | Revised: 19 June 2023 | Accepted: 21 June 2023

DOI: 10.1111/jcal.12847

ORIGINAL ARTICLE

Journal of Computer Assisted Learning WILEY

20 years of interactive tasks in large-scale assessments: Process data as a way towards sustainable change?

Matthias Stadler[1,2] | **Laura Brandl**[1] | **Samuel Greiff**[3]

[1]Department of Psychology, Ludwig Maximilians University Munich, Munich, Germany

[2]Institut für Didaktik und Ausbildungsforschung in der Medizin, LMU Klinikum, Ludwig-Maximilians-Universität München, Munich, Germany

[3]Department of Behavioral and Cognitive Sciences, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Correspondence
Matthias Stadler, Klinikum der Ludwig-Maximilians-Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, Pettenkoferstr 8a, München 80336, Germany.
Email: matthias.stadler@med.uni-muenchen.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: COSIMA; DFG-Forschungsgruppe 2385

Abstract

Background: Over the last 20 years, educational large-scale assessments have undergone dramatic changes moving away from simple paper-pencil assessments to innovative, technology-based assessments. This comprehensive switch has led to some rather technical improvements such as identifying early guessing or improving standardization.

Objectives: At the same time, process data on student interaction with items has been shown to carry value for obtaining, reporting, and interpreting additional results on student skills in international comparisons. In fact, on the basis of innovative simulated assessment environments, news about student rankings, under- and overperforming countries, and novel ideas on how to improve educational systems are prominently featured in the media. At the same time, few of these efforts have been used in a sustainable way to create new knowledge (i.e., on a scientific level), to improve learning and instruction (i.e., on a practical level), and to provide actionable advice to political stakeholders (i.e., on a policy level).

Methods: This paper will adopt a meta-perspective and discuss recent and current developments with a focus on these three perspectives. There will be a particular emphasis on new assessment environments that have been recently employed in large-scale assessments.

Results and Conclusions: Most findings remain very task specific. We propose a necessary steps that need to be taken in order to yield sustainable change from analysing process data on all three levels.

Implications: New technologies might be capable of contributing to the research-policy-practitioner gap when it comes to utilizing the results from large-scale assessments to increase the quality of education around the globe but this will require a more systematic approach towards researching them.

KEYWORDS
large-scale assessments, process data, replication, research-policy-practitioner gap, technology-based education

1 | INTRODUCTION

Over the last 20 years, educational large-scale assessments have undergone dramatic changes moving away from simple paper-pencil assessments to innovative, technology-based assessments (von Davier et al., 2019). One example of the emergence of technology-based large-scale assessments is the Programme for International Student Assessment (PISA). This program, arguably the most extensive international educational assessment program, started to partially collect their data through technology-based assessment of science literacy back in 2006 (OECD, 2010). In 2009 PISA already administered another core competence using technology (digital reading assessment; OECD, 2012). This was extended in 2012 for the third core competence (mathematical literacy) plus adding a technology-based problem-solving assessment (OECD, 2013). In 2015, technology-based assessment was the primary mode of assessment in PISA (OECD, 2017). One reason was the inability to design authentic, interactive, and dynamic tasks for 21st-century skills with traditional paper-pencil formats (OECD, 2010). Technology-based assessments make the use of multimedia, simulations, interactive tasks, and virtual reality possible (Goldhammer et al., 2020). In addition to allowing for the operationalization of previously unobtainable competencies, using technology-based assessments allows for continuous measurement of the response process (i.e., process data), instead of only discrete states of responses depicted through the answers given to a task (i.e., product data; Thille et al., 2014).

This comprehensive switch from paper-pencil assessments to technology-based assessments has led to some rather technical improvements such as identifying early guessing (e.g., Kong et al., 2007) or improving standardization of assessment and scoring (e.g., Goldhammer et al., 2020). At the same time, process data on student interaction with items have been shown to carry value for obtaining, reporting, and interpreting additional results on student skills in international comparisons (e.g., Reis Costa et al., 2021; Xiao et al., 2021). Process data was used to relate behaviour to cognitive processes (Greiff et al., 2016), to validate score interpretations (Kane & Mislevy, 2017), and led to a better theoretical understanding of the construct under investigation (Goldhammer et al., 2017; Goldhammer & Zehner, 2017).

However, few of these efforts have been used sustainably to decrease global inequalities, and realize universal quality education (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2015) by creating new knowledge, improving learning and instruction, and providing actionable advice to political stakeholders (Dawson et al., 2019). This paper will adopt a meta-perspective and discuss recent and current developments focusing on these three perspectives. There will be a particular emphasis on new assessment environments that have been recently employed in large-scale assessments and how they might contribute to the research-policy-practitioner gap when it comes to utilizing the results from large-scale assessments to increase the quality of education around the globe.

2 | INTERACTIVE TASKS IN LARGE-SCALE ASSESSMENT

2.1 | New types of assessment

One of the driving forces behind the fast and comprehensive switch from paper-pencil assessments to technology-based assessments in international large-scale assessments has been the need to assess so-called 21st-century skills (Care et al., 2012). These 21st-century skills encompass a set of skills deemed critically important to student success in today's world, particularly as students move on to college, the workforce, and adult life, such as solving complex problems individually and collaboratively or possessing the media literacy to utilize and critically evaluate digital sources of information. Assessing these competencies requires assessment tools that respond to the test-taskers' inputs to allow for adequately complex and realistic tasks. Unlike conventional tasks (such as multiple-choice questions), these interactive tasks change, while the test-taker is trying to solve them, providing feedback to interventions or new information (Stadler et al., 2015). For instance, collaborative problem-solving tasks would hardly be valid if there was no interaction between the test-takers and the collaboration partners (Stadler, Herborn, et al., 2020). Likewise, assessments of hyper-text reading (reading and understanding digital text organized in a non-linear hypertext format) need to allow the test-takers to choose what information they want to read actively and in what order (Hahnel et al., 2023).

All of these new forms of assessment share that the interaction between test-takers and the assessment are expressed in observable actions (e.g., mouse clicks, eye-movements, keyboard inserts). Researchers are, thus, no longer limited to measuring the final outcome of an assessment (i.e., product data) but can also investigate the steps and actions resulting in the specific outcome through analyzes of test-taking behaviours (i.e., process data; Greiff et al., 2016; He et al., 2021).

2.2 | Process data and sequence data

In contrast to product data, process data and sequence data is seen as empirical information depicting behaviour that leads to the measured outcome (Goldhammer & Zehner, 2017). Typical process data are response times or the number of actions taken, whereas sequence data, as a special form of process data, describes the qualitative action sequences that lead to a specific result (Pohl et al., 2021; von Davier et al., 2019). Sequence data hence includes timing data, adding a quantitative dimension. Analysing process data and sequence data instead of only product data allows insights into the process leading to the eventual outcome. Researchers have already used process data to answer research questions as diverse as the detection of early guessing behaviour (e.g., Kong et al., 2007), validation of product data (e.g., Kane & Mislevy, 2017), early identification of students at risk to show inadequate performance (e.g., Wolff et al., 2013), analyses of incorrect responses and reasons (e.g., Ulitzsch et al., 2021) and a better theoretical understanding of the construct under investigation

(Goldhammer et al., 2017; Goldhammer & Zehner, 2017). Accordingly Pohl et al. (2021) argue that test-taking behaviour is not a nuisance factor that may confound measurement, but an aspect that provides important information on how examinees approach tasks, which is relevant for real-life outcomes.

Regarding the use of sequence data, Greiff et al. (2018) reported that students might show similar overall performance and yet can be distinguished according to their strategic behaviours in the tasks. These results indicate that process indicators depict individual differences in the ability that are not necessarily depicted in product data. This interpretation was further corroborated on laboratory data by Stadler, Hofer, and Greiff (2020), who found that participants solving a set of complex problem-solving tasks systematically differed in both time-on-task and number of clicks despite having reached the same outcome. This difference in behaviour was systematic and represented differences in ability as indicated by significant relations to an external criterion (participants' GPA). Moreover, the differences in behaviour could be explained by adjusting their effects on participants' GPA for individual differences in general problem-solving ability, which reduced them to negligible levels. He and von Davier (2016) used sequence data from the Programme for the International Assessment of Adult Competencies (PIAAC) studying how action sequences from problem-solving tasks are related to task performance finding several distinct action sequences that were related to correct responses (such as actions related to using software-tools).

While there is a surge of interest among researchers in harnessing process data, this rich resource's full utilization through dedicated analyses remains in its embryonic stages (Stadler et al., 2019). Significant improvements have been made in employing process data to enhance scoring accuracy and reporting in educational large-scale assessments (Pohl et al., 2021). However, the real value of integrating interactive tasks into these programs lies in their unique ability to capture action sequences that facilitate an exploration of the underlying reasons for students' success and failure (von Davier et al., 2019).

These interactive tasks provide a distinctive opportunity to contribute to the development of more sophisticated models of student cognition. By yielding detailed sequence data, we gain a more granular understanding of how students approach and navigate through different tasks. This allows us to observe the evolution of their problem-solving strategies over time, providing empirical evidence that can validate or challenge existing cognitive theories. Such insights can then directly inform the design of more nuanced, targeted instructional methods and learning materials, thus enriching the teaching-learning process.

Despite the evident value, comprehensive analyses employing this resource are scarce, often restricted to single or a few selected items with little common theoretical underpinning and minimal attempts at replicating findings. In the second part of this paper, we will discuss how these missed opportunities have resulted in a lack of sustainable change in education at the scientific, practical, and political levels. As we move forward, it is essential to shift our focus from merely improving scoring and extending reporting, towards fully exploiting the potential of interactive tasks in generating refined cognitive models that can transform educational practices and theories.

3 | ISSUES LIMITING SUSTAINABLE CHANGE

3.1 | Scientific level

A substantial obstacle preventing sustainable change, brought about by the use of interactive tasks in large-scale assessments at the scientific level, is the strong task-specificity of findings. Replications, already a rarity in educational research (Makel & Plucker, 2014), are virtually non-existent when it comes to sequence data from interactive tasks (c.f., Brooks et al., 2015 for a positive example). Several reasons may account for this, such as the relative infancy of the field. However, we contend that the lack of generalizability of findings and a missing relation between data and theory strongly limit the replicability of research on sequence data from interactive tasks in educational large-scale assessments, and thus, its scientific value.

Interactive tasks are often highly complex, involving multiple interrelated variables, usually embedded in a certain semantic context. These contexts only permit specific interactions between them and the test-takers. Thus, directly relating specific interactions with one item to interactions with other items becomes a challenging task, especially if these items do not even allow for these particular interactions. Many studies interested in comparing processes across items are therefore forced to rely on relatively low-level metric analyses (Ihantola et al., 2015), such as relating time-on-task or the number of interactions to the latent construct being assessed (e.g., Greiff et al., 2016).

Drawing on Mislevy's (2019) view, we suggest an explicit differentiation between low-level features and higher-level features. Low-level features, such as time-on-task or the number of interactions, can be more idiosyncratic and may not convey the same meaning across different tasks (Stadler, Radkowitsch, et al., 2020). On the other hand, higher-level features, derived from low-level ones, can present robust evidence that is pertinent across different tasks, thereby providing the possibility of conceptual replication even when items differ between studies.

An inventive solution to this predicament was offered by He et al. (2021), who related the performance on several PIAAC tasks to the distance of the observed behaviour sequence from an ex-ante defined ideal sequence. This approach allows for generalizing findings across various tasks that do not need to be similar as long as it is possible to determine an ideal sequence of actions. However, these findings would still exist within a theoretical vacuum as long as the ideal sequence is not linked to a theory-based definition of the construct.

As an application of the approach of distinguishing low-level and higher-level features, (Brandl et al., 2021) coded the interactions between learners and a training simulation for medical diagnoses based on theoretically defined diagnostic activities (Fischer et al., 2014). This focus on higher-level features allowed the study to move beyond task-specificity, thereby enabling the generalization of findings across various diagnostic tasks. Aggregating the process data in this way allows to train machine-learning algorithms to predict successful diagnoses in various diagnostic tasks. This study makes it

apparent how relating process data to established theoretical concepts can make the findings generalizable. The diagnostic activities used to code the interactions are not specific to any individual task, and the same method could be applied to any diagnostic training simulation regardless of context. Lotz et al. (2017) demonstrated how intelligence relates to individual differences in interaction frequency and quality changes in a computer-based problem-solving task. The authors find that, while all test-takers improve their test-taking behaviour across tasks on average, individual differences in intelligence predict the speed and range of this improvement. This example illustrates how rather basic process information can still be related to theoretically defined latent constructs.

In conclusion, to truly advance scientific knowledge through process data analyses, it's essential to generalize findings from specific items and link them to established constructs (see also Kroehne & Goldhammer, 2018). This requires the capacity for replication, systematic testing of theories, and the understanding of the difference between low-level and higher-level features. Recognizing this differentiation sets a crucial theoretical consideration for the level of abstraction in analysing process data, paving the way for more substantial scientific progress. Additionally, as underscored by (Goldhammer et al., 2021), there is a pressing need to validate the interpretation of measures based on process data. Even when a theoretical link between data and theory is postulated, this link necessitates substantiation with theoretical and empirical arguments to ensure its validity. Therefore, the integration of theoretical considerations, the differentiation of low-level and higher-level features, and the validation of interpretations collectively form the pillars of more rigorous and impactful research in this domain.

3.2 | Practical level

Despite the scientific challenges described above, there are many high-quality studies on the use of process data in educational large-scale assessments, demonstrating the benefit of modelling new data sources and incorporating process data in the statistical modelling of multiple possible assessment data (He & von Davier, 2016; Jiang et al., 2021; Pohl et al., 2021; von Davier et al., 2019). Process data can help validate and facilitate measuring response accuracy and provide supplementary information in understanding test-takers' behaviours, the reasons for missing data, and links with motivation studies.

However, with the evolution of educational large-scale assessment from a paper-based technology to an electronic one, the focus of these assessments has evolved, too (Bennett, 2015). Over the past several decades, the most common use of educational assessment has been for institutional purposes such as state school accountability. Accordingly, lots of research on the use of process data has concentrated on this use of assessment. However, in recent years, the value of assessment as a feedback tool informing individual learning (formative assessment) has been realized (e.g., Chudowsky & Pellegrino, 2003; van der Kleij et al., 2015). Whereas testing to serve institutional purposes may not diminish in absolute terms, there is reason to believe it will diminish in relative terms as assessment to serve individual learning purposes becomes more frequent. The increasing prominence of formative assessment is being driven by many factors, including advances in measurement and data science and the emergence of electronic learning environments.

Obviously, international large-scale assessments are primarily designed to facilitate group-based assessments and comparisons across large populations, not individuals (von Davier et al., 2019). Nonetheless, they can and should support learning (Chudowsky & Pellegrino, 2003). Especially interactive tasks inherently offer a type of feedback to the test-takers through the evolution of the task in response to their interactions (Greiff et al., 2016). This feedback, as we conceive it here, does not correspond to traditional, evaluative feedback, but instead refers to the changing state of the task according to the decisions made by the test-takers. Essentially, the task environment responds and adapts based on the actions of the test-takers, thus providing them with an implicit form of feedback about the consequences of their actions within the task scenario. This results in learning opportunities that can be used more or less efficiently, which needs to be considered when using these tasks as a means of standardized testing. Rather than trying to reduce these learning opportunities by limiting the tasks responsiveness, it may be beneficial to assess individual learning rather than the mere ability to solve the task. To benefit individually from an assessment situation, especially from a complex interactive task, learners require individualized scaffolding (e.g., Azevedo et al., 2004). For example, educators could use process-data from computer-based assessments to differentiate specific behaviours in students and use this information to provide individualized support (e.g., Li et al., 2020). Accordingly, process data analyses have long been considered a promising tool to detect a need for scaffolding and provide individualized support, yet most of this potential remains essentially untapped in day-to-day teaching practice (Bakharia et al., 2016). Most previous studies have drawn on historical data to identify patterns in students' process data and related these patterns to academic performance, retention, or other institutional outcomes. Utilizing process data for individual learning purposes requires understanding the pedagogical context that influences student activities and how identifying patterns in students' learning behaviours can help influence and contribute to more positive learning experiences (Gašević et al., 2016; Lockyer et al., 2013). An essential next step in advancing the practical relevance of new assessment technologies in educational large-scale assessments will, therefore, be to align the design of assessment with learning design to use assessments not only for institutional information but also as a source of individual learning.

3.3 | Policy level

Finally, modern educational large-scale assessments are an increasingly important part of the educational research and policy landscape internationally (Rutkowski et al., 2013). For instance, PISA claims to have become “the world's premier yardstick for evaluating the quality,

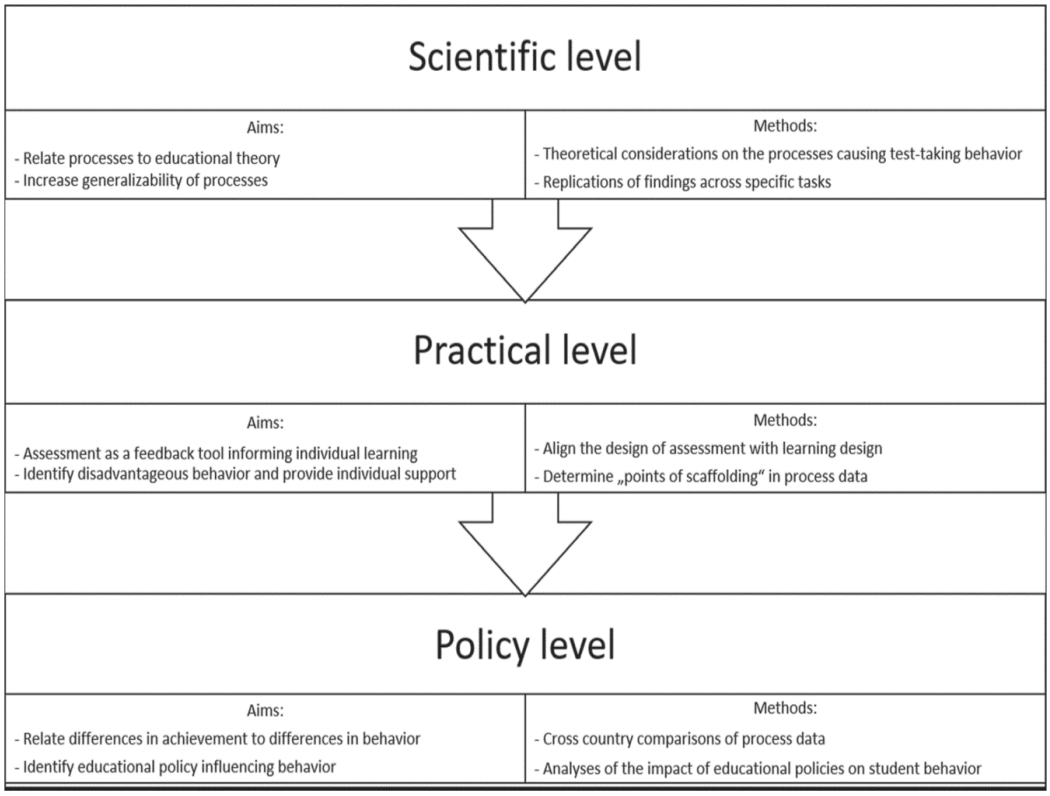


FIGURE 1 The use of process data in educational large-scale assessment for sustainable change.

equity and efficiency of school systems" (OECD, 2016, p. 2). In fact, despite many criticisms and potential issues with comparability across countries (see e.g., Winthrop & Simons, 2013) as well as methodological constraints (e.g., Rutkowski & Rutkowski, 2016), there are several examples of how results from educational large-scale assessments have been converted into educational policy such as the formation, expansion and improvements to national assessment and evaluation systems, the revision of curriculum standards, often to include and emphasize PISA-like competencies, or promoting equity through school financing (Breakspear, 2012; Wagemaker, 2014).

However, in international large-scale assessments, the focus lies mainly on achievement test scores as a measure of competence, which makes sense when discussing performance on an aggregated level for quality monitoring (Skedsmo & Huber, 2017). Unfortunately, the idea that the criterion of competence is what someone can do often downplays the importance of how the person arrives at this competence (Havnes & Prøitz, 2016; Oliveri & Davier, 2014). In other words, merely relating aggregated sum scores to differences in educational systems such as the number of all-day schools or integrative schools without a very good understanding of the behaviour on which the test values are based seems problematic (e.g., Gür et al., 2012; Kuhlmann & Tillmann, 2009) and is unlikely to yield sustainable changes (Pohl et al., 2021).

Analyses of process data can help provide this understanding. For instance, Greiff et al. (2016), analysed log-file data of a complex problem-solving tasks for students from 44 countries and economies. The authors find that there were different levels of non-mastery that ranged from applying no systematic strategic behaviour to actually applying the appropriate strategy but still failing to solve the task. On the backdrop of these results, they discuss implications and future potentials of log-file analyses in educational large-scale assessments for researchers, teachers, and policy makers. This study demonstrates how for policy makers, interesting comparisons between educational systems might emerge from the relation between actual behaviour and overall proficiency. In the PISA 2012 cycle, for instance, Polish students performed reasonably well in mathematics (518 points in the international comparison of the PISA scale), science (526 points), and reading (518 points) but performed considerably worse in complex problem-solving compared with other countries or economies (481 points). Log-file analyses revealed how this performance drop could be explained by (a lack of) specific actions, for instance, because Polish students never learned the principle of isolated variation or because they were too reluctant to explore a problem situation comprehensively. This provides interesting starting points for policy decisions and educational priorities (see Greiff et al., 2015).

Questions such as what exactly it is that students do better in one country compared with another may provide insights into how teaching

practices foster or neglect certain behaviours but need to consider cultural differences in learning and teaching (Huang et al., 2016).

4 | CONCLUSION

In summary, we posit that educational large-scale assessments, particularly through their evolution from simple paper-pencil tests to innovative technology-based and simulated environments, hold tremendous potential to advance research, educational practice, and policy-making. Despite this potential, the sustainable utilization of the rich information these assessments provide has been hindered, impacting their potential to enhance global education quality. As illustrated in Figure 1, there are necessary changes to be undertaken at the scientific level in how we analyse process data to foster sustainable changes at the practical and policy levels. Primarily, linking process data to educational theory is crucial for enhancing the generalizability of our findings. This link not only enables the utilization of assessment results as individual learning feedback tools but also allows the identification of disadvantageous behaviours, paving the way for targeted individual support.

To achieve this, the alignment of assessment design with learning theories is paramount. The process data emanating from such theoretically grounded and practically meaningful assessments can then elucidate achievement disparities across countries or educational systems. Policy predicated on such robust data can have a lasting, sustainable impact on students' education. This exploration underscores a fundamental need for further research dedicated to the sustainable, theory-driven utilization of process data from interactive tasks in large-scale assessments. Our aspiration is that this research will lead to systemic changes that bridge the gap between research, practice, and policy in education, ultimately contributing to the quality of education worldwide.

ACKNOWLEDGEMENTS

The contributions by Matthias Stadler and Laura Brandl was supported by a grant from the Deutsche Forschungsgesellschaft DFG (COSIMA; DFG-Forschungsgruppe 2385). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/jcal.12847.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study

ORCID

Matthias Stadler https://orcid.org/0000-0001-8241-8723

REFERENCES

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology*, *29*(3), 344–370. https://doi.org/10.1016/j.cedpsych.2003.09.002

Bakharia, A., Corrin, L., de Barba, P., Kennedy, G., Gašević, D., Mulder, R., Williams, D., Dawson, S., & Lockyer, L. (2016). A conceptual framework linking learning design with learning analytics. In D. Gašević, G. Lynch, S. Dawson, H. Drachsler, & C. Penstein Rosé (Eds.), *Proceedings of the sixth international conference on Learning Analytics & Knowledge - LAK '16* (pp. 329–338). ACM Press. https://doi.org/10.1145/2883851.2883944

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370–407. https://doi.org/10.3102/0091732X14554179

Brandl, L., Richters, C., Radkowitsch, A., Obersteiner, A., Fischer, M. R., Schmidmaier, R., Fischer, F., & Stadler, M. (2021). Simulation-based learning of complex skills: Predicting performance with theoretically derived process features. *Psychological Test and Assessment Modeling*, *63*(4), 542–560 https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2021-4/PTAM__4-2021_6_kor.pdf

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance* (OECD Education Working Papers No. 71). https://doi.org/10.1787/5k9fdfqffr28-en

Brooks, C., Baker, R., & Andres, J. M. L. (2015). Infrastructure for replication in learning analytics. In *Nature*. Advance Online Publicatio. https://doi.org/10.1038/nature.2015.17433

Care, E., Griffin, P., & McGaw, B. (2012). *Assessment and teaching of 21st century skills*. Springer.

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: What will it take? *Theory Into Practice*, *42*(1), 75–83. https://doi.org/10.1353/tip.2003.0002

Dawson, S., Joksimovic, S., Poquet, O., & Siemens, G. (2019). Increasing the impact of learning analytics. In *Proceedings of the 9th international conference on Learning Analytics & Knowledge* (pp. 446–455). ACM. https://doi.org/10.1145/3303772.3303784

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, 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

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

Goldhammer, F., Hahnel, C., & Kroehne, U. (2020). Analysing log file data from PIAAC. In D. B. Maehler & B. Rammstedt (Eds.), *Methodology of educational measurement and assessment. Large-scale cognitive assessment* (pp. 239–269). Springer International Publishing. https://doi.org/10.1007/978-3-030-47515-4_10

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: On validating the interpretation of process indicators based on log data. *Large-scale Assessments in Education*, *9*(1), 1–25. https://doi.org/10.1186/s40536-021-00113-5

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.), *Methodology of educational measurement and assessment. Competence assessment in education* (pp. 407–425). Springer International Publishing. https://doi.org/10.1007/978-3-319-50030-0_24

Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3–4), 128–132. https://doi.org/10.1080/15366367.2017.1411651

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem

environments: A latent class approach. *Computers & Education*, *126*, 248–263. https://doi.org/10.1016/j.compedu.2018.07.013

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., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Gür, B. S., Çelik, Z., & Özoğlu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy*, *27*(1), 1–21. https://doi.org/10.1080/02680939.2011.595509

Hahnel, C., Ramalingam, D., Kroehne, U., & Goldhammer, F. (2023). Patterns of reading behaviour in digital hypertext environments. *Journal of Computer Assisted Learning*, *39*(3), 737–750. https://doi.org/10.1111/jcal.12709

Havnes, A., & Prøitz, T. S. (2016). Why use learning outcomes in higher education? Exploring the grounds for academic resistance and reclaiming the value of unexpected learning. *Educational Assessment, Evaluation and Accountability*, *28*(3), 205–223. https://doi.org/10.1007/s11092-016-9243-z

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, *166*, 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-grams. 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* (pp. 750–777). Information Science Reference, an imprint of IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch029

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: Language, curriculum or culture. *Educational Psychology*, *36*(2), 378–390. https://doi.org/10.1080/01443410.2014.946890

Ihantola, P., Vihavainen, A., Ahadi, A., Butler, M., Börstler, J., Edwards, S. H., Isohanni, E., Korhonen, A., Petersen, A., Rivers, K., Rubio, M. Á., Sheard, J., Skupas, B., Spacco, J., Szabo, C., & Toll, D. (2015). Educational data mining and learning analytics in programming. In N. Ragonis & P. Kinnunen (Eds.), *Proceedings of the 2015 ITiCSE on working group reports* (pp. 41–63). ACM. https://doi.org/10.1145/2858796.2858798

Jiang, Y., Gong, T., Saldivia, L. E., Cayton-Hodges, G., & Agard, C. (2021). Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment. *Large-scale Assessments in Education*, *9*(1), 1–31. https://doi.org/10.1186/s40536-021-00095-4

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *The NCME applications of educational measurement and assessment book series. Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 11–24). Routledge Taylor & Francis Group. https://doi.org/10.4324/9781315708591-2

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*(2), 527–563. https://doi.org/10.1007/s41237-018-0063-y

Kuhlmann, C., & Tillmann, K.-J. (2009). Mehr Ganztagsschulen als Konsequenz aus PISA? Bildungspolitische Diskurse und Entwicklungen in den Jahren 2000 bis 2003. In F.-U. Kolbe, S. Reh, T.-S. Idel, B. Fritzsche, & K. Rabenstein (Eds.), *Ganztagsschule als symbolische Konstruktion* (pp. 23–45). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91354-4_2

Li, H., Kim, M. K., & Xiong, Y. (2020). Individual learning vs. interactive learning: A cognitive diagnostic analysis of MOOC Students' learning behaviors. *American Journal of Distance Education*, *34*(2), 121–136. https://doi.org/10.1080/08923647.2019.1697027

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *American Behavioral Scientist*, *57*(10), 1439–1459. https://doi.org/10.1177/0002764213479367

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2017). Intelligence in action – Effective strategic behaviors while solving complex problems. *Intelligence*, *64*, 98–112. https://doi.org/10.1016/j.intell.2017.08.002

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513

Mislevy, R. J. (2019). Advances in measurement and cognition. *The Annals of the American Academy of Political and Social Science*, *683*(1), 164–182. https://doi.org/10.1177/0002716219843816

OECD. (2010). *PISA computer-based assessment of student skills in science*. OECD Publishing. https://doi.org/10.1787/9789264082038-en

OECD. (2012). *PISA 2009 technical report*. OECD Publishing. https://doi.org/10.1787/9789264167872-en

OECD. (2013). *PISA 2012 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/9789264190511-en

OECD. (2016). *PISA in focus* (Vol. 67). Organisation for Economic Co-Operation and Development (OECD). https://doi.org/10.1787/aa9237e6-en

OECD. (2017). *PISA 2015 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*(1), 1–21. https://doi.org/10.1080/15305058.2013.825265

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science (New York, N.Y.)*, *372*(6540), 338–340. https://doi.org/10.1126/science.abd3300

Reis Costa, D., Bolsinova, M., Tijmstra, J., & Andersson, B. (2021). Improving the precision of ability estimates using time-on-task variables: Insights from the PISA 2012 computer-based assessment of mathematics. *Frontiers in Psychology*, *12*, 579128. https://doi.org/10.3389/fpsyg.2021.579128

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252–257. https://doi.org/10.3102/0013189X16649961

Rutkowski, L., von Davier, M., & Rutkowski, D. (2013). *Handbook of international large-scale assessment*. Chapman and Hall/CRC. https://doi.org/10.1201/b16061

Skedsmo, G., & Huber, S. G. (2017). Policies and practices related to student assessment and learning outcomes—Combining different purposes and ideals. *Educational Assessment, Evaluation and Accountability*, *29*(3), 225–228. https://doi.org/10.1007/s11092-017-9268-y

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., 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., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education*, *157*, 103964. https://doi.org/10.1016/j.compedu.2020.103964

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. *Computers in Human Behavior*, *111*, 106442. https://doi.org/10.1016/j.chb.2020.106442

Stadler, M., Radkowitsch, A., Schmidmaier, R., Fischer, M. R., & Fischer, F. (2020). Take your time: Invariance of time- on-task in problem solving tasks across expertise levels. *Psychological Test and Assessment Modeling*, *65*(4), 517–525.

Thille, C., Kizilee, R. F., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The future of data–enriched assessment. *Research & Practice in Assessment*, *9*, 5–16.

Ulitzsch, E., He, Q., & Pohl, S. (2021). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*, *47*(1), 3–35. https://doi.org/10.3102/10769986211010467

United Nations Educational, Scientific and Cultural Organization. (2015). *Education 2030 Incheon declaration and framework for action: Towards inclusive and equitable quality education and lifelong learning for all*. https://unesdoc.unesco.org/ark:/48223/pf0000245656

van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 324–343. https://doi.org/10.1080/0969594X.2014.999024

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics*, *44*(6), 671–705. https://doi.org/10.3102/1076998619881789

Wagemaker, H. (2014). *International large-scale assessments: From research to policy. Handbook of international large-scale assessment* (pp. 11–36). Background, Technical Issues, and Methods of Data Analysis.

Winthrop, R., & Simons, K. A. (2013). Can international large-scale assessments inform a global learning goal? Insights from the learning metrics task force. *Research in Comparative and International Education*, *8*(3), 279–295. https://doi.org/10.2304/rcie.2013.8.3.279

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving retention. In D. Suthers, K. Verbert, E. Duval, & X. Ochoa (Eds.), *Proceedings of the third international conference on learning analytics and knowledge - LAK '13* (p. 145). ACM Press. https://doi.org/10.1145/2460296.2460324

Xiao, Y., He, Q., Veldkamp, B., & Liu, H. (2021). Exploring latent states of problem-solving competence using hidden Markov model on process data. *Journal of Computer Assisted Learning*, *37*(5), 1232–1247. https://doi.org/10.1111/jcal.12559

How to cite this article: 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*, 1–8. https://doi.org/10.1111/jcal.12847

PAPER 2

Collaborative Problem-Solving in Knowledge-Rich Domains: A Multistudy Structural Equation Model

3

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

Reference: Brandl, L., Stadler, M., Richters, C., Radkowitsch, A., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2024). Collaborative Problem-Solving in Knowledge-Rich Domains: A Multi-Study Structural Equation Model. International Journal of Computer-Supported Collaborative Learning, 19(3), 341–368. https://doi.org/10.1007/s11412-024-09425-4

Intern. J. Comput.-Support. Collab. Learn (2024) 19:341–368
https://doi.org/10.1007/s11412-024-09425-4

Collaborative Problem-Solving in Knowledge-Rich Domains: A Multi-Study Structural Equation Model

Laura Brandl[1] · Matthias Stadler[1,2] · Constanze Richters[1] · Anika Radkowitsch[3] · Martin R. Fischer[2] · Ralf Schmidmaier[4] · Frank Fischer[1]

Received: 18 September 2023 / Accepted: 9 May 2024 / Published online: 24 June 2024

Abstract

Collaborative skills are crucial in knowledge-rich domains, such as medical diagnosing. The Collaborative Diagnostic Reasoning (CDR) model emphasizes the importance of high-quality collaborative diagnostic activities (CDAs; e.g., evidence elicitation and sharing), influenced by content and collaboration knowledge as well as more general social skills, to achieve accurate, justified, and efficient diagnostic outcomes (Radkowitsch et al., 2022). However, it has not yet been empirically tested, and the relationships between individual characteristics, CDAs, and diagnostic outcomes remain largely unexplored. The aim of this study was to test the CDR model by analyzing data from three studies in a simulation-based environment and to better understand the construct and the processes involved ($N = 504$ intermediate medical students) using a structural equation model including indirect effects. We found various stable relationships between individual characteristics and CDAs, and between CDAs and diagnostic outcome, highlighting the multidimensional nature of CDR. While both content and collaboration knowledge were important for CDAs, none of the individual characteristics directly related to diagnostic outcome. The study suggests that CDAs are important factors in achieving successful diagnoses in collaborative contexts, particularly in simulation-based settings. CDAs are influenced by content and collaboration knowledge, highlighting the importance of understanding collaboration partners' knowledge. We propose revising the CDR model by assigning higher priority to collaboration knowledge compared with social skills, and dividing the CDAs into information elicitation and sharing, with sharing being more transactive. Training should focus on the development of CDAs to improve CDR skills.

Keywords Collaborative Problem-solving · Simulation-based Learning Environment · Diagnostic Activities · Diagnostic Reasoning · Medical Education

Introduction

Collaborative skills are highly relevant in many situations, ranging from computer-supported collaborative learning to collaborative problem-solving in professional practice (Fiore et al., 2018). While several broad collaborative problem-solving frameworks exist

Extended author information available on the last page of the article

(OECD, 2017), most of them are situated in knowledge-lean settings. However, one example of collaborative problem-solving of knowledge-rich domains is collaborative diagnostic reasoning (CDR; Radkowitsch et al., 2022)—which aligns closely with medical practice—as this is a prototypical knowledge-rich domain requiring high collaboration skills in daily practice. In daily professional practice, physicians from different specialties often need to collaborate with different subdisciplines to solve complex problems, such as diagnosing, that is, determining the causes of a patient's problem. Moreover, research in medical education and computer-supported collaborative learning suggests that the acquisition of medical knowledge and skills is significantly enhanced by collaborative problem-solving (Hautz et al., 2015; Koschmann et al., 1992). For problem-solving and learning, it is crucial that all relevant information (e.g., evidence and hypotheses) is elicited from and shared with the collaboration partner (Schmidt & Mamede, 2015). However, CDR is not unique to the medical field but also relevant in other domains, such as teacher education (Heitzmann et al., 2019).

The CDR model has been the basis of empirical studies and describes how individual characteristics and the diagnostic process are related to the diagnostic outcome. However, it has not yet been empirically tested, and the relationships between individual characteristics, diagnostic process, and diagnostic outcome remain mostly unexplored (Fink et al., 2023). The aim of this study is to test the CDR model by analyzing data from three studies with similar samples and tasks investigating CDR in a simulation-based environment. By undertaking these conceptual replications, we aspire to better understand the construct and the processes involved. As prior research has shown, collaboration needs to be performed at a high quality to achieve accurate problem solutions respectively learning outcomes (Pickal et al., 2023).

Collaborative Diagnostic Reasoning (CDR) Model

Diagnosing can be understood as the process of solving complex diagnostic problems through "goal-oriented collection and interpretation of case-specific or problem-specific information to reduce uncertainty" in decision-making through performing diagnostic activities at a high quality (Heitzmann et al., 2019, p. 4). To solve diagnostic problems, that is, to identify the causes of an undesired state, it is increasingly important to collaborate with experts from different fields, as these problems become too complex to be solved individually (Abele, 2018; Fiore et al., 2018). Collaboration provides advantages such as the division of labor, access to diverse perspectives and expertise, and enhanced solution quality through collaborative sharing of knowledge and skills (Graesser et al., 2018).

The CDR model is a theoretical model focusing on the diagnostic process in collaborative settings within knowledge-rich domains (Radkowitsch et al., 2022). The CDR model is based on scientific discovery as a dual-search model (SDDS; Klahr & Dunbar, 1988) and its further development by van Joolingen and Jong (1997). The SDDS model describes individual reasoning as the coordinated search through hypothetical evidence and hypotheses spaces and indicates that for successful reasoning it is important not only that high-quality cognitive activities within these spaces are performed but also that one is able to coordinate between them (Klahr & Dunbar, 1988). In the extended SDDS model (van Joolingen & Jong, 1997) focusing on learning in knowledge-rich domains, a learner hypothesis space was added including all the hypotheses one can search for without additional knowledge. Although Dunbar (1995) found that conceptual change occurs more often in groups than in individual work, emphasizing the importance of collaborative

processes in scientific thinking and knowledge construction, the SDDS model has hardly been systematically applied in computer-supported collaborative learning and collaborative problem-solving.

Thus, the CDR model builds upon these considerations and describes the relationship between individual characteristics, the diagnostic process, and the diagnostic outcome. As in the SDDS model we assume that CDR involves activities within an evidence and hypotheses space; however, unlike the SDDS in the CDR model, these spaces are understood as cognitive storages of information. Which aligns more to the extended dual search space model of scientific discovery learning (van Joolingen & Jong, 1997). In summary we assume that coordinating between evidence (data) and hypothesis (theory) is essential for successful diagnosing. Further, the CDR model is extended to not only individual but also collaborative cognitive activities and describes the interaction of epistemic activities (F. Fischer et al., 2014) and collaborative activities (Liu et al., 2016) to construct a shared problem representation (Rochelle & Teasley, 1995) and effectively collaborate. Thus, we define CDR as a set of skills for solving a complex problem collaboratively "by generating and evaluating evidence and hypotheses that can be shared with, elicited from, or negotiated among collaborators" (Radkowitsch et al., 2020, p. 2). The CDR model also makes assumptions about the factors necessary for successful CDR. First, we look at what successful CDR means, why people differ, and what the mediating processes are.

Diagnostic Outcome: Accuracy, Justification, and Efficiency

The primary outcome of diagnostic processes, such as CDR, is the accuracy of the given diagnosis, which indicates problem-solving performance or expertise (Boshuizen et al., 2020). However, competence in diagnostic reasoning, whether it is done individually or collaboratively, also includes justifying the diagnosis and reaching it effectively. This is why, in addition to diagnostic accuracy, diagnostic justification and diagnostic efficiency should also be considered as secondary outcomes of the diagnostic reasoning process (Chernikova et al., 2022; Daniel et al., 2019). Diagnostic justification makes the reasoning behind the decision transparent and understandable for others (Bauer et al., 2022). Good reasoning entails a justification including evidence, which supports the reasoning (Hitchcock, 2005). Diagnostic efficiency is related to how much time and effort is needed to reach the correct diagnosis; this is important for CDR, as diagnosticians in practice are usually under time pressure (Braun et al., 2017). Both diagnostic justification and diagnostic efficiency are thus indicators of a structured and high-quality reasoning process. So, while in many studies, the focus of assessments regarding diagnostic reasoning is on the accuracy of the given diagnosis (Daniel et al., 2019), the CDR model considers all three facets of the diagnostic outcome as relevant factors.

Individual Characteristics: Knowledge and Social Skills

Research has shown that content knowledge, social skills, and, in particular, collaboration knowledge are important prerequisites for, and outcomes of, computer-supported collaborative learning (Jeong et al., 2019; Vogel et al., 2017). CDR has integrated these dependencies into its model structure. Thus, the CDR model assumes that people engaging in CDR differ with respect to their content knowledge, collaboration knowledge, and domain general social skills.

Content knowledge refers to conceptual and strategic knowledge in a specific domain (Förtsch et al., 2018). Conceptual knowledge encompasses factual understanding of domain-specific concepts and their interrelations. Strategic knowledge entails contextualized knowledge regarding problem-solving during the diagnostic process (Stark et al., 2011). During expertise development, large amounts of content knowledge are acquired and restructured through experience with problem-solving procedures and routines (Boshuizen et al., 2020). Research has repeatedly shown that having high conceptual and strategic knowledge is associated with the diagnostic outcome (e.g., Kiesewetter et al., 2020; cf. Fink et al., 2023).

In addition to content knowledge, the CDR model assumes that collaborators need collaboration knowledge. A key aspect of collaboration knowledge (i.e., being aware of knowledge distribution in the group; Noroozi et al., 2013) is the pooling and processing of non-shared information, as research shows that a lack of collaboration knowledge has a negative impact on information sharing, which in turn has a negative impact on performance (Stasser & Titus, 1985).

Finally, general social skills influence the CDR process. These skills mainly influence the collaborative aspect of collaborative problem-solving and less the problem-solving aspect (Graesser et al., 2018). Social skills are considered particularly important when collaboration knowledge is low (F. Fischer et al., 2013). CDR assumes that in particular the abilities to share and negotiate ideas, to coordinate, and to take the perspective are relevant for the diagnostic process and the diagnostic outcome (Radkowitsch et al., 2022; see also Liu et al., 2016, and Hesse et al., 2015).

Diagnostic Process: Collaborative Diagnostic Activities

The diagnostic process is thought to mediate the effect of the individual characteristics on the diagnostic outcome and is described in the CDR model using collaborative diagnostic activities (CDAs), such as evidence elicitation, evidence sharing, and hypotheses sharing (Heitzmann et al., 2019; Radkowitsch et al., 2022). One of the main functions of CDAs is to construct a shared problem representation (Rochelle & Teasley, 1995) by sharing and eliciting relevant information, as information may not be equally distributed among all collaborators initially. To perform these CDAs at a high quality, each collaborator needs to identify information relevant to be shared with the collaboration partner as well as information they need from the collaboration partner (OECD, 2017).

Evidence elicitation involves requesting information from a collaboration partner to access additional knowledge resources (Weinberger & Fischer, 2006). Evidence sharing and hypothesis sharing involve identifying the information needed by the collaborator to build a shared problem representation. This externalization of relevant information can be understood as the novelty aspect of transactivity (Vogel et al., 2023). However, challenges arise from a lack of relevant information due to deficient sharing, which can result from imprecise justification and insufficient clustering of information. In particular, research has shown that collaborators often lack essential information-sharing skills, such as identifying information relevant for others from available data, especially in the medical domain (Kiesewetter et al., 2017; Tschan et al., 2009).

It is crucial for the diagnostic outcome that all relevant evidence and hypotheses are elicited and shared for the specific collaborators (Tschan et al., 2009). However, diagnostic outcomes seem to be influenced more by the relevance and quality of the shared information than by their quantity (Kiesewetter et al., 2017; Tschan et al., 2009). In addition, recent research has shown that the diagnostic process is not only an embodiment of individual characteristics but also adds a unique contribution to diagnostic outcome (Fink et al., 2023). However, it remains difficult to assess and foster CDAs.

Collaboration in Knowledge-Rich Domains: Agent-Based Simulations

There are several challenges when it comes to modelling collaborative settings in knowledge-rich domains for both learning and research endeavors. First, many situations are not easily accessible, as they may be scarce (e.g., natural disasters) or too critical or overwhelming to be approached by novices (e.g., some medical procedures). In these cases, the use of simulation-based environments allows authentic situations approximating real-life diagnostic problems to be provided (Cook et al., 2013; Heitzmann et al., 2019). Further, the use of technology-enhanced simulations allows data from the ongoing CDR process to be collected in log files. This enables researchers to analyze process data without the need for additional assessments with dedicated tests. Analyzing process data instead of only product data (the assessment's outcome) permits insights into the problem-solving processes leading to the eventual outcome (e.g., Goldhammer et al., 2017). Second, when using human-to-human collaboration, the results of one individual are typically influenced by factors such as group composition or motivation of the collaboration partner (Radkowitsch et al., 2022). However, we understand CDR as an individual set of skills enabling collaboration, as indicated by the broader definition of collaborative problem-solving (OECD, 2017). Thus, the use of simulated agents as collaboration partners allows a standardized and controlled setting to be created that would otherwise be hard to establish in collaborations among humans (Rosen, 2015). There is initial research showing that performance in simulations using computerized agents is moderately related to collaborative skills in other operationalizations (Stadler & Herborn et al., 2020). Thus, computerized agents allow for enhanced control over the collaborative process without significantly diverging from human-to-human interaction (Graesser et al., 2018; Herborn et al., 2020). Third, in less controlled settings it is hard to ensure a specific process is taking place during collaborative problem-solving. For example, when using a human-to-human setting, it is possible that, even though we envision measuring or fostering a specific activity (i.e. hypotheses sharing), it is not performed by the student. Through using an agent-based simulated collaboration partner, we can ensure that all required processes are taking place while solving the problem (Rosen, 2015).

Summarizing, by fostering a consistent and controlled setting, simulated agents facilitate the accurate measurement and enhancement of collaborative problem-solving. Evidential support for the application of simulated agents spans a variety of contexts, including tutoring, collaborative learning, knowledge co-construction, and collaborative problem-solving itself, emphasizing their versatility and effectiveness in educational settings (Graesser et al., 2018; Rosen, 2015).

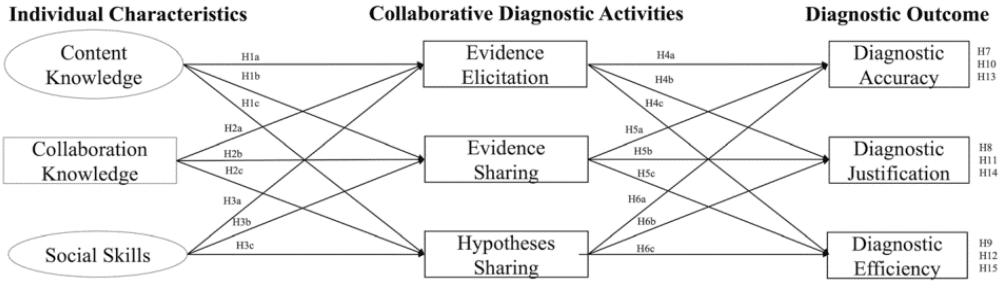


Fig. 1 Visualization of hypothesized relationships between individual characteristics, collaborative diagnostic activities, and diagnostic outcome

Research Question and Current Study

In computer-supported collaborative learning there has been the distinction between approaches addressing collaboration to learn and approaches focusing on learning to collaborate. Our study is best understood as addressing the second approach, learning to collaborate. We want to better understand CDR to be able to facilitate collaborative problem-solving skills in learners. Thus, in this paper, we examine what it takes to be able to collaborate in professional practice of knowledge-rich domains, such as medical diagnosing.

When solving diagnostic problems, such as diagnosing a patient, it is often necessary to collaborate with experts from different fields (Radkowitsch et al., 2022). In CDR, the diagnostic outcome depends on effectively eliciting and sharing relevant evidence and hypotheses among collaborators, who often lack information-sharing skills (Tschan et al., 2009). Thus, the CDR model emphasizes the importance of high-quality CDAs influenced by content and collaboration knowledge as well as social skills to achieve accurate, justified, and efficient diagnostic outcomes (Radkowitsch et al., 2022).

This study reviews the relationships postulated in CDR model across three studies to test them empirically and investigate the extent to which the relationships in the CDR model are applicable across studies. By addressing this research question, the current study contributes to a better understanding of the underlying processes in collaborative problem-solving.

We derived a model (Fig. 1) from the postulated relationships made by the CDR model. We assume that the individual characteristics are positively related to the CDAs (Hypotheses 1–3), as well as that the CDAs are positively related to the diagnostic outcome (Hypotheses 4–6). Further, we expect that the relationship between the individual characteristics and the diagnostic outcome is partially mediated by the CDAs (Hypotheses 7–15).

We used data from three studies with similar samples and tasks investigating CDR in an agent-based simulation in the medical domain. The studies can therefore be considered conceptual replication studies. Furthermore, we decided to use an agent-based simulation of a typical collaboration setting in diagnostic reasoning, namely the interdisciplinary collaboration between an internist and a radiologist (Radkowitsch et al., 2022).

Table 1 Sample description per study

Study	N	Gender	Age	Year of study
Study A	157	Male = 49 Female = 108	$M = 25.1$ $SD = 3.1$	$M = 5.3$ $SD = 0.9$
Study B	155	Male = 44 Female =111	$M = 25.3$ $SD = 3.0$	$M = 5.4$ $SD = 0.8$
Study C	192	Male = 62 Female = 130	$M = 23.3$ $SD = 3.4$	$M = 3.2$ $SD = 0.4$

N = sample size. M = mean, SD = standard deviation

Methods

Sample

To test the hypotheses, three studies were analyzed.[1] Study A was carried out in a laboratory setting in 2019 and included medical students in their third to sixth years. Study B included medical students in their fifth to sixth years. Data collection for this study was online due to the pandemic situation in 2020 and 2021. In both studies, participation was voluntary, and participants were paid 10 per hour. Study C was embedded as an online session in the curriculum of the third year of medical school in 2022. Participation was mandatory, but permission to use the data for research purposes was given voluntarily. All participants took part in only one of the three studies. All three studies received ethical approval from LMU Munich (approval numbers 18-261, 18-262 & 22-0436). For a sample description of each study, see Table 1. We would like to emphasize that none of the students were specializing in internal medicine, ensuring that the study results reflect the competencies of regular medical students without specialized expertise.

Procedure

Each of the three studies was organized in the same way, with participants first completing a pretest that included a prior knowledge test, socio-demographic questions, and questions about individual motivational-affective characteristics (e.g., social skills, interest, and motivation). Participants then moved on to the CDR simulation and worked on the patient case. The patient case was the same for studies B and C, but was different for study A. The complexity and difficulty of the patient case did not vary systematically between the patient cases.

Simulation and Task

In the CDR simulation, which is also used as a learning environment, the task was to take over the role of an internist and to collaborate with an agent-based radiologist to obtain further information by performing radiological examinations to diagnose fictitious patient

[1] Please note that the data employed in this study have been used in previous publications (e.g., Brandl et al., 2021; Radkowitsch, et al., 2021; Richters et al., 2022). However, the research question and the results reported in this study are completely unique to this study.

Table 2 Overview of the number of questions in the content knowledge test

Study	Conceptual knowledge in internal medicine	Conceptual knowledge in radiology	Strategic knowledge in internal medicine	Strategic knowledge in radiology
Study A	20	15	24 8 cases 3 questions per case	16 8 cases 2 questions per case
Study B	20	15	24 8 cases 3 questions per case	16 8 cases 2 questions per case
Study C	13	12	24 8 cases 3 questions per case	12 6 cases 2 questions per case

Table 3 Example items for each subscale for measuring social skills

Subscale	Item
Direct Measurement	I enjoy working with others.
Perspective taking	It is easy for me to put myself in the position of my collaboration partners.
Information sharing	When I collaborate with others, I purposefully share relevant information.
Negotiating	I can negotiate compromises when working with others.
Coordination	When I work with others, we have a clear common goal in mind.

cases with the chief symptom of fever. Medical experts from internal medicine, radiology, and general medicine constructed the patient cases. Each case was structured in the same way: by studying the medical record individually, then collaborating with an agent-based radiologist, and finally reporting the final diagnosis and its justification again individually. For a detailed description on the development and validation of the simulation, see Radkowitsch and colleagues (2020).

Before working within the simulation, participants were presented with an instruction for the simulated scenario and informed what they were to do with it. Then, we instructed participants how to access further information in the medical record by clicking on hyperlinks, as well as how they could use the toolbar to make notes for the later in the process. Furthermore, we acquainted the students with how they could request further information through collaborating with a radiologist.

During the collaboration with an agent-based radiologist, participants were asked to fill out request forms to obtain further evidence from radiological examinations needed to diagnose the patient case. To effectively collaborate with radiologists, it is crucial for internists to clearly communicate the type of evidence required to reduce uncertainty (referred to as "evidence elicitation") and share any relevant patient information such as signs, symptoms, and medical history (referred to as "evidence sharing") as well as suspected diagnoses under investigation (referred to as "hypotheses sharing") that may impact the radiologists' diagnostic process. Only when participants shared evidence and hypotheses appropriately for their requested examination did they receive a description and evaluation of the radiologist's radiologic findings. What was considered appropriate was determined by medical

Table 4 Means, standard deviations, and internal consistency for individual characteristics, collaborative diagnostic activities, and diagnostic outcome per study

Variable	Study A			Study B			Study C		
	M	SD	ω	M	SD	ω	M	SD	ω
Conceptual knowledge in internal medicine[c]	0.61	0.12	.37	0.65	0.13	.53	0.49	0.18	.49
Conceptual knowledge in radiology[c]	0.67	0.15	.55	0.68	0.16	.62	0.52	0.18	.39
Strategic knowledge in internal medicine[c]	0.58	0.12	.53	0.61	0.13	.54	0.48	0.14	.40
Strategic knowledge in radiology[c]	0.42	0.12	.38	0.47	0.14	.45	0.44	0.16	.44
Collaboration knowledge[c]	0.70	0.09	.83	0.72	0.09	.83	0.65	0.10	.82
Direct measurement[d]	4.45	0.63	.79	4.36	0.64	.74	4.59	0.67	.81
Perspective taking[d]	4.36	0.57	.70	4.35	0.57	.62	4.50	0.58	.62
Information sharing[d]	4.55	0.52	.59	4.49	0.55	.62	4.56	0.51	.66
Negotiating[d]	4.76	0.57	.46	4.71	0.57	.51	4.71	0.49	.28
Coordination[d]	4.51	0.58	.72	4.51	0.60	.71	4.58	0.59	.76
Evidence elicitation[a,c]	0.49	0.50	–[b]	0.70	0.46	–[b]	0.67	0.47	–[b]
Evidence sharing[c]	0.55	0.19	–[b]	0.60	0.22	–[b]	0.53	0.24	–[b]
Hypotheses sharing[a,c]	0.61	0.49	–[b]	0.62	0.49	–[b]	0.51	0.50	–[b]
Diagnostic accuracy[a,c]	0.82	0.38	–[b]	0.90	0.30	–[b]	0.92	0.27	–[b]
Diagnostic justification[c]	0.34	0.23	–[b]	0.43	0.22	–[b]	0.41	0.20	–[b]
Diagnostic efficiency[c]	0.07	0.02	–[b]	0.07	0.04	–[b]	0.05	0.02	–[b]

[a]Binary indicator

[b]Single measure item

[c]Ranging from 0 to 1

[d]Ranging from 1 to 6

experts for each case and examination in preparation of the cases. Therefore, this scenario involves more than a simple division of tasks, as the quality of one person's activity (i.e., description and evaluation of the radiologic findings) depends on the collaborative efforts (i.e., CDAs) of the other person (OECD, 2017)

Measures—Individual Characteristics

The individual characteristics were measured in the pretest. The internal consistencies of each measure per study are displayed in Table 4 in the Results section. We want to point out that the internal consistency of knowledge as a construct—determined by the inter-correlations among knowledge pieces—typically exhibits a moderate level. Importantly, recent research argues that a moderate level of internal consistency does not undermine the constructs' capacity to explain a significant amount of variance (Edelsbrunner, 2024; Stadler et al., 2021; Taber, 2018).

Content knowledge was separated into radiology and internal medicine knowledge, as these two disciplines play a major role in the diagnosis of the simulated patient cases. For each discipline, conceptual and strategic knowledge was assessed (Kiesewetter et al., 2020; Stark et al., 2011). The items in each construct were presented in a randomized way in each study. However, the items for study C were shortened due to the embedding of the data

collection in the curriculum. Therefore, items with a very high or low item difficulty in previous studies were excluded (Table 2).

Conceptual knowledge was measured using single-choice questions including five options adapted from a database of examination questions from the Medical Faculty of the LMU Munich, focusing on relevant and closely related diagnoses of the patient cases used in the simulation. A mean score of 0–1 was calculated, representing the percentage of correct answers and indicating the average conceptual knowledge of the participant per medical knowledge domain.

Strategic content knowledge was measured contextually using key features questions (M. R. Fischer et al., 2005). Short cases were introduced followed by two to three follow up questions (e.g., What is your most likely suspected diagnosis?, What is your next examination?, What treatment do you choose?). Each question had eight possible answers, from which the learners were asked to choose one. Again, a mean score of 0–1 was calculated, representing the percentage of correct responses, indicating the average strategic content knowledge of the participant per domain.

The measure of collaboration knowledge was consistent across the three studies and specific to the simulated task. Participants were asked to select all relevant information for seven different patient cases with the cardinal symptom of fever (internal medicine). The patient cases were presented in a randomized order and always included 12 pieces of information regarding the chief complaints, medical history, and physical examination of the patient cases. We then assessed whether each piece of information was shared correctly (i.e. whether relevant information was shared and irrelevant information was not shared) and assigned 1 point and divided it by the maximum of 12 points to standardized the range of measure to 0–1. Then we calculated a mean score for each case and then across all cases, resulting in a range of 0–1 indicating the participants' collaboration knowledge

The construct of social skills was consistent across the three data collections and was measured on the basis of self-report on a 6-point Likert scale ranging from total disagreement to total agreement. The construct was measured using 23 questions divided into five subscales; for example items, see Table 3. Five questions aimed to measure the overall construct, and the other four subscales were identified using the complex problem-solving frameworks of Liu et al. (2016) and Hesse et al. (2015): perspective taking (four questions), information sharing (five questions), negotiation (four questions), and coordination (five questions). For the final score, the mean of all subcategories was calculated, ranging from 1 to 6, representing general social skills.

Measures—Collaborative Diagnostic Activities (CDA)

We operationalize CDAs in the pretest case in terms of quality of evidence elicitation, evidence sharing, and hypotheses sharing. The internal consistencies of each measure per study are displayed in Table 4 in the Results section.

The quality of evidence elicitation was measured by assessing the appropriateness of the requested radiological examination for the indicated diagnosis. An expert solution was developed to indicate which radiological examinations were appropriate for each of the possible diagnoses. If participants requested an appropriate radiological examination for the indicated diagnoses, they received 1 point for that request attempt. Finally, a mean score across all request attempts (maximum of 3) was calculated and scored. The final mean score was transformed into a binary indicator, with 1 indicating that all requested radiological examinations were appropriated and 0 indicating that inappropriate radiological

examinations were also requested, due to the categorical nature of the original data and its skewed distribution, with a majority of responses concentrated in a single category.

The quality of evidence sharing was measured using a precision indicator. This was calculated as the proportion of shared relevant evidence out of all shared evidence. Relevant evidence is defined per case and per diagnosis and indicated by the expert solution. The precision indicator was first calculated per radiological request. We then calculated the mean score, summarizing all attempts in that patient case. This resulted in a range from 0 points, indicating that only irrelevant evidence was shared, to 1 point, indicating that only relevant evidence was shared.

The quality of hypotheses sharing was also measured using a precision indicator. For each patient case, the proportion of diagnoses relevant for the respective patient case to all shared diagnoses was calculated. Which diagnoses were considered relevant for a specific case was determined by an expert solution. As with evidence elicitation, this score was evaluated and converted into a binary variable, where 1 indicated that only relevant diagnoses were shared and 0 indicated that also irrelevant diagnoses were shared, due to the categorical nature of the original data and its skewed distribution, with a majority of responses concentrated in a single category.

Measures—Diagnostic Outcome

We operationalize diagnostic outcome in the pretest case in terms of diagnostic accuracy, diagnostic justification, and diagnostic efficiency.

For diagnostic accuracy, a main diagnosis was assigned to each patient case as expert solution. After working on the patient case and requesting the radiological examination, participants indicated their final diagnosis. To do this, they typed in the first three letters of their desired diagnosis and then received suggestions from a list of 249 possible diagnoses. Diagnostic accuracy was then calculated by coding the agreement between the final diagnosis given and the expert solution. Accurate diagnoses (e.g., hospital-acquired pneumonia) were coded as 1, correct but inaccurate diagnoses (e.g., pneumonia) were coded as 0.5, and incorrect diagnoses were coded as 0. A binary indicator was used for the final diagnostic accuracy score, with 0 indicating an incorrect diagnosis and 1 indicating an at least inaccurate diagnosis, due to the categorical nature of the original data and its skewed distribution, with a majority of responses concentrated in a single category.

A prerequisite for diagnostic justification and diagnostic efficiency is the provision of at least an inaccurate diagnosis. If a participant provided an incorrect diagnosis (coded as 0), diagnostic justification and diagnostic efficiency were immediately scored as 0.

After choosing a final diagnosis, participants were asked to justify their decision in an open text field. Diagnostic justification was then calculated as the proportion of relevant reported information out of all relevant information that would have fully justified the final accurate diagnosis. Again, medical experts agreed on an expert solution that included all relevant information to justify the correct diagnosis. The participants' solution was coded by two independent coders, each coding the full data, and differences in coding were discussed until the coders agreed. We obtained a range from 0 points, indicating a completely inadequate justification, to 1 point, indicating a completely adequately justified final diagnosis.

Diagnostic efficiency was defined as diagnostic accuracy (non-binary version) divided by the minutes required to solve the case.

Statistical Analyses

To answer the research question, a structural equation model (SEM) was estimated using MPlus Editor, version 8 (Muthén & Muthén, 2017). We decided to use a SEM, as it is a comprehensive statistical approach widely used in psychology and educational sciences for its ability to model complex relationships among observed and latent variables while accounting for measurement error (Hilbert & Stadler, 2017). SEM support the development and verification of theoretical models, enabling scholars to refine theories and interventions in psychology and education based on empirical evidence, as not only can one relationship be investigated but a system of regressions is also considered simultaneously (Nachtigall et al., 2003).

We included all links between the variables and applied a two-step approach, using mean-adjusted and variance-adjusted unweighted least squares (ULSMV, Savalei & Rhemtulla, 2013) as the estimator and THETA for parametrization, first examining the measurement model and then the structural model. The assessment of model fit was based on chi-square ($\chi 2$), root mean square error of approximation (RMSEA), and comparative fit index (CFI). Model fit is generally indicated by small chi-squared values; RMSEA values of $<$ 0.08 (acceptable) and $<$ 0.06 (excellent), and CFI values $\geq$ 0.90. We do not consider standardized root mean squared residual (SRMR), because, according to the definition used in MPlus, this index is not appropriate when the sample size is 200 or less, as natural variation in such small samples contributes to larger SRMR values (Asparouhov & Muthén, 2018). For hypotheses 1–6, we excluded path coefficients $<$ 0.1 from our interpretation, as they are relatively small. In addition, at least two interpretable path coefficients, of which at least one is statistically significant, are required to find support for the hypothesis. For hypotheses 7–15, specific indirect effects (effect of an individual characteristic on diagnostic outcome through a specific CDA) and total indirect effects (mediation of the effect of an individual characteristic on diagnostic outcome through all mediators) were estimated.

We reported all measures in the study and outlined differences between the three samples. All data and analysis code have been made publicly available at the Open Science Framework (OSF) and can be accessed at https://osf.io/u8t62. Materials for this study are available by email through the corresponding author. This study's design and its analysis were not pre-registered.

Results

The descriptive statistics of each measure per study are displayed in Table 4. The intercorrelations between the measures per study can be found in Appendix Table 7.

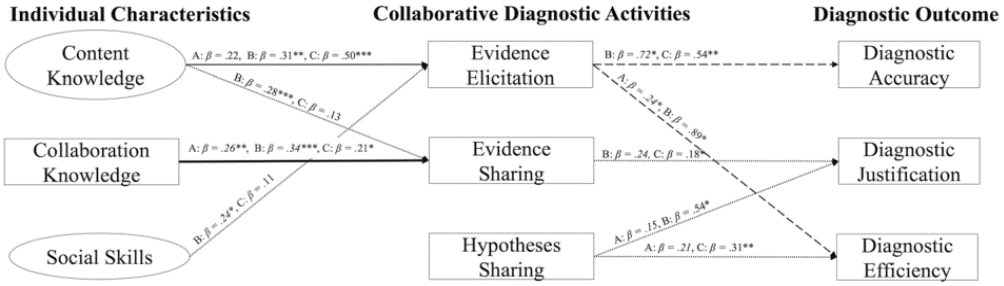
Overall Results of the SEM

All loadings were in the expected directions and statistically significant, except for conceptual knowledge in internal medicine in study C ($\lambda = 0.241$, $p = .120$), conceptual knowledge in radiology in study A ($\lambda = 0.398$, $p = .018$), and strategic knowledge in internal medicine ($\lambda = 0.387$, $p = .206$) and radiology ($\lambda = -0.166$, $p = .302$) in study B. Standardized factor loadings of the measurement model are shown in Appendix Table 8.

Table 5 Standardized paths coefficients (β) and standard errors (SE) for paths between individual characteristics, collaborative diagnostic activities, and diagnostic outcome per study

Hypothesis	Path between		Study A		Study B		Study C		Conclusion
	Independent variable	Dependent variable	β	SE	β	SE	β	SE	
1a	Content knowledge	EE	.22	0.12	.31**	0.12	.50***	0.10	Supported
1b		ES	.08	0.08	.28***	0.08	.13	0.09	Supported
1c		HS	.12	0.12	-.14	0.12	.1	0.12	Not supported
2a	Collaboration knowledge	EE	-.03	0.10	.16	0.10	-.04	0.08	Not supported
2b		ES	.26**	0.08	.34***	0.07	.21*	0.08	Fully supported
2c		HS	.06	0.10	.09	0.10	.03	0.09	Not supported
3a	Social skills	EE	.01	0.11	.24*	0.11	.11	0.10	Supported
3b		ES	.00	0.08	.23**	0.08	-.03	0.08	Not supported
3c		HS	-.1	0.11	-.1	0.11	.11	0.11	Not supported
4a	EE	Diagnostic accuracy	-.25	0.16	.72*	0.37	.54**	0.19	Supported
4b		Diagnostic justification	.03	0.12	-.69	0.37	.25	0.16	Not supported
4c		Diagnostic efficiency	.24*	0.12	.89*	0.40	.08	0.18	Supported
5a	ES	Diagnostic accuracy	-.02	0.12	-.27	0.25	.20	0.20	Not supported
5b		Diagnostic justification	-.19	0.10	.24	0.21	.18*	0.09	Supported
5c		Diagnostic efficiency	.09	0.09	-.18	0.22	.05	0.07	Not supported
6a	HS	Diagnostic accuracy	.13	0.16	-.22	0.33	.13	0.17	Not supported
6b		Diagnostic justification	.15	0.12	.54*	0.27	-.18	0.12	Supported
6c		Diagnostic efficiency	.21	0.11	-.4	0.28	.31 **	0.11	Supported

EE evidence elicitation, *ES* evidence sharing, *HS* hypotheses sharing. *$p < .05$, **$p < .01$, and ***$p < .001$



Only coefficients > 0.10 are displayed. The full thick line represents a positive statistically significant relationship in three studies; the full line represents a positive relationship in three studies, of which two are statistically significant; the dashed line represents a positive statistically significant relationship in two studies; and the dotted line represents a positive statistically significant relationship in two studies, of which one is statistically significant. $p < .05$. $^{**}p < .01$. $^{***}p < .001$

Fig. 2 Evidence on supported relationships between individual characteristics, collaborative diagnostic activities, and diagnostic outcome

The SEM has a good fit for study A [$X^2(75) = 74.086$, $p = .508$, RMSEA = 0.00, CFI = 1.00], study B [$X^2(75) = 68.309$, $p = .695$, RMSEA = 0.000, CFI = 1.00], and study C [$X^2(75) = 93.816$, $p = .070$, RMSEA = 0.036, CFI = 1.00].

Paths between Individual Characteristics, CDAs, and Diagnostic Outcome

The standardized path coefficients and hypotheses tests for the theoretical model are reported in Table 5. An overview of the paths supported by the data is shown in Fig. 2.

Overall, the R^2 for the CDAs ranged from medium to high for evidence elicitation and evidence sharing, depending on the study, and were consistently low for hypotheses sharing across all three studies. Looking at diagnostic outcome, R^2 is consistently large for diagnostic accuracy and medium to large for diagnostic justification and diagnostic efficiency (Table 6).

Table 6 R^2 for collaborative diagnostic activities and diagnostic outcome per study

Variable	Study A		Study B		Study C	
	R^2	SE	R^2	SE	R^2	SE
EE [a]	.051	0.05	.166	0.09	.289	0.11
ES [a]	.077	0.04	.234***	0.07	.061	0.04
HS [a]	.023	0.03	.035	0.04	.027	0.03
Diagnostic accuracy [a]	.286	0.13	.343	0.20	.332	0.17
Diagnostic justification	.165	0.08	.246	0.19	.146	0.07
Diagnostic efficiency	.141	0.07	.422	0.22	.143	0.10

EE evidence elicitation, ES evidence sharing, HS hypotheses sharing. *p < .05, **p < .01, and ***p < .001

[a]For dichotomous criterions, MPlus computes a pseudo-R^2

The path from content knowledge to evidence elicitation was positive and > 0.1 in all three studies, as well as statistically significant in two of them; therefore, we consider Hypothesis 1a supported. The path from content knowledge to evidence sharing was positive and > 0.1 in two studies, as well as statistically significant in one of them; therefore, Hypothesis 1b is also supported. In contrast, the path from content knowledge to hypotheses sharing was indeed also positive in two studies, but as neither was statistically significant, we conclude that Hypothesis 1c was not supported. The path from collaboration knowledge to evidence elicitation was positive and > 0.1 in only one study, but also not statistically significant. Thus, we found that Hypothesis 2a was not supported. For the path from collaboration knowledge to evidence sharing, we found relevant positive and statistically significant coefficients in all three studies. Hypothesis 2b is therefore fully supported by the data. This is not the case for Hypothesis 2c, for which we found no coefficient > 0.1 for the path from collaboration knowledge to hypotheses sharing. For the path from social skills to evidence elicitation, we found positive coefficients > 0.1 in two out of three studies, of which one was also statistically significant. Thus, we consider Hypothesis 3a to be supported. For the path from social skills to evidence sharing, we again found one statistically significant positive coefficient, but in the other two studies it was < 0.1. Therefore, we do not consider Hypothesis 3b to be supported by the data. The same applies to the path from social skills to hypotheses sharing, where the coefficient is < 0.1 in two studies. We therefore do not consider Hypothesis 3c to be supported.

The path from evidence elicitation to diagnostic accuracy was statistically significant and large in magnitude in two out of three studies. Hypothesis 4a is therefore supported. The path from evidence elicitation to diagnostic justification was only positive and > 0.1 in one study, which was also not statistically significant. Therefore, we find no support for Hypothesis 4b. In contrast, the path from evidence elicitation to diagnostic efficiency was positive and statistically significant in two out of three studies, with one large effect. Hypothesis 4c is therefore supported. The path from evidence sharing to diagnostic accuracy was only positive and reasonably large in one study. Therefore, we do not find support for Hypothesis 5a. The path from evidence sharing to diagnostic justification was positive and > 0.1 in two studies as well as statistically significant in one of them, so Hypothesis 5b is supported. In contrast, we did not find a positive coefficient > 0.1 for the path from evidence sharing to diagnostic efficiency. Therefore, Hypothesis 5c is not supported by the data. Although we found coefficients > 0.1 in two studies for the path from hypotheses sharing to diagnostic accuracy, we found no support for Hypothesis 6a, as none of these was statistically significant. This is different for Hypothesis 6b, as we found two positive paths from hypotheses sharing to diagnostic justification, one of which was statistically significant and large. Finally, we found two positive paths from evidence sharing to diagnostic efficiency in three studies, one of which was statistically significant. Hypothesis 6c is therefore supported.

Indirect Effects between Individual Characteristics, CDA, and Diagnostic Outcome

Indirect effects of CDAs on the effect of individual characteristics on the diagnostic outcome in CDR were estimated to test hypotheses 7–15. Although we found a mediating effect of all CDAs ($\beta = .31$, $p = .008$), and specifically for evidence elicitation ($\beta = .27$, $p = .021$) from content knowledge on diagnostic accuracy in study C, and some significant overall and direct effects for other relationships (Appendix Table 9), none of these were consistent across all of the studies. Thus, we conclude no consistent support for any of the Hypotheses 7–15.

Discussion

The aim of the current study was to investigate the extent to which the relationships specified in the CDR model (Radkowitsch et al., 2022) are applicable across studies, to better understand the processes underlying CDR in knowledge-rich domains. Not only is this exploration crucial for the medical field or collaborative problem-solving in knowledge-rich domains, but it also offers valuable insights for computer-supported collaborative learning research. Despite CDR's specific focus, the principles and findings have relevant implications for understanding and enhancing collaborative processes in various educational and professional settings.

Specifically, we investigated how individual learner characteristics, the CDAs, and the diagnostic outcome are related. We therefore analyzed data from three independent studies, all from the same context, a simulation-based environment in the medical domain. Our study found positive relationships between content knowledge and the quality of evidence elicitation as well as the quality of evidence sharing, but not for the quality of hypotheses sharing. Furthermore, collaboration knowledge is positively related to the quality of evidence sharing, but not to the quality of evidence elicitation and the quality of hypotheses sharing. Social skills are only positively related to the quality of evidence elicitation. This underscores the multifaceted nature of collaborative problem-solving situations. Thus, effective CDR, a form of collaborative problem-solving, necessitates a nuanced understanding of the interplay between individual characteristics and CDAs.

The relevance of content knowledge for diagnostic competence is well established in research (Chernikova et al., 2020). To develop any diagnostic skills in knowledge-rich domains, learners need to acquire large amounts of knowledge and to restructure it through experience with problem-solving procedures and routines (Boshuizen et al., 2020). In the case of CDR this enables the diagnostician to come up with an initial suspected diagnosis, which is likely to be relevant information for the collaboration partner and to guide the further CDAs effectively. The finding that content knowledge only has a relation to the quality of evidence elicitation but none of the other CDAs can be explained by the fact that evidence elicitation is the least transactive CDA within the collaborative decision-making process. When eliciting evidence, the collaboration partner is used as an external knowledge resource (Weinberger & Fischer, 2006). So, despite being a collaborative activity, evidence elicitation is about what information from the collaboration partner is needed rather than what the collaboration partner needs. Thus, elicitation is less transactive than sharing, which is focused at what the collaboration partner needs.

Not only content knowledge but also collaboration knowledge is related to the quality of evidence sharing. This finding implies that collaboration knowledge may influence the CDR above and beyond individual content knowledge. It also supports the differentiation of knowledge types made in the CDR model (Radkowitsch et al., 2022). Thus, it is important to learn not only the conceptual and strategic medical knowledge that is required for diagnosing but also knowledge about what information is relevant for specific collaboration partners when diagnosing collaboratively. This finding underpins the importance of being aware of the knowledge distribution among collaboration partners and the relevance of the transactive memory (Wegner, 1987). Thus, for collaborative problem-solving in knowledge-rich domains—as for computer-supported collaborative learning more generally—knowledge and information awareness is crucial (Engelmann & Hesse, 2010).

Thus, the relevance of collaboration knowledge in collaborative problem-solving is an important finding of our study, highlighting that it is critical in facilitating effective

collaborative processes and outcomes. The current findings emphasize the need for educational strategies that explicitly target the development of collaborative knowledge to ensure that learners have the knowledge and skills necessary to participate in productive collaborative problem-solving and computer-supported collaborative learning processes. In doing so, the CDR model emphasizes the need for learners to master collaborative skills and build shared problem representations to take full advantage of collaborative learning opportunities.

As CDR is conceptualized to be an interplay of cognitive and social skills (Hesse et al., 2015), we also assumed that social skills are related to CDAs. However, we only found evidence of the expected relationship between social skills and CDAs for the quality of evidence elicitation. One explanation could be that collaboration knowledge was relatively high in all three samples, outweighing the influences of general skills. This is consistent with the assumption of the CDR model that the influence of more general social skills is reduced with an increasing level of professional collaboration knowledge (Radkowitsch et al., 2022). When collaboration knowledge is available to the diagnosticians, it becomes more important than social skills. This finding again underlines the importance of collaboration knowledge, which can be seen as a domain- and profession-specific development of social skills. However, another explanation could be that, when collaborating with an agent, the effect of social skills decreases, as the agent was not programmed to respond to social nuances. The design of the simulation would thus buffer against the effect of social skills. Although the study by Herborn et al. (2020) found no differences between human-to-human and human-to-agent collaboration, this does not necessarily invalidate the potential variability in outcomes associated with the social skills incorporated into the agent. For a thorough investigation into the impact of social skills, the agent would need variable social abilities, enabling the variation of the importance of basic social skills for successful collaboration.

Further, we need to conclude that there is no support for a relationship between the individual characteristics and hypotheses sharing, as we found no stable support for the relationship between any of the individual characteristics and the quality of hypotheses sharing. One possible explanation could be that the binary precision measure used to operationalize quality in hypotheses sharing is not sensitive enough or is not capturing the relevant aspect of quality in that activity. Another explanation could be that there is no direct relationship between the individual characteristics and hypotheses sharing, as this relationship is mediated by evidence sharing and thus influenced by the activated knowledge scripts (Schmidt & Rikers, 2007).

Looking at the relationships between CDAs and the diagnostic outcome, the current results highlight the need to distinguish between primary (diagnostic accuracy) and secondary (diagnostic justification and efficiency) outcomes of diagnostic reasoning (Daniel et al., 2019). Achieving diagnostic accuracy, a purely quantitative outcome measure, is less transactive than other aspects of the diagnostic outcome. This is also where we find the link to evidence elicitation, as we consider this to be the least transactive CDA within the collaborative decision-making process. However, the ability to justify and reach this decision efficiently is then highly dependent on evidence sharing and hypotheses sharing, activities that are more focused on transactivity within CDR (Weinberger & Fischer, 2006).

Although individual learner characteristics are found to have an effect on CDAs, and CDAs impact the diagnostic outcome, the effect is not mediated by CDAs across studies. Thus, we assume that, for effective collaborative problem-solving in knowledge-rich domains, such as CDR, it is not enough to have sufficient content and collaboration knowledge; it is also necessary to be able to engage in high quality CDAs to achieve a

high-quality diagnostic outcome. This is consistent with research on individual diagnostic reasoning, which shows that diagnostic activities have a unique contribution to the diagnostic outcome after controlling for content knowledge (Fink et al., 2023).

In summary, we explored evidence elicitation, evidence sharing, and hypotheses sharing as crucial CDAs. The findings revealed diverse associations of these CDAs with individual characteristics and facets of the diagnostic outcome, supporting the notion that the CDR-process involves a variety of different skills (instead of being one overarching skill). On the basis of these results, we propose categorizing CDAs into activities primarily focused on individual goals and needs (e.g., elicitation) and more transactive activities directly targeted at the collaborator (e.g., sharing). To enhance quality in CDAs, instructional support should be considered. For instance, providing learners with an adaptive collaboration script has been shown to improve evidence sharing quality and promote the internalization of collaboration scripts, fostering the development of collaboration knowledge (Radkowitsch et al., 2021). Further, group awareness tools, such as shared concept maps, should be considered to compensate for deficits in one's collaboration knowledge (Engelmann & Hesse, 2010). However, what is required to engage in high-quality CDAs remains an open question. One starting point is domain-general cognitive skills. These could influence CDAs, particularly in the early stages of skill development (Hetmanek et al., 2018). Previous research showed that, in diagnostic reasoning, instructional support is more beneficial when being domain-specific than domain-general (Schons et al., 2022). Thus, there is still a need for further research on how such instructional support might look like.

Future Research

Although we used data from three studies, all of them were in the same domain; thus, it remains an open question whether these findings are applicable across domains. The CDR model claims that the described relationships are not limited to the medical domain, but rather are valid across domains for collaboratively solving complex problems in knowledge-rich domains. Future research should explore generalizability, for example, for teacher education, which is a distinct field that also requires diagnosing and complex problem-solving (Heitzmann et al., 2019).

Regardless of domain, the non-mediating relationship of CDAs between individual characteristics and diagnostic outcomes, as well as the found effects of the CDAs in the current study, suggests that an isolated analysis of CDAs does not fully represent the complex interactions and relationships among activities, individual characteristics, and diagnostic outcomes. Future studies might assess CDAs as a bundle of necessary activities, including a focus on their possible non-linear interactions. We propose to use process data analysis to account for the inherent complexity of the data, as different activities in different sequences can lead to the same outcome (Y. Chen et al., 2019). More exploratory analyses of fine-grained, theory-based sequence data are needed to provide insights into more general and more specific processes involved in successful solving complex problems collaboratively (Stadler et al., 2020).

As our results have shown, collaboration knowledge and thus awareness of the knowledge distribution among collaboration partners is highly relevant. While a recent meta-analyses showed a moderate effect of group awareness of students' performance in computer-supported collaborative learning (D. Chen et al., 2024), it has so far not been systematically investigated in collaborative problem-solving. Thus, more research on the influence collaboration knowledge in collaborative problem-solving is needed.

Further, additional factors associated with success in collaborative problem-solving—not yet incorporated into the model and thus not yet investigated systematically—include communication skills (OECD, 2017), the self-concept of problem-solving ability (Scalise et al., 2016), and positive activating emotions during problem-solving tasks (Camacho-Morles et al., 2019).

Limitations

There are, however, some limitations to be considered. One is that we have only considered CDAs and how they relate to individual characteristics and outcomes. However, the CDR model also introduces individual diagnostic activities, such as the generation of evidence and the drawing of conclusions. These occur before and after the CDAs and may therefore also have an impact on the described relationships. However, we decided to focus on the CDAs within the CDR process because they are particularly relevant for constructing a shared problem representation, being central to CDR. Future research might consider these individual diagnostic activities, as they could, for example, further explain the how content knowledge is related to the diagnostic outcome.

Another limitation of the current analyses is the operationalization of quality for the CDAs. We chose the appropriateness of radiological examination for the indicated diagnosis for quality of evidence elicitation and precision for quality of evidence sharing and hypotheses sharing. However, all of these only shed light on one perspective of each activity, while possibly obscuring others. For example, it may be that content knowledge is not related to the precision of hypotheses sharing, but this may be different when looking at other quality indicators, such as sensitivity or specificity. However, we decided to use the precision aspect of activities, as research shows that collaborators often fail to identify relevant information, and the amount of information is not related to performance (Tschan et al., 2009). Future research may explore a broader variety of quality indicators to be able to assess the quality of CDAs as comprehensively as possible. It should also be noted that in study B a suppression effect (Horst, 1941) between hypothesis sharing and evidence elicitation artificially inflated the observed effect size. This is to be expected with process data that can be highly correlated and needs to be considered when interpreting the effect sizes.

In addition, it should be noted that the omega values obtained for the conceptual and strategic knowledge measures were below the commonly accepted threshold of 0.7. While we chose to use omega values as a more appropriate measure of reliability in our context, given the complex and multifaceted nature of the knowledge constructs, these lower-than-expected values raise important questions about the quality of the data and the robustness of the findings. Thus, it is important to understand that knowledge constructs, by their very nature, may not always exhibit high levels of internal consistency due to the diverse and interrelated components they encompass (Edelsbrunner, 2024; Stadler et al., 2021; Taber, 2018). This complexity may be reflected in the moderate omega values observed, which, while seemingly counterintuitive, does not invalidate the potential of the constructs to account for substantial variance in related outcomes. However, findings related to these constructs should be interpreted with caution, and the results presented should be considered tentative. Future research should further explore the implications of using different reliability coefficients in assessing complex constructs within the learning sciences, potentially providing deeper insights into the nuanced nature of knowledge and its measurement.

Another limitation of this study is related to the agent-based collaboration, as a predictive validation of collaborative problem-solving for later human-to-human collaboration in comparable contexts has not yet been systematically conducted. Although the agent-based collaboration situation used has been validated in terms of perceived authenticity, it still does not fully correspond to a real collaboration situation (Rosen, 2015). This could be an explanation for the low influence of social skills, as the setting might not require the application of a broad set of social skills (Hesse et al., 2015; Radkowitsch et al., 2020). In a real-life collaboration, the effects of social skills might be more pronounced. However, research showed that the human-to-agent approach did not lead to different results in collaborative problem-solving than the human-to-human approach in the 2015 PISA study, and correlations with other measures of collaborative skills have been found (Herborn et al., 2020; Stadler, Herborn et al., 2020). Future studies should specifically test the relevance of social skills for CDR in a human-to-human setting to strengthen the generalizability of our findings.

Conclusion

In conclusion, the current study highlights the importance of individual characteristics and CDAs as independent predictors for achieving good diagnoses in collaborative contexts, at least in the simulation-based settings we used in the studies included in our analysis. Collaboration knowledge emerged as a critical factor, demonstrating its importance over early acquired, general social skills. Therefore, it is imperative to revise the CDR approach by giving higher priority to the proficiency of collaboration knowledge compared with social skills. Furthermore, we conclude that, in simulation-based CDR, content knowledge does not play such a crucial role in predicting diagnostic success compared with many other educational settings, most probably because of the endless opportunities for retrying and revising in simulation-based learning environments.

With respect to CDAs, we suggest refining the perspective on the quality of CDAs and consider revising the CDR model by summarizing CDAs as information elicitation and information sharing, with the former being less transactive, and thus, less demanding than the latter. Adequate performance in both types of CDA is presumed to result in a high-quality shared problem representation, resulting in good diagnostic outcome. Collaborative problem-solving skills are highly relevant in professional practice of knowledge-rich domains, highlighting the need to strengthen these skills in students engaged in CDR and to provide learning opportunities accordingly. Further, the ability to effectively collaborate and construct shared problem representations is important, not only in CDR but also in collaborative problem-solving and computer-supported collaborative learning more in general, highlighting the need for integrating such skills into curricula and instructional design.

By emphasizing these aspects, we can improve the diagnostic skills of individuals in collaborative settings. Through advancing our understanding of CDR, we are taking a key step forward in optimizing collaborative problem-solving and ultimately contributing to improved diagnostic outcomes in various professional domains beyond CDR in medical education. In particular, integrating collaboration knowledge and skills into computer-supported collaborative learning environments can enrich learning experiences and outcomes in various knowledge-rich domains.

Appendix

Table 7 Intercorrelations (Pearson), for the latent and manifest variables

	Study A								
	Variable	1	2	3	4	5	6	7	8	9
1	Content Knowledge	–								
2	Collaboration Knowledge	0.21**	–							
3	Social Skills	0.13	−0.03	–						
4	EE	0.12	−0.03	0.04	–					
5	ES	0.14	0.26***	0.01	0.04	–				
6	HS	0.06	0.05	−0.08	−0.12	−0.05	–			
7	Diagnostic Accuracy	0.23**	0.02	0.02	−0.09	0.00	0.13	–		
8	Diagnostic Justification	0.21*	0.05	0.15	0.4	−0.15	0.14	–[a]	–	
9	Diagnostic Efficiency	−0.11	0.10	−0.16	0.10	0.11	0.11	–[a]	−0.00	–
	Study B									
	Variable	1	2	3	4	5	6	7	8	9
1	Content Knowledge	–								
2	Collaboration Knowledge	0.16*	–							
3	Social Skills	−0.04	0.06	–						
4	EE	0.26**	0.14	0.16*	–					
5	ES	0.25**	0.34***	0.19*	0.46***	–				
6	HS	−0.03	0.09	−0.07	0.29***	−0.01	–			
7	Diagnostic Accuracy	0.05	0.12	−0.06	0.25**	0.11	0.07	–		
8	Diagnostic Justification	0.10	0.05	0.04	−0.11	−0.02	0.12	–[a]	–	
9	Diagnostic Efficiency	0.10	0.14	−0.17	0.22*	0.20*	0.07	–[a]	−0.26**	–
	Study C									
	Variable	1	2	3	4	5	6	7	8	9
1	Content Knowledge	–								
2	Collaboration Knowledge	0.09	–							
3	Social Skills	0.17*	0.16*	–						
4	EE	0.29***	−0.03	0.15*	–					
5	ES	0.09	0.19**	0.02	0.03	–				
6	HS	0.07	0.00	0.08	0.25***	−0.06	–			
7	Diagnostic Accuracy	0.05	−0.03	0.01	0.25**	0.09	0.14	–		
8	Diagnostic Justification	0.16	0.15	0.05	0.14	0.24**	−0.08	–[a]	–	
9	Diagnostic Efficiency	0.12	0.03	−0.01	0.16*	0.07	0.26***	–[a]	−0.08	–

EE evidence elicitation, *ES* evidence sharing, *HS* hypotheses sharing. *$p < .05$, **$p < .01$, and ***$p < .001$

[a]These correlations cannot be calculated as if diagnostic accuracy were 0; diagnostic justification and efficiency are coded as NA

Table 8 Standardized factor loadings (λ) and standard errors (SE) in the measurement model

Factor	Observed variable	Study A		Study B		Study C	
		λ	SE	λ	SE	λ	SE
Content knowledge	Conceptual knowledge in internal medicine	0.709***	0.168	0.741***	0.166	0.241	0.155
	Conceptual knowledge in radiology	0.398*	0.168	0.626***	0.148	0.424**	0.136
	Strategic knowledge in internal medicine	0.525**	0.169	0.387	0.206	0.797***	0.118
	Strategic knowledge in radiology	0.755**	0.217	−0.166	0.302	0.651***	0.113
Social skills	Direct measurement	0.612***	0.05	0.625***	0.056	0.681***	0.065
	Perspective taking	0.595***	0.062	0.642***	0.06	0.639***	0.059
	Information sharing	0.671***	0.047	0.719***	0.056	0.627***	0.068
	Negotiating	0.611***	0.055	0.614***	0.051	0.511***	0.065
	Coordination	0.685***	0.05	0.65***	0.054	0.754***	0.052

SE standard errors. *$p < .05$, **$p < .01$, and ***$p < .001$

Table 9 Mediation between individual characteristics and diagnostic outcomes by collaborative diagnostic activities

Mediation		Study A		Study B		Study C	
Name	Type	β	SE	β	SE	β	SE
Content knowledge—diagnostic accuracy	Total	.45***	0.111	.26	0.151	.13	0.134
	Total indirect	-.04	0.063	.18	0.145	.31**	0.118
	Via EE	-.06	0.05	.22	0.148	.27*	0.117
	Via ES	0	0.01	-.08	0.076	.03	0.033
	Via HS	.01	0.023	.03	0.052	.01	0.022
	Direct	.49***	0.122	.09	0.173	-.18	0.166
Content knowledge—diagnostic justification	Total	.26*	0.109	.16	0.101	.19	0.104
	Total indirect	.01	0.043	-.22	0.149	.13	0.086
	Via EE	.01	0.027	-.22	0.147	.12	0.085
	Via ES	-.01	0.021	.07	0.061	.03	0.02
	Via HS	.02	0.021	-.07	0.079	-.02	0.025
	Direct	.26*	0.119	.38*	0.182	.06	0.139
Content knowledge—diagnostic efficiency	Total	-.13	0.102	.09	0.11	.17	0.104
	Total indirect	.09	0.05	.28	0.17	.08	0.093
	Via EE	.05	0.039	.28	0.187	.04	0.089
	Via ES	.01	0.012	-.05	0.065	.01	0.011
	Via HS	.03	0.029	.06	0.062	.03	0.037
	Direct	-.21*	0.104	-.2	0.198	.09	0.143
Collaboration knowledge—diagnostic accuracy	Total	.04	0.119	.18	0.146	-.08	0.241
	Total Indirect	.01	0.047	0	0.096	.02	0.071
	Via EE	.01	0.025	.12	0.092	-.02	0.046
	Via ES	0	0.033	-.09	0.089	.04	0.045
	Via HS	.01	0.016	-.02	0.038	0	0.011
	Direct	.03	0.127	.18	0.152	-.1	0.241
Collaboration knowledge—diagnostic justification	Total	.05	0.095	.05	0.08	.16	0.086
	Total indirect	-.04	0.037	.02	0.09	.03	0.031
	Via EE	0	0.005	-.11	0.09	-.01	0.021
	Via ES	-.05	0.031	.08	0.076	.04	0.024
	Via HS	.01	0.017	.05	0.057	0	0.016
	Direct	.1	0.1	.03	0.122	.14	0.091
Collaboration knowledge—diagnostic efficiency	Total	.12	0.1	.13	0.11	.04	0.092
	Total indirect	.03	0.04	.05	0.1	.01	0.035
	Via EE	-.01	0.024	.14	0.113	0	0.01
	Via ES	.03	0.026	-.06	0.077	.01	0.016
	Via HS	.01	0.023	-.04	0.047	0	0.028
	Direct	.09	0.103	.09	0.141	.04	0.089
Social skills—diagnostic accuracy	Total	-.05	0.128	-.1	0.125	.05	0.142
	Total indirect	-.02	0.035	.14	0.112	.07	0.061
	Via EE	0	0.027	.18	0.123	.06	0.054
	Via ES	0	0.001	-.06	0.059	-.01	0.018
	Via HS	-.01	0.022	.02	0.041	.01	0.023
	Direct	-.03	0.131	-.23	0.164	-.02	0.141

Table 9 (continued)

Mediation		Study A		Study B		Study C	
Name	Type	β	SE	β	SE	β	SE
Social Skills—Diagnostic Justification	Total	.14	0.104	.03	0.1	-.03	0.089
	Total indirect	-.02	0.027	-.17	0.109	0	0.037
	Via EE	0	0.004	-.17	0.114	.03	0.03
	Via ES	0	0.016	.05	0.049	0	0.016
	Via HS	-.01	0.021	-.06	0.069	-.02	0.022
	Direct	.15	0.101	.2	0.125	-.04	0.086
Social Skills—Diagnostic Efficency	Total	-.14	0.094	-.2*	0.082	-.06	0.096
	Total indirect	-.02	0.032	.22	0.13	.04	0.04
	Via EE	0	0.026	.22	0.143	.01	0.022
	Via ES	0	0.008	-.04	0.051	0	0.004
	Via HS	-.02	0.025	.04	0.057	.03	0.036
	Direct	-.12	0.097	-.42**	0.136	-.1	0.089

EE evidence elicitation, *ES* evidence sharing, *HS* hypotheses sharing. *$p < .05$, **$p < .01$, and ***$p < .001$

Funding Open Access funding enabled and organized by Projekt DEAL. The research presented in this contribution was funded by a grant of the Deutsche Forschungsgemeinschaft (DFG, FOR 2385) to Frank Fischer, Martin R. Fischer and Ralf Schmidmaier (FI 792/11-1 & FI 792/11-2)

Declarations

Conflict of interest statement On behalf of all authors, the corresponding author states that there is no conflict of interest.

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

Asparouhov, T., & Muthe´n, B. (2018). *SRMR in Mplus*. https://www.statmodel.com/download/SRMR2.pdf

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*, Article 977631. https://doi.org/10.3389/feduc.2022.977631

Boshuizen, H. P., 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

Brandl, L., Richters, C., Radkowitsch, A., Obersteiner, A., Fischer, M. R., Schmidmaier, R., Fischer, F., & Stadler, M. (2021). Simulation-based learning of complex skills: Predicting performance with theoretically derived process features. *Psychological Test and Assessment Modeling, 63*(4), 542–560.https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2021-4/PTAM__4-2021_6_kor.pdf

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

Camacho-Morles, J., Slemp, G. R., Oades, L. G., Morrish, L., & Scoular, C. (2019). The role of achievement emotions in the collaborative problem-solving performance of adolescents. *Learning and Individual Differences, 70*, 169–181. https://doi.org/10.1016/j.lindif.2019.02.005

Chen, D., Zhang, Y., Luo, H., Zhu, Z., Ma, J., & Lin, Y. (2024). Effects of group awareness support in CSCL on students' learning performance: a three-level meta-analysis. *International Journal of Computer-Supported Collaborative Learning, 19*(1), 97–129. https://doi.org/10.1007/s11412-024-09418-3

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem-solving process data: an event history analysis approach. *Frontiers in Psychology*, *10*, Article 486. https://doi.org/10.3389/fpsyg.2019.00486

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., Opitz, A., Seidel, T., & Fischer, F. (2022). A theoretical framework for fostering diagnostic competences with simulations in higher education. In F. Fischer & A. Opitz (Eds.), *Learning to Diagnose with Simulations*. Springer, Cham. https://doi.org/10.1007/978-3-030-89147-3_2

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. M. (2013). Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Academic Medicine: Journal of the Association of American Medical Colleges, 88*(6), 872–883. https://doi.org/10.1097/ACM.0b013e31828ffdcf

Daniel, M., Rencic, J., Durning, S. J., Holmboe, E. S., Santen, S. A., Lang, V., Ratcliffe, T., Gordon, D., Heist, B., Lubarsky, S., Estrada, C. A., Ballard, T., Artino, A. R., Da Sergio Silva, A., Cleary, T., Stojan, J., & Gruppen, L. D. (2019). Clinical reasoning assessment methods: a scoping review and practical guidance. *Academic Medicine: Journal of the Association of American Medical Colleges, 94*(6), 902–912. https://doi.org/10.1097/ACM.0000000000002618

Dunbar, K. (1995). How scientists really reason: scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 365–395). MIT Press.

Edelsbrunner, P. A. (2024). *Does interference between intuitive conceptions and scientific concepts produce reliable inter-individual differences?* Science & Education. Advance online publication. https://doi.org/10.1007/s11191-024-00500-8

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

Fink, M. C., Heitzmann, N., Reitmeier, V., Siebeck, M., Fischer, F., & Fischer, M. R. (2023). Diagnosing virtual patients: the interplay between knowledge and diagnostic activities. *Advances in Health Sciences Education : Theory and Practice*, 1–20. https://doi.org/10.1007/s10459-023-10211-4

Fiore, S. M., Graesser, A. C., & Greiff, S. (2018). Collaborative problem-solving education for the twenty-first-century workforce. *Nature Human Behaviour, 2*(6), 367–369. https://doi.org/10.1038/s41562-018-0363-y

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. J., 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, 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

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M. R., Girwidz, R., Obersteiner, A., Reiss, K., Stürmer, K., Siebeck, M., Schmidmaier, R., Seidel, T., Ufer, S., Wecker, C., & Neuhaus, B. J. (2018). Systematizing professional knowledge of medical doctors and teachers: development of an

interdisciplinary framework in the context of diagnostic competences. *Education Sciences, 8*(4), 207. https://doi.org/10.3390/educsci8040207

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.), *Methodology of educational measurement and assessment. competence assessment in education* (pp. 407–425). Springer International Publishing. https://doi.org/10.1007/978-3-319-50030-0_24

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: A Journal of the American Psychological Society, 19*(2), 59–92. https://doi.org/10.1177/1529100618808244

Hautz, W. E., Kämmer, J. E., Schauber, S. K., Spies, C. D., & Gaissmaier, W. (2015). Diagnostic performance by medical students working individually or in teams. *JAMA, 313*(3), 303–304. 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. J., 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. *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

Hesse, F. W., 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* (pp. 37–56). Springer.

Hetmanek, A., Engelmann, K., Opitz, A., & Fischer, F. (2018). Beyond intelligence and domain knowledge. In F. Frank, C. Clark A., E. Katharina, O. Jonathan, F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific reasoning and argumentation* (pp. 203–226). Routledge. https://doi.org/10.4324/9780203731826-12

Hilbert, S., & Stadler, M. (2017). Structural equation models. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual differences* (pp. 1–9). Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1285-1

Hitchcock, D. (2005). Good reasoning on the Toulmin model. *Argumentation, 19*(3), 373–391. https://doi.org/10.1007/s10503-005-4422-y

Horst, P. (1941). The prediction of personnel adjustment. *Socia LScience Research and Council Bulletin* (48), 431–436.

Jeong, H., Hmelo-Silver, C. E., & Jo, K. (2019). Ten years of computer-supported collaborative learning: a meta-analysis of CSCL in STEM education during 2005–2014. *Educational Research Review, 28*, 100284. https://doi.org/10.1016/j.edurev.2019.100284

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

Kiesewetter, J., Sailer, M., Jung, V. M., Schönberger, R., Bauer, E., Zottmann, J. M., Hege, I., Zimmermann, H., Fischer, F., & Fischer, M. R. (2020). Learning clinical reasoning: how virtual patient case format and prior knowledge interact. *BMC Medical Education, 20*(1), 73–83. https://doi.org/10.1186/s12909-020-1987-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

Koschmann, T. D., Feltovich, P. J., Myers, A. C., & Barrows, H. S. (1992). Implications of CSCL for problem-based learning. *ACM SIGCUE Outlook, 21*(3), 32–35. https://doi.org/10.1145/130893.130902

Liu, L., Hao, J., Davier, A. A. von, Kyllonen, P., & Zapata-Rivera, J.-D. (2016). A tough nut to crack: measuring collaborative problem solving. In J. Keengwe, 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

Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical Analysis with Latent Variables: User's Guide* (Version 8) [Computer software]. Authors.

Nachtigall, C., Kroehne, U., Funke, F., & Steyer, R. (2003). (Why) should we use SEM? Pros and cons of structural equation modeling. *Methods of Psychological Research Online, 8*(2), 1–22.

Noroozi, O., Biemans, H. J., 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

OECD. (2017). PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, revised edition. PISA, OECD Publishing. https://doi.org/10.1787/9789264281820-en

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 *Proceedings of the International Conference on Computer-supported for Collaborative Learning, 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/cscl2023.638463

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

Radkowitsch, 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*. Springer, Cham. https://doi.org/10.1007/978-3-030-89147-3_10

Radkowitsch, 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

Richters, C., Stadler, M., Radkowitsch, 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 (Chair), International Society of the Learning Sciences. Hiroshima, Japan.

Rochelle, J., & Teasley, S. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer-Supported Collaborative Learning* (pp. 66–97). Springer.

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

Savalei, V., & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology, 66*(2), 201–223. https://doi.org/10.1111/j.2044-8317.2012.02049.x

Scalise, K., Mustafic, M., & Greiff, S. (2016). Dispositions for collaborative problem solving. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Methodology of Educational Measurement and Assessment. Assessing Contexts of Learning* (pp. 283–299). Springer International Publishing. https://doi.org/10.1007/978-3-319-45357-6_11

Schmidt, H. G., & Mamede, S. (2015). How to improve the teaching of clinical reasoning: a narrative review and a proposal. *Medical Education, 49*(10), 961–973. https://doi.org/10.1111/medu.12775

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

Schons, C., Obersteiner, A., Reinhold, F., Fischer, F., & Reiss, K. (2022). *Developing a simulation to foster prospective mathematics teachers' diagnostic competencies: the effects of scaffolding*. Advance online publication. https://doi.org/10.1007/s13138-022-00210-0

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: an investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education, 157*, 103964. https://doi.org/10.1016/j.compedu.2020.103964

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: log data indicates ability differences despite equal scores. *Computers in Human Behavior, 111*, 106442. https://doi.org/10.1016/j.chb.2020.106442

Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: a good alpha is not always better. *New Ideas in Psychology*, *60*. https://doi.org/10.1016/j.newideapsych.2020.100832

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

Stasser, G., & Titus, W. (1985). Pooling of unshared infomration in group decision making: biased information sampling during discussion. *Journal of Personality and Social Psychology, 48*(6), 1467–1478.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

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

van Joolingen, W. R., & de Jong, T. (1997). An extended dual search space model of scientific discovery learning. *Instructional Science, 25*(5), 307–346. https://doi.org/10.1023/A:1002993406499

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., Radkowitsch, A., Schmidmaier, R., . . . Noroozi, O. (2023). Transactivity and knowledge co-construction in collaborative problem solving. In *Proceedings of the International Conference on Computer-supported for Collaborative Learning, 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

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 New York. https://doi.org/10.1007/978-1-4612-4634-3_9

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71–95. https://doi.org/10.1016/j.compedu.2005.04.003

Please note that the data employed in this study have been used in previous publications (e.g., Brandl et al., 2021; Radkowitsch, et al., 2021; Richters et al., 2022). However, the research question and the results reported in this study are completely unique to this study. An initial version of this article is presented as a poster at ISLS 2024.

Authors and Affiliations

Laura Brandl[1] · Matthias Stadler[1,2] · Constanze Richters[1] · Anika Radkowitsch[3] · Martin R. Fischer[2] · Ralf Schmidmaier[4] · Frank Fischer[1]

✉ Laura Brandl
L.Brandl@psy.lmu.de

[1] Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 Munich, Germany

[2] Institute of Medical Education, LMU University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany

[3] IPN Leibniz Institute for Science and Mathematics Education, Department of Mathematics Education, Kiel, Germany

[4] Medizinische Klinik und Poliklinik IV, LMU University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany

PAPER 3

Simulation-Based Learning of Complex Skills: Predicting Performance With Theoretically Derived Process Features

4

Laura Brandl * Constanze Richters * Anika Radkowitsch *
Andreas Obersteiner * Martin R. Fischer *
Ralf Schmidmaier * Frank Fischer * Matthias Stadler

Reference: Brandl, L., Richters, C., Radkowitsch, A., Obersteiner, A., Fischer, M. R., Schmidmaier, R., Fischer, F., & Stadler, M. (2021). Simulation-Based Learning of Complex Skills: Predicting Performance With Theoretically Derived Process Features. Psychological Test and Assessment Modeling, 63(4), 542–560. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2021-4/PTAM__4-2021_6_kor.pdf

Psychological Test and Assessment Modeling, Volume 63, 2021 (4), 542-560

Simulation-Based Learning of Complex Skills: Predicting Performance With Theoretically Derived Process Features

Laura Brandl[1], Constanze Richters[1]; Anika Radkowitsch[2], Andreas Obersteiner[3], Martin R. Fischer[4], Ralf Schmidmaier[5], Frank Fischer[1], Matthias Stadler[1]*

Abstract

Simulation-based learning is often used to facilitate complex problem-solving skills, such as collaborative diagnostic reasoning (CDR). Simulations can be especially effective if additional instructional support is provided. However, adapting instructional support to the learners' needs remains a challenge when performance is only assessed as the outcome after using the simulation. Researchers are, therefore, increasingly interested in whether process data analyses can predict outcomes of simulated learning tasks and whether such analyses allow early identification of the need for support. This study developed a random forest classification model based on theoretically derived process indicators to predict success in a simulated learning environment. The context of the simulated learning environment was medicine. Internists interacted with a simulated radiologist to identify possible causes of an illness. Participants' CDR was conceptualized via log-data, coded on a broad, domain-general level for better generalizability. Results showed a satisfactory prediction rate for CDR performance, indicated by diagnostic accuracy. The model predicted accurate and inaccurate diagnoses and was therefore suitable for making statements about the performance by only using process data of CDR. The findings

[1] Department Psychologie, Ludwig-Maximilians-Universität München, Munich, Germany

[2] IPN - Leibniz Institut für die Pädagogik der Naturwissenschaften und Mathematik, Abteilung Didaktik der Mathematik, Kiel, Germany

[3] Heinz Nixdorf-Stiftungslehrstuhl für Didaktik der Mathematik, TUM School of Social Sciences and Technology, Technische Unversität München, Munich, Germany

[4] Institut für Didaktik und Ausbildungsforschung in der Medizin, LMU Klinikum, Ludwig-Maximilians-Universität München, Munich, Germany

[5] LMU Klinikum, Medizinische Klinik und Poliklinik IV, Ludwig-Maximilians-Universität München, Munich, Germany

*Correspondence concerning this article should be addressed to Laura Brandl, Leopoldstr. 13, 80802 München, Germany. E-Mail: L.Brandl@psy.lmu.de

contribute to the development of more adaptive instructional support within simulation-based learning through being able to predict the individuals' learning outcomes already during the process.

Keywords: simulation-based learning, complex problem solving, learning analytics, process-based performance prediction, adaptive instructional support

Simulation-based learning is thought to facilitate complex problem-solving skills (Chernikova, Heitzmann, Fink, et al., 2020). Simulations represent relevant aspects of real-life problems (Grossman, 2021) and can be especially effective if they provide adaptive instructional support (Leutner, 1993). Adaptivity of instructional support is understood as the provision of support adjusted to individuals' specific needs. The aim of adaptive instructional support is twofold: Enhancing learning outcomes and enhancing self-regulation skills concerning learning processes. When a simulation can identify the needs of learners to better self-regulate their learning process and provide adaptive instructional support accordingly, this can allow learners to progress in their learning more efficiently than with non-adaptive support (Plass & Pawar, 2020).

Methods from the field of Learning Analytics seem to be helpful to enable adaptive instructional support because they focus on predicting future outcomes based on behavioral data during the assessment or training process, rather than solely observing the outcome of assessments (Baker & Siemens, 2014). One application of Learning Analytics is the prediction of learning performance using process data, thereby identifying learners at risk of showing inadequate performance (e.g., Gašević, Jovanovic, Pardo, & Dawson, 2017). The present study aims to apply Learning Analytics to the context of collaborative diagnostic reasoning (CDR) in simulation-based learning environments. CDR is an example of a complex problem-solving skill (Fiore et al., 2018) and refers to individual and collaborative skills that enable diagnosticians to diagnose problem states of specific systems (e.g., patients) while working together in teams, based on their conceptual and strategic knowledge (Radkowitsch, M.-R. Fischer, Schmidmaier, & F. Fischer, 2020).

Particularly, we predict CDR performance, indicated by diagnostic accuracy, based on the collaborative diagnostic process derived from existing theoretical models. In addressing this goal, the study serves as preparatory research for developing more adaptive instructional support within simulation-based learning.

Simulation-Based Learning of Complex Skills

Although most complex tasks require intensive training to be performed expertly, many are not easily accessible as training situations as they may be scarce (e.g., natural disasters) or too critical to be approached by novices (e.g., some medical procedures). Simulation-based learning enables the deliberate practice of complex tasks that learners cannot solve (Ericsson, 2004), with the opportunity to provide additional

instructional support. It represents a promising instructional approach to facilitate the development of complex skills by providing authentic situations approximating real-life diagnostic problems (Cook, Brydges, Zendejas, Hamstra, & Hatala, 2013; Heitzmann et al., 2019). As Chernikova, Heitzmann, Stadler, et al. (2020) report in a recent meta-analysis, simulation-based learning significantly fosters complex problem-solving skills.

Complex problem solving is a multidimensional set of skills needed to solve complex problems (Dörner & Funke, 2017). Complex problems require active knowledge acquisition to create a mental representation of the problem (Stadler, Niepel & Greiff, 2019). If complex problems are solved with another person or simulated agent, this process is called collaborative problem solving (Fiore et al., 2018; Stadler, Herborn, Mustafić & Greiff, 2020). One example is CDR, which can be conceptualized as the set of skills to solve a problem, such as diagnosing a patient, "by generating and evaluating evidences and hypotheses that can be shared with, elicited from, or negotiated among collaborators" based on their conceptual and strategic knowledge (Radkowitsch et al., 2020, p. 2). The first entails declarative knowledge about constructs (e.g., diagnoses and symptoms) and their relation, the second is about knowledge of how to apply strategic knowledge through problem-solving (Stark, Kopp & M.-R. Fischer, 2011). The goal of CDR is to reduce the uncertainty of decision-making by diagnosing a phenomenon, such as a patient's symptoms, in a collaborative effort. As such, CDR requires individual diagnostic as well as collaborative processes. To successfully solve a diagnostic problem, diagnosticians draw inferences from latent or hidden patterns of a phenomenon based on their current knowledge (Heitzmann et al., 2019). Heitzmann et al. (2019) described the process of *individual* diagnosing using the scientific reasoning and argumentation framework by F. Fischer et al. (2014), stating that, similar to scientific reasoning, diagnosing can be described with eight epistemic activities (e.g., evidence evaluation, evidence generation, hypothesis generation). In an attempt to extend these considerations to *collaborative* diagnostic processes, Radkowitsch et al. (2020) proposed the CDR model. The CDR model is based on the scientific discovery as dual search model by Klahr and Dunbar (1988) and its further development by Van Joolingen and De Jong (1997) and describes how individual diagnostic processes (F. Fischer et al. 2014) and collaborative activities (Liu et al., 2015) interact with each other. Liu and colleagues (2015) suggest four social skills (sharing ideas, negotiating ideas, regulating problem-solving, and maintaining communication) to describe collaborative activities. One of the main functions of the collaborative activities is to construct a shared problem representation (Roschelle & Teasley, 1995) through sharing and eliciting relevant information, as information might not be distributed equally between all collaborators. Hence, it is crucial to accurately share all relevant information to diagnose the patient‘s illness. These activities seem particularly relevant in a field such as medicine in which physicians from different fields of expertise collaborate frequently. In such situations, it is crucial for an accurate diagnosis of the patient's problem that all relevant evidence and hypotheses for the specific collaborators are shared (Kiesewetter, F. Fischer , & M.-R. Fischer, 2017).

The CDR model specifies such collaborative diagnostic processes by suggesting collaborative diagnostic activities (CDAs). CDAs combine individual and collaborative diagnostic activities such as evidence elicitation, evidence sharing, and hypotheses sharing. Evidence and hypotheses, which are results of individual diagnostic processes and stored in an individual's cognitive storage (see Klahr & Dunbahr, 1988), can become part of *collaborative* cognitive processes by, for instance, sharing or eliciting them. In the medical context, evidence is, for example, patient information about symptoms and other parameters which are identified as relevant for a diagnosis. A hypothesis is a suspected diagnosis that refers to an underlying illness that could explain the patient´s symptoms. Evidence elicitation is, then, the activity of collaboratively generating new information, for example, by conducting medical examinations like radiological tests (Radkowitsch et al., 2020). Adequate performance of CDR in the context of medicine is defined as performing those activities with high quality resulting in an accurate diagnosis (Tschan et al., 2009). However, there is currently no assumption about the linearity and sequence of the performance of CDAs required to reach an accurate diagnosis, and not all CDAs might be necessary for all collaborative diagnostic scenarios.

In summary, simulation-based learning offers a promising approach for the training of complex problem-solving skills, such as CDR, by providing authentic diagnostic situations for learners to engage in (Chernikova, Heitzmann, Stadler, et al., 2020; H. G. Schmidt & Rikers, 2007) while allowing to provide adequate instructional support. However, adapting these support measures (such as prompts or worked-out examples; Belland, 2017) to the learners' needs remains a challenge because it requires assessing the learner's current knowledge during the simulation rather than after using the simulation. Analyzing data stemming from the CDR process to inform a learner model (Ding, Zhu, & Guo, 2018) while the learner is still working on the simulation might lead to more timely support when necessary.

Learning Analytics and Process Data in Simulation-Based Learning

Using technologically-enhanced simulations that store data on the learning process immediately in log-files allows analyzing process data without the need for additional assessments with dedicated tests. Analyzing process data instead of only product data (the assessment's outcome) allows insights into the process leading to the eventual outcome (e.g., Goldhammer, Naumann, Rölke, Stelter, & Tóth, 2017). Widely used process data is often not at all straightforward to interpret. For example, more time spent on a task may indicate cognitive factors (i.e., the tasks are challenging) or motivational factors (i.e., tedious tasks). Nevertheless, process data analyses can increase understanding of the analyzed process (Greiff, Niepel, Scherer, & Martin, 2016). The results can be used to improve the theoretical understanding of the processes involved and approaches to assessing and facilitating them (Goldhammer, Naumann, Stelter, Tóth, Rölke & Klieme, 2014).

Using process data allows for the prediction of performance, enabling researchers to identify learners at risk to show inadequate performance, such as to benefit little from engaging in a learning activity (e.g., Leitner, Khalil, & Ebner, 2017), and to provide them with additional instructional support (e.g., scaffolding, Tabak & Kyza, 2018). Such support is ideally timed and adapted to the learners' needs (Plass & Pawar, 2020). Previous research has shown that the number of clicks and the time on task can be predictive for task success (Goldhammer et al., 2017). Stadler, Hofer, and Greiff (2020) analyzed differences between the time-on-task and the number of clicks of participants having the same outcome in a simulation of complex problem-solving. Despite having equal scores, participants differed in both time-on-task and number of clicks. The results indicate that process indicators depict individual differences in the ability not depicted in product data. This illustrates the need to take process data into account to assess learners' abilities. This is also in line with the assumption that complex problem-solving is not only about a task's outcome but also about the process to get there (Dörner & Funke, 2017).

However, it is difficult to deduct information on specific problems a learner might have with a task or what instructional support might be beneficial using process data. Therefore, researchers have called for a more robust link from process data to learning theories to understand better and facilitate learning (Gašević, Dawson, & Siemens, 2015). The identification of suitable features for the prediction of learning outcomes within process data should always be supported with theoretical models (Tomasevic, Gvozdenovic, & Vranes, 2020) in order to make findings replicable and generalizable beyond idiosyncratic learning environments.

Goal and Research Question

The current study uses activities theoretical derived from the CDR model (Radkowitsch et al., 2020) and constructed from process data to predict the performance of complex problem-solving skills, such as CDR, in simulation-based learning. It addresses the research question to what extent theoretically derived process indicators are suitable to predict learners' diagnostic accuracy in the context of simulation-based learning of CDR. Since CDR frequently occurs in medical settings and has been identified to be a significant challenge for physicians (e.g., Tschan et al., 2009; Brady et al., 2012), the simulation was embedded in the context of medical education and developed based on the CDR model. Three CDAs proposed in the CDR model are particularly relevant in the simulated situation: evidence elicitation, evidence sharing, and hypotheses sharing. Hence, the current study investigates to what extent diagnostic accuracy can be predicted using the CDAs constructed from process data of a simulated learning environment in medical education. In addressing this research question, the current study contributes to developing more adaptive instructional support within simulation-based learning through showing the possibilities of learning analytics methods, being able to predict the outcome already in the process.

Method

Simulation and Learner Task

The simulation was integrated into CASUS (https://www.instruct.eu/; M. R. Fischer, Aulinger, & Baehring, 1999), a case-based learning platform, where learners worked on five different patient cases within the simulation. Medical experts from internal medicine, radiology, and general medicine constructed the patient cases. In the simulation, the learners' task was to interact with an agent-based (i.e., simulated) radiologist to diagnose fictitious patient cases suffering from unknown fever. To that end, learners requested further information about the patient from the radiologist who conducted radiological examinations. This required learners to engage in the CDA evidence elicitation, evidence sharing, and hypotheses sharing. Medical experts who supported the development of the situation considered these CDAs as particularly important for the specific collaborative diagnostic situation. The collaboration took place after the learners studied a health record (containing all current information about the patient). The collaboration consisted of filling out a radiological request form and receiving the requested results from the simulated radiologist only if the request form contained sufficient evidence and hypotheses relevant for the radiologist to conduct and interpret the radiologic test. Specifically, learners needed to elicit evidence by choosing an exam method to be performed by the radiologist, sharing evidence by choosing information from the health record relevant for the radiologist, and sharing suspected diagnoses as hypotheses. After the collaboration, learners were asked to indicate their final diagnosis individually. For a detailed description of the simulation's development and validation, see Radkowitsch et al. (2020).

Sample and Design

Data for this study was taken from a more extensive experimental study conducted within the COSIMA Project. The study's design was an experimental setting with four groups investigating the effect of different kinds of instructional support. One group received an adaptive collaboration script; one was encouraged to have reflection phases, one both kinds of support, and the control group received none of them. In order to avoid confounding effects of the experimental conditions for the current study, only the control condition was used for the current analyses. Data was collected online from 9 male and 26 female intermediate learners from the 4^{th} – 6^{th} year of medical studies. In total, the study program includes six years of studying. Learners had an average age of $M = 25.43$ years ($SD = 2.54$ years) and studied medicine on average in their $M = 5^{th}$ year ($SD = 0.76$ years). Learners were recruited through an email distribution list and flyers. For full participation, learners received 10€ compensation per hour of testing. In line with the university's ethics requirements, participation was voluntary, and learners could terminate participation at any time. Given the focus of the study on the CDR process, the unit of analysis was the patient case and

not the participant. As learners worked on five patient cases, this led to a total of $n = 167$ after excluding missing data on diagnostic accuracy. The ethics committee of the medical faculty of LMU Munich declared ethical clearance prior to data collection (approval number 18-262).

Measures

Diagnostic Accuracy

Each patient case is assigned to one primary diagnosis, consented by experts. After working on the patient case and requesting a radiological examination, the learners indicated their final diagnosis using a free text field with suggested options out of a list of 249 diagnoses, based on the first letters entered, to shorten and standardize the input. Diagnostic accuracy was calculated by coding the final diagnosis's compliance with the expert solution. To that end, two independent coders each coded the complete data. Differences in the coding were discussed until all codes were identical. Accurate diagnoses were coded with 1, while inaccurate diagnoses were coded with 0. For example, when the patient suffers from hospital-acquired pneumonia, this diagnosis would be coded with 1, while only pneumonia or any other diagnosis would be coded with 0.

Process Data

Every click in the simulation leading to an interaction with the system was stored with the corresponding timestamp in log file data allowing for analyzing process data. Based on the CDR model, the CDAs were coded depending on the learners' entries to a radiological request form during the collaboration with the simulated radiologist. Every activity where the learners selected a radiological examination by choosing a method and the body part to examine was coded as *evidence elicitation*. Every activity where the learners shared information from the health record to justify the radiological examination was coded as *evidence sharing*. Every activity where the learners indicated a potential diagnosis was coded as *hypotheses sharing*. Diagnoses were entered using a free text field with suggested options out of a list of 249 diagnoses, based on the first letters entered, to shorten and standardize the input. To illustrate this process, we will give an example of how a learner could have filled out the request form: The learner started to fill out the request form by choosing an x-ray of the chest as a radiological examination (evidence elicitation). This requires the learner to make two clicks in the simulation, one for selecting a method and another for selecting the respective body part. Next, the learner justified the decision for the examination method by ticking information presented in the health record (evidence sharing). In this example, the learner shared that the patient has decreased breathing sound, fever, is

male, and is a smoker. The learner identified and ticked the respective box to share that evidence, including this information. Lastly, the learner typed 'pneu' into the free text field on the bottom of the form. The system offered possible diagnoses starting with 'pneu' (e.g., pneumonia; community-acquired pneumonia; hospital-acquired pneumonia), the learner chose the share 'pneumonia' as a hypothesis with the simulated radiologist. Before sending the form, the learner decided to additionally share the evidence that the patient has an increased lymphocyte value.

First, the clicks in the simulation were coded automatically according to the CDAs using spreadsheet software. Then, each coded activity was decomposed into the number of seconds a participant spent on the activity. The activities coded in units of seconds were then summarized into behavioral strings that indicated, per learner and case, which CDA was performed, how long, and what activity followed. This information was stored in a string variable.

Analyses

The proper selection of features is essential in prediction models. When process data depicts long sequences, exploratory approaches such as the *n-gram* method proposed by Damashek (1995) can be helpful. Here the process of activities is summarized as a sequence of n consecutive elements. This allows representing the sequence of activities as well as their frequency. For this study, we chose bigrams (n = 2) to ensure there are not too many different features in our prediction models. The bigrams represented either consistent activity (two instances of the same activity) or transitions from one behavior to another (two different activities). To apply the n-gram method, the string variable representing an individual's sequence of activities was separated in bigrams using the *n-gram* package in R (3.0.4; D. Schmidt & Heckendorf, 2017), leading to nine features constructed from the three theoretical derived activities, each summarizing how often this specific bigram occurred in the string variable.

Referring back to the previous example, the learner spent 60 seconds on evidence elicitation, which resulted in 59 instances of the EE.EE bigram. Further, the learner spent 200 seconds at the beginning and 6 seconds with evidence sharing when they returned to that activity after sharing the hypothesis resulting in 204 instances of the ES.ES bigram. Spending 150 seconds with hypotheses sharing results in 149 instances of the HS.HS bigram. Those three bigrams indicate consistent activity. Looking at transitions, the learner had a value of one on the bigrams EE.ES, ES.HS, HS.ES indicating changes between evidence elicitation and evidence sharing, evidence sharing and hypotheses sharing, as well as hypotheses sharing and evidence sharing, respectively.

For predicting diagnostic accuracy using bigrams of CDAs, the statistical software R (RStudio Team, 2020) was used. The essential packages were *ranger* (0.12.1; Wright & Ziegler, 2017) and *caret* (6.0-86; Kuhn, 2008). A random forest classification model (ranger algorithm; Wright & Ziegler, 2017) was developed to answer the research question. This model was chosen as it is highly accurate and able to deal with

relatively large numbers of features and few data points while considering complex interactions among the features. In contrast to more interpretable logistic regression models, random forest classification models are also less affected by multicollinearity issues (Breiman, 2001; Fernández-Delgado, Cernadas, Barro, & Amorim, 2014).

First, the data set was split into a training set (including 75 % of the data) and a testing set (including 25 % of the data). The training set was then used to fit the prediction model. To increase the model fit, hyperparameters were tuned automatically. A 10x3 cross-validation was applied to identify the hyperparameters to decrease the risk of overfitting. For the ranger algorithm, only the number of randomly selected predictors (mtry), the split rule (gini or extra trees), and the minimum node size needed to be determined (Kuhn, 2008). The prediction model was evaluated in the testing set and the training set using a confusion matrix (Buskirk, Kirchner, Eck, & Signorino, 2018). To assess classification quality of the prediction model classification accuracy (the total percentage of correct classifications), sensitivity (true positive classification relative to all positive classifications), and specificity (true negative classification relative to all negative specification), no-information rate (always predicting the most common class), and a one-sided significance test to see whether the developed model outperforms the no-information rate was evaluated (Alpaydin, 2010; Kuhn, 2008). Kappa, the agreement between predicted values and the actual data in relation to expected values by chance, is assessed, with a value of greater than .61 indicating sufficient strength of agreement (Landis & Koch, 1977).

Finally, a closer look into how each feature influenced the classification was done using feature importance. Due to complex interactions among different features, the interpretation of importance is not always straightforward and can only be done in relation to other features in the model, not by applying standardized cut off values (Kuhn, 2008; Liaw & Wiener, 2002; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). The dataset and the code for the analyses are uploaded to the open science framework (OSF) repository and can be retrieved from https://osf.io/y6bfx/

Results

Before looking at the predictability of diagnostic accuracy using process data, the used features are presented descriptively in Table 1. The bivariate correlation between the features and diagnostic accuracy is only minor, ranging from -.06 to .11.

Table 1

Descriptive Results of the Features used for Prediction Diagnostic Accuracy

Feature	Accurate Diagnoses		Inaccurate Diagnoses		r_p
	Median	Range	Median	Range	
EE.EE	25.5	6 - 342	36.0	2 - 429	-.05
ES.ES	146.0	0 - 587	135.5	0 - 581	.11
HS.HS	79.0	0 - 568	67.5	0 - 520	-.02
EE.ES	1.0	0 - 6	1.0	0 - 6	.01
EE.HS	0.0	0 - 5	0.0	0 - 4	.03
ES.EE	0.0	0 - 3	0.0	0 - 4	-.04
ES.HS	0.0	0 - 4	0.0	0 - 5	.05
HS.EE	0.0	0 - 4	1.0	0 - 4	-.02
HS.ES	0.0	0 - 2	0.0	0 - 3	.06

Note. EE = evidence elicitation, ES = evidence sharing, HS = hypotheses sharing

r_p = Pearson correlation between feature and diagnostic accuracy

Investigating the predictability of diagnostic accuracy using process indicators, depicted through bigrams of CDAs, the identified random forest classification model (mtry = 2, splitrule = extra trees, min node size = 1) performed well. Classification accuracy of .98 (95 % CI [.93; 1.00]) was found for the training set, indicating strong predictive power. The results of the one-sided hypothesis test indicated that the developed model was significantly better than the no-information rate model (accuracy of .54, $p < .001$). The kappa for the model was .95, implying high agreement between the predicted values by the model and the actual data (Landis & Koch, 1977). Further evaluation revealed a sensitivity of .95 and a specificity of 1.00, indicating that the model could correctly predict accurate and inaccurate diagnoses in most cases.

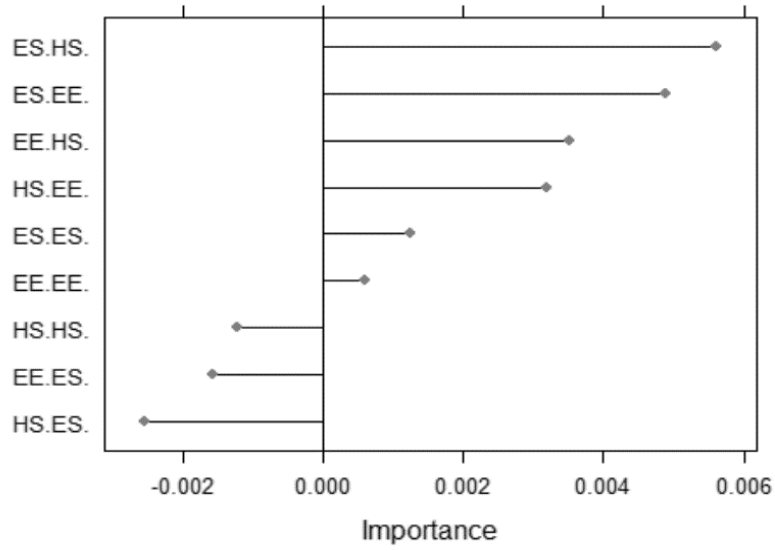
When using the testing set, the results supported the good ability of the model to predict diagnostic accuracy, with a predictive accuracy of .95 (95 % CI [.84;.99]) and a no-information rate of .73. The classification was also significantly better than the no-information model ($p <.001$) for the testing set. Results implied, again, a high agreement between the predicted values by the model that was trained based on the training

sample and the data of the testing sample with a kappa of .88. The additional evaluation metrics further indicated a sensitivity of .91 and specificity of .97, slightly worse than in the training set. Nevertheless, both measures indicated a high capacity of the model to predict accurate and inaccurate diagnoses in both the training and the testing data set.

Looking at the importance of the different features (see Figure 1), the most important one was the transition from evidence sharing to hypothesis sharing. This is followed by the transition from evidence sharing to evidence elicitation and the transition from evidence elicitation to hypotheses sharing. The fourth most important feature is the transition from hypothesis sharing to evidence elicitation. All those transitions are entailed in the process of CDR.

Figure 1

Importance of Features Predicting Diagnostic Accuracy Using Process Data



Note. EE = evidence elicitation, ES = evidence sharing, HS = hypotheses sharing

Discussion

The current study aimed at investigating to what extent theoretically derived process indicators are suitable to predict learners' diagnostic accuracy (performance measure) in the context of simulation-based learning of CDR. A random forest algorithm

classified accurate and inaccurate diagnoses correctly based on bigrams of CDAs. The model predicted a large percentage of accurate and inaccurate diagnoses and is, therefore, suitable to support statements about the performance only using process data. This is in line with former research (e.g., Mahboob, Irfan, & Karamat, 2016), indicating that algorithms from the field of Learning Analytics are suitable for performance prediction.

Learning performance and its enhancement are widely investigated in Learning Analytics (Leitner et al., 2017). However, most of the studies lacked a theoretical grounding of their approach (Gašević et al., 2015). The present study used features for the prediction of diagnostic accuracy that were derived from the CDR model by Radkowitsch et al. (2020), which is theoretically rooted in well established theoretical frameworks (e.g., Klahr & Dunbahr, 1988; F. Fischer et al., 2014; Liu et al., 2015). The current results underline the relevance of epistemic activities, such as CDAs, and their sequences for diagnostic processes. However, so far, the CDR model does not consider predictions about the relation between the CDAs and diagnostic accuracy. It is only conceptualizing CDAs as part of the CDR process, which needs to be performed with high quality to draw an accurate final decision. Using Learning Analytics, we showed that the CDAs are relevant for diagnostic accuracy, being a performance indicator of CDR, even though the bivariate correlations between the bigrams and diagnostic accuracy were only minor.

The clear benefit of using machine-learning prediction models instead of traditional statistical models is the change of perspective. While the latter is concerned about explaining causal relationships and therefore has a retrospective view on the data, the former has the goal of predicting future data and therefore has a prospective view (Yarkoni & Westfall, 2017). Accordingly, predictive accuracy is the primary goal, and the ratio of bias and variance, which minimize the occurring error the best, should be chosen. In order to achieve this, one must be willing to allow for bias and nonlinearity for the sake of accurate prediction (Molnar et al., 2020; Yarkoni & Westfall, 2017). This focus on predictive accuracy can make prediction models, especially ensemble methods such as random forests, highly complex, resulting in accurate predictions but lacking an explanation of how they were achieved, leading to less transparent models, also known as black boxes (Molnar et al., 2018; Yarkoni & Westfall, 2017). There is a need to investigate non-linear relations between process indicators to enhance theoretical models. The current results highlight the relevance of theoretically derived process indicators for the performance of CDR in simulation-based learning and can be used to predict the performance of complex problem-solving skills in simulation-based learning already in the process. Such predictions may help provide learners with inadequate performance with additional (adaptive) instructional support. From the feature importance plot, we can see that the consistent features (e.g., time spent with evidence elicitation) and transitions from evidence elicitation to evidence sharing and from hypotheses sharing to evidence sharing are relatively unimportant. Future analyses should therefore focus less on these processes and more on the transitions from evidence sharing to hypotheses sharing, from evidence sharing to evidence elicitation and from evidence elicitation to hypotheses sharing and hypotheses

sharing to evidence elicitation. The most important feature is the transition from evidence sharing to hypotheses sharing. However, the feature did not differ considerably regarding accurate and inaccurate diagnoses. Therefore, a non-linear relationship or a complex interaction with one or more other features is assumed, which needs to be further investigated. However, we currently do not know precisely what indicates inadequate performance and how to foster it accordingly, as the interpretation of black-box models and feature importance is not straightforward, and the prediction is not linear but a result of complex interactions.

Nevertheless, the current study was able to show that predicting the performance in complex simulation-based learning environments based on theoretically derived indicators of behavior is possible, even if there are no linear correlations between behavior and performance. Since we were able to demonstrate a relation between the theoretically derived process indicators and the performance of CDR in simulation-based learning, the next step should be to investigate sequences of activities in depth, e.g., with sequence clustering (Piccarreta, 2017), allowing not only to identify learners who need additional instructional support but also to provide this support.

The current results are not limited to learning of CDR in the medical context but likely generalize to related fields such as teacher education (Heitzmann et al., 2019) and complex problem-solving skills in different domains as the indicators of behavior were coded on a domain-general level.

Limitations and Future Research

The current study is not without limitations, which must be kept in mind when interpreting the results. First, it must be considered that all patient cases were analyzed independently, regardless of the order in which they appeared in the simulation, thus ignoring potential learning effects between the cases. Statistically speaking, this approach risks ignoring non-negligible random effects due to the clustered nature of the data. Extensions of the random forest algorithm have been proposed that consider clustered data (Hajjem, Bellavance, & Larocque, 2014). However, since our model performed exceptionally well, the intra-class correlation among participants is likely very low even without this extension. Another limitation is that only data from learners with an intermediate level of expertise was collected, limiting the observation of full expert and novice behavior. However, data showed a balanced frequency of accurate and inaccurate diagnoses. Future research might investigate whether participants of different expertise levels employ different strategies for their collaborative diagnosing, which would likely require an algorithm capable of including this information as an additional level of data.

Another potential limitation lies in the decision to observe only bigrams rather than n-grams that are more complex. N-grams that are more complex might provide further insights into more advanced strategies and might be more interpretable towards necessary support. However, the number of features increases exponentially with the length of observed n-grams. Even trigrams might have resulted in too many ($3^3 = 27$)

features for our limited sample. Future research might investigate longer n-grams using larger samples. An alternative would be the theoretical definition of specific sequences as predictors to explicitly test hypotheses on strategic behavior in a simulation. In line with the current results, the focus should be on the transitions between activities rather than on consistent behavior.

To help learners who potentially show inadequate performance as early as possible, future research will also need to investigate how early it is possible to predict the performance of complex problem-solving skills using process data. In addition, future research may also investigate how additional instructional support could look like. For example, Azevedo, Moos, Cromley, and Greene (2011) demonstrated that a combination of content and process-oriented adaptive scaffolding is suitable to facilitate self-regulated learning.

Currently, there is only little known about sequences of CDAs and their relation with diagnostic accuracy. However, we could show in this study that there are non-linear relations between those process indicators and learning performance. Future research should deepen this by investigating the transitions between activities to make further claims on refining existing theoretical process models. This is in line with the call for *explanatory learner models* that focus on optimal predictions using black-box models but use more interpretable methods to gain deeper insights into learning (Rosé et al., 2019). One approach in this context is the use of different kinds of data, such as process (e.g., log-file data), product (e.g., the outcome of a task), and learner data (e.g., self-report measures) using dispositional learning analytics (Buckingham Shum & Crick, 2021). This combination of data sources allows improving the design of adaptive scaffolding and interventions as it provides more profound insights into the origins of underperforming (Gašević et al., 2017). For example, Tempelaar, Rienties, and Nguyen (2021) combined this approach with a person-oriented type of research (instead of the traditional variable-oriented type) to identify five different learning profiles based on only learner data at the beginning and then by including more process data in a stepwise manner. This allows providing instructional support not only for a group of learners or an average learner but also for a specific individual learner, that is, personalized learning support.

Conclusion

This study aimed to predict CDR performance using process data, indicated by diagnostic accuracy. Results show that using a Learning Analytics approach, a random forest prediction model, is suitable for predicting performance using process indicators theoretically derived and constructed from process data. Using Learning Analytics enables researchers to provide practical solutions such as identifying learners at risk to show inadequate performance in need of adaptive instructional support. The findings contribute to the development of more adaptive instructional support within simulation-based learning through being able to predict the individuals' learning outcomes already during the process.

Acknowledgments

This research was supported by a grant from the Deutsche Forschungsgemeinschaft DFG (COSIMA; DFG-Forschungsgruppe 2385).

References

Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). Adaptive computation and machine learning. Cambridge Mass.: MIT Press.

Azevedo, R., Moos, D. C., Cromley, J. G., & Greene, J. A. (2011). Adaptive Content and Process Scaffolding: A key to facilitating students' self-regulated learning with hypermedia. Psychological Test and Assessment Modeling, 53(1), 106–140.

Baker, R., & Siemens, G. (2014). Educational Data Mining and Learning Analytics. In R. K. Sawyer (Ed.), The Cambridge Handbook of the Learning Sciences (pp. 253–272). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.016

Belland, B. R. (2017). Instructional Scaffolding in STEM Education. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-02565-0

Brady, A., Laoide, R. Ó., McCarthy, P., & McDermott, R. (2012). Discrepancy and error in radiology: Concepts, causes and consequences. Ulster Medical Journal, 81(1), 3.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Buckingham Shum, S., & Deakin Crick, R. (2012). Learning dispositions and transferable competencies. In S. Dawson & C. Haythornthwaite (Eds.), Proceedings of the 2nd international conference on learning analytics and knowledge - lak '12 (pp. 1–10). ACM Press. https://doi.org/10.1145/2330601.2330629

Buskirk, T. D., Kirchner, A., Eck, A., & Signorino, C. S. (2018). An Introduction to Machine Learning Methods for Survey Researchers. Survey Practice, 11(1), 1–10. https://doi.org/10.29115/SP-2018-0004

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., & Fischer, F. (2020). Facilitating Diagnostic Competencies 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

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. M. (2013). Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. Academic Medicine: Journal of the Association of American Medical Colleges, 88(6), 872–883. https://doi.org/10.1097/ACM.0b013e31828ffdcf

Damashek, M. (1995). Gauging Similarity with n-Grams: Language-Independent Categorization of Text. Science (New York, N.Y.), 267(5199), 843–848. https://doi.org/10.1126/science.267.5199.843

Ding, W., Zhu, Z., & Guo, Q. (2018). A New Learner Model in Adaptive Learning System. In 3rd International Conference on Computer and Communication Systems (ICCCS) (pp. 440–443). IEEE. https://doi.org/10.1109/CCOMS.2018.8463316

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

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, 3133–3181.

Fiore, S. M., Graesser, A., & Greiff, S. (2018). Collaborative problem-solving education for the twenty-first-century workforce. Nature Human Behaviour, 2(6), 367–369. https://doi.org/10.1038/s41562-018-0363-y

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., . . . 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, M. R., Aulinger, B., & Baehring, T. (1999). Computer-based-Training (CBT). Fallorientiertes Lernen am PC mit dem CASUS/ProMediWeb-System [Computer-based training (CBT). Case-oriented learning on the PC with CASUS/ProMediWeb System]. Deutsche medizinische Wochenschrift (1946), 124(46), 1401. https://doi.org/10.1055/s-2007-1024550

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

Gašević, D., Jovanovic, J., Pardo, A., & Dawson, S. (2017). Detecting Learning Strategies with Analytics: Links with Self-reported Measures and Academic Performance. Journal of Learning Analytics, 4(2), 113–128. https://doi.org/10.18608/jla.2017.42.10

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.), Methodology of Educational Measurement and Assessment. Competency Assessment in Education (pp. 407–425). Cham: Springer International Publishing. 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

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

Grossman, P. (2021). Teaching core practices in teacher education. Harvard Education Press.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation, 84(6), 1313–1328. https://doi.org/10.1080/00949655.2012.741599

Heitzmann, N., Seidel, T., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., . . . Opitz, A. (2019). Facilitating Diagnostic Competences in Simulations in Higher Education A Framework and a Research Agenda. Frontline Learning Research, 1–24. https://doi.org/10.14786/flr.v7i4.384

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

Klahr, D., & Dunbar, K. (1988). Dual Space Search During Scientific Reasoning. Cognitive Science, 12(1), 1–48. https://doi.org/10.1207/s15516709cog1201_1

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software,, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics. (33), 159–174.

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning Analytics in Higher Education—A Literature Review. In A. Peña-Ayala (Ed.), Studies in Systems, Decision and Control. Learning Analytics: Fundaments, Applications, and Trends (Vol. 94, pp. 1–23). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-52977-6_1

Leutner, D. (1993). Guided discovery learning with computer-based simulation games: Effects of adaptive and non-adaptive instructional support. Learning and Instruction, 3(2), 113–132. https://doi.org/10.1016/0959-4752(93)90011-N

Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. R News. 2(3), 18–22.

Liu, L., Hao, J., Davier, A. von, Kyllonen, P., & Zapata-Rivera, D. (2015). 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). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.c

Mahboob, T., Irfan, S., & Karamat, A. (2016). A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms, 2016 19th International Multi-Topic Conference (INMIC), pp. 1-8. https://doi.org/10.1109/INMIC.2016.7840094.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning – a brief history, state-of-the-art and challenges. In I. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Guidotti, Ö. Özgöbek, R. P. Ribeiro, R. Gavaldà, J. Gama, L. Adilova, Y. Krishnamurthy, P. M. Ferreira, D. Malerba, I. Medeiros, M. Ceci, G. Manco, E. Masciari, . . . J. A. Gulla (Eds.), Communications in Computer and Information Science. ECML PKDD 2020 Workshops (Vol. 1323, pp. 417–431). Springer International Publishing. https://doi.org/10.1007/978-3-030-65965-3_28

Piccarreta (2017). Joint Sequence Analysis: Association and Clustering. Sociological Methods and Research, 46(2), 252–287. doi: 10.1177/0049124115591013.

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

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

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

Rosé, C. P., McLaughlin, E. A., Liu, R., and Koedinger, K. R. (2019). Explanatory Learner Models: Why Machine Learning (Alone) Is Not the Answer. Br. J. Educ. Technol. 50, 2943–2958. https://doi.org/10.1111/bjet.1285

RStudio Team (2020). RStudio: Integrated Development for R. RStudio. Bosten, MA: PBC: PBC. Retrieved from http://www.rstudio.com/

Schmidt, D., & Heckendorf, C. (2017). ngram: Fast n-Gram Tokenization. Retrieved from https://cran.r-project.org/package=ngram

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

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. Computers & Education, 157, 103964.

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. Computers in Human Behavior, 111, 106442. https://doi.org/10.1016/j.chb.2020.106442

Stadler, M., Niepel, C., & Greiff, S. (2019). Differentiating between static and complex problems: A theoretical framework and its empirical validation. Intelligence, 72, 1-12. https://doi.org/10.1016/j.intell.2018.11.003

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

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8. https://doi.org/10.1186/1471-2105-8-25

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 (pp. 191–200). New York, NY : Routledge, 2018.: Routledge. https://doi.org/10.4324/9781315617572-19

Tempelaar, D., Rienties, B. & Nguyen, Q. (2021). Dispositional Learning Analytics for Supporting Individualized Learning Feedback. Frontiers in Education, 6. https://doi.org/10.3389/feduc.2021.703773

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & Education, 143, 103676. 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

Van Joolingen, W. R., & De Jong, T. (1997). An extended dual search space model of scientific discovery learning. Instructional Science, 25(5), 307-346. https://doi.org/10.1023/A:1002993406499

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: A Journal of the Association for Psychological Science, 12(6), 1–23. https://doi.org/10.1177/1745691617693393

GENERAL DISCUSSION

5

Laura Brandl

The objective of this thesis was to facilitate the leveraging of process data for the assessment and support of collaborative problem-solving. Two sub-goals were addressed: This thesis examined the potential of process data analyses to (1) facilitate theoretical advancements and (2) inform learning and instruction of collaborative problem-solving in the context of collaborative diagnostic reasoning. To achieve these goals, three papers were presented, focusing on different but related aspects. This chapter provides a brief summary of the results of each paper, followed by a joint discussion of the theoretical and practical implications. The transferability of these implications is then reflected upon, before concluding with a discussion of the limitations of the presented papers and suggestions for future research.

5.1 Summary of Results

The first paper (Stadler et al., 2023), a theoretical perspective paper, adopted a meta-perspective to discuss recent and current developments in the use of process data in large-scale assessments, as well as the scientific, practical, and policy-level issues that impede sustainable use. From a scientific standpoint, the findings from process data analyses are currently not widely generalizable due to their task-specific nature. Furthermore, there is a lack of replication studies, which hinders the establishment of robust evidence. The utilization of higher-level features (see 1.4.1) has the potential to yield robust evidence that is applicable across different studies, thereby facilitating the possibility of conceptual replication even when items differ between studies. From a practical standpoint, the utilization of process data has resulted in a transition from a purely summative assessment approach to a more formative one, with an emphasis on providing feedback (see 1.2.3; Pellegrino et al., 2001). For instance, with regard to the CoSiMed simulation (see 1.3.3), the way in which learners complete the radiological request form leads to a response by the agent-based radiologist, which can be perceived as a form of feedback. It can be argued that assessment tasks can provide opportunities for learning when feedback is provided. This has led to an emerging call to use process data as a measure of ability that goes beyond the ability to solve the problem. Additionally, it is argued that such process data can be used to provide individualized instructional support, e.g. adaptive scaffolding (e.g. Azevedo et al., 2004). This allows learners to benefit from the interaction with the assessment task. Nevertheless, in order to leverage interactive tasks for the purpose of identifying individuals in need of instructional support and facilitating personalized learning and adaptive support (see 1.4.2), it is imperative to ensure a coherent alignment between the design of assessment and the instructional design. In addition, at the policy level, the leveraging of process data from large-scale assessments enables a shift in focus from product data to the processes that contribute to the generation of these products (see 1.4.1), thereby facilitating the formulation of

informed educational decisions. However, to ensure sustainable use, it is essential to consider the issues outlined at the scientific and practical levels, particularly when utilizing process data from large-scale assessments, and to take cultural differences in learning and teaching into account in educational decision-making. In summary, there are necessary changes to be undertaken at the scientific level in how process data are analyzed to foster sustainable changes at the practical and policy levels. First and foremost, establishing a connection between process data and educational theory is vital for enhancing the generalizability of our findings and, consequently, facilitating theoretical advancements. Secondly, there is a need to align the design of assessment with that of instructional design, with the aim of informing learning and instruction.

Building on these considerations, the second paper (Brandl et al., 2024) aimed to investigate how process data analyses can facilitate theoretical advancements, particularly in the context of validating theoretical models in educational research using quality measures derived from high-level features constructed from collaborative diagnostic activities. The objective of this empirical study was to evaluate the CDR-M (Radkowitsch et al., 2022) in a simulation-based environment by analyzing data from three studies in the medical domain. The CDR-M describes the relations between individual characteristics (i.e., content knowledge, collaboration knowledge, and social skills), collaborative diagnostic activities (i.e., evidence elicitation, evidence sharing, and hypotheses sharing), and diagnostic outcomes (i.e., diagnostic accuracy, diagnostic justification, and diagnostic efficiency). The results indicate that the hypothesized relations in the CDR-M can be partially applied across studies. Content knowledge enables the diagnostician to formulate an initial suspected diagnosis, which is likely to be relevant information for the collaboration partner and to guide the subsequent collaborative diagnostic activities effectively. The observed relation between content knowledge and the quality of evidence elicitation, but not the other collaborative diagnostic activities, can be explained by the fact that evidence elicitation represents the least transactive collaborative diagnostic activity within the collaborative diagnostic reasoning process. In the process of evidence elicitation, the collaboration partner is utilized as an external knowledge resource, requiring minimal collaborative effort (Weinberger & Fischer, 2006). This perspective on the transactivity of collaborative diagnostic activities is further reinforced by the necessity to differentiate between primary (i.e., diagnostic accuracy) and secondary (i.e., diagnostic justification and diagnostic efficiency) outcomes of diagnostic reasoning (Daniel et al., 2019). Achieving diagnostic accuracy, which is related to evidence elicitation, the least transactive collaborative diagnostic activity, requires less collaboration and therefore less transactivity than secondary outcomes. The ability to justify and reach this

decision efficiently relies on evidence sharing and hypotheses sharing, which are more focused on transactivity within collaborative diagnostic reasoning (Weinberger & Fischer, 2006). Furthermore, in contrast to traditional diagnostic contexts (Boshuizen et al., 2020), simulation-based tasks reduce the relevance of content knowledge that can be obtained from various sources due to the iterative nature of collaborative diagnostic activities in simulations. The results provide further support for the differentiation of content and collaboration knowledge as outlined in the CDR-M (see 1.3.1; Radkowitsch et al., 2022). This underscores the importance of being aware of the knowledge distribution among collaboration partners and the relevance of a transactive memory (see 1.3.1; Wegner, 1987). While individual characteristics have been shown to influence the quality of collaborative diagnostic activities and the quality of collaborative diagnostic activities affect diagnostic outcomes, the effect is not mediated by collaborative diagnostic activities across studies. Therefore, in order to achieve effective collaborative problem-solving in knowledge-rich domains, such as collaborative diagnostic reasoning, it is not sufficient to possess sufficient content and collaboration knowledge; it is also necessary to enact collaborative diagnostic activities with high quality in order to achieve successful diagnostic outcomes. In summary, the second paper found that two factors, in addition to content knowledge, are crucial for successful collaborative problem solving in knowledge-rich domains: (1) knowledge about the domain of the collaboration partner and (2) collaborative diagnostic activities.

While the second paper focused at using process data to better understand collaborative diagnostic reasoning, the third paper focused on informing learning and instruction of collaborative diagnostic reasoning. The objective of the third paper and second empirical study (Brandl et al., 2021) was to investigate the extent to which process data of collaborative diagnostic reasoning from simulation-based learning can be utilized to predict collaborative problem-solving performance. If differences in behavior are related to success in simulated learning environments, they can be used to identify support needs at an early stage, thus enabling learners to progress in their learning more efficiently than with non-adaptive support (see 1.4.2; Plass & Pawar, 2020). The developed random forest classification model (Breiman, 2001) predicted a high percentage of accurate and inaccurate diagnoses with a classification accuracy of greater than 0.90, using high-level features constructed from log-file data. The nine high-level features were categorized as bigrams (n = 2; Damashek, 1995) of collaborative diagnostic activities. These represented either consistent behavior, with two instances of the same collaborative diagnostic activity, or transitions from one collaborative diagnostic activity to another, with two instances of different collaborative diagnostic activities. The results indicated that the consistent

features (e.g., time spent with evidence elicitation) and transitions from evidence elicitation to evidence sharing and from hypotheses sharing to evidence sharing are of less importance. In contrast, the transition from evidence sharing to hypotheses sharing is the most important feature. However, no notable distinction was observed in the feature with regard to accurate and inaccurate diagnoses. It can thus be assumed that a non-linear relation or complex interaction with one or more other features is involved. Furthermore, additional analyses employing partial dependence plots (Friedman, 2001) facilitate a more comprehensive understanding of the prediction model (Brandl et al., 2022, August/September). The results for the most important feature indicated that the absence of a transition from evidence sharing to hypotheses sharing was associated with an increased likelihood of an inaccurate diagnosis, while the presence of at least one transition was associated with a decreased probability (see Appendix 8.5). Therefore, the absence of this transition may serve as an indicator of the necessity for adaptive instructional support. In conclusion, the findings of Paper 3 contribute to the development of more adaptive instructional support within simulation-based learning by enabling the prediction of individual learning outcomes at an early stage of the process. Moreover, the results can serve as a foundation for more generalizable insights by employing theoretical-derived process indicators and high-level features. It is reasonable to hypothesize that the identified features are not exclusive to collaborative diagnostic reasoning in medical contexts. Instead, they are likely to be applicable to the development of collaborative problem-solving skills across various domains (Mislevy, 2019).

The three papers presented in this thesis aimed to improve the use of process data to assess and support collaborative problem-solving through the advancement of theoretical models and the provision of insights for learning and instruction. The two empirical contributions, which build on the views presented in the first paper, provide insights into how the full potential of process data analyses can be utilized not only to gain deeper insights into collaborative diagnostic reasoning, but also to predict performance of collaborative diagnostic reasoning to identify learners in need of additional instructional support.

5.2 Implications for Understanding Collaborative Diagnostic Reasoning

The first sub-goal of this thesis was to investigate how process data can facilitate theoretical advancements. More specifically, it examined how theoretical advancements in the context of collaborative diagnostic reasoning in agent-based simulations using process data can be facilitated.

A first conclusion concerns the distinction between content and collaboration knowledge proposed in the CDR-M (Radkowitsch et al., 2022). As outlined in Paper 2, the relations

between content and collaboration knowledge and the different collaborative diagnostic activities vary. This can be attributed to the varying degrees of transactivity associated with the respective collaborative diagnostic activities. Evidence elicitation represents the least transactive collaborative diagnostic activity within the context of the collaborative decision-making process. In the process of evidence elicitation, the collaboration partner is utilized as an external knowledge resource, requiring minimal collaborative effort (Weinberger & Fischer, 2006). Accordingly, different relations of content and collaboration knowledge to collaborative diagnostic activities in Paper 2 support the assumption from the CDR-M that content and collaboration knowledge are two distinct constructs. However, Paper 2 also raises questions about the role of content knowledge in simulation-based learning environments, where repeated attempts and revisions are possible. However, in a subsequent study (Vogel et al., 2023), we demonstrated that successful diagnosticians had more conceptual knowledge and spent less time with hypotheses sharing than those who were unsuccessful. Therefore, the creation of a shared problem representation during collaborative diagnostic reasoning necessitates both content knowledge and the externalization of information. Accordingly, an adequate initial problem representation, such as activated illness scripts, appears to be pivotal for the success of collaborative diagnostic reasoning (Charlin et al., 2007). Finally, the multi-study structural equation model in Paper 2 demonstrated that, alongside content knowledge—a traditional focus of expertise research—collaboration knowledge plays a crucial role in effective collaborative diagnostic reasoning. The importance of collaboration knowledge has also been demonstrated in the study conducted by Radkowitsch et al. (2021). The researchers demonstrated that providing learners with external collaboration scripts enhanced the performance of collaborative diagnostic activities. One potential explanation is that collaboration knowledge stored in internal collaboration scripts provides information about appropriate actions, which leads to enhanced information processing and reduced coordination effort (Kollar et al., 2006). Consequently, a reduction in collaboration load is achieved, which is defined as the working memory capacity required to engage in collaborative activities (F. Kirschner et al., 2009). These findings support the assumption of the CDR-M that content and collaboration knowledge are two distinct factors that are positively associated with collaborative diagnostic reasoning.

A further conclusion concerns the collaborative diagnostic activities proposed in the CDR-M. As stated in the CDR-M, collaborative diagnostic activities need to be enacted with high quality for a successful collaborative diagnostic reasoning, as indicated by their predictive accuracy to collaborative diagnostic reasoning performance in Paper 3. In addition, collaborative diagnostic activities account for additional variance beyond that explained by individual

characteristics as shown in Paper 2. This aligns with the findings of research on individual diagnostic reasoning (Fink et al., 2023), which investigated the extent to which individual diagnostic activities and content knowledge uniquely explain the variance in diagnostic success among 106 medical students. Moreover, Paper 3 illustrated that collaborative diagnostic activities derived from log-file data are effective in predicting collaborative diagnostic reasoning performance. These findings thus support the assumption of the CDR-M that collaborative diagnostic activities are of crucial importance for collaborative diagnostic reasoning.

Along this assumption, another conclusion concerns the differentiation of collaborative diagnostic activities. As proposed by the CDR-M, sharing and elicitation of evidence represent two distinct collaborative diagnostic activities, each requiring a different set of underlying collaboration skills (F. Fischer et al., 2002). This is indicated by the differing relations between individual characteristics and diagnostic outcomes as presented in Paper 2. In considering the differentiation between the evidence sharing and hypotheses sharing, the findings of Paper 2 indicated an absence of consistent support for a relation between any of the individual characteristics and the quality of hypotheses sharing. Leading to the conclusion that this might be either to the operationalization of quality in hypotheses sharing or that there is no direct relation between the individual characteristics and hypotheses sharing, as this relation is mediated by evidence sharing. Supporting the second possibility, Paper 3 revealed that the transition from evidence sharing to hypotheses sharing is crucial for reaching diagnostic accuracy. Specifically, when participants make this transition at least once, they significantly reduce the likelihood of arriving at incorrect diagnoses. Subsequent transitions from evidence sharing to hypotheses sharing have a less substantial impact than the initial one. This suggests that some form of data-driven reasoning, as opposed to hypotheses-driven reasoning, is a relevant factor in achieving an accurate diagnosis (Patel et al., 2005).

In order to reach a final conclusion regarding the role of process data in facilitating the understanding of collaborative diagnostic reasoning, it is necessary to investigate the complex, non-linear interactions between collaborative diagnostic activities. This approach is more appropriate than analyzing these interactions in an isolated or linear manner. The non-mediating relation of collaborative diagnostic activities between individual characteristics and diagnostic outcomes is in contradiction with the assumptions of the CDR-M. Along the effects of the collaborative diagnostic activities in Paper 2, an isolated analysis of these activities does not fully represent the complex interactions and relations among activities, individual characteristics, and diagnostic outcomes. In Paper 3, we used bigrams of the collaborative diagnostic activities, representing short sequences of either spending time with or transitioning between collaborative

diagnostic activities, as features in a random forest prediction model. This model allows for complex-nonlinear relations among the features. The results indicate that consistent features (e.g., two instances of evidence elicitation indicating time spent with evidence elicitation) are found to be less important. However, while the results indicated that the transition from evidence sharing to hypotheses sharing is the most important feature for predicting diagnostic accuracy, they are unable to produce actionable information such as the direction of the relation, beyond that this feature is important (see 1.4.3; Yarkoni & Westfall, 2017). Subsequent analyses have revealed that when this transition is performed at least once, the likelihood of an inaccurate diagnosis is decreased. Therefore, the CDR-M will need to consider not only isolated collaborative diagnostic activities, but also the complex interactions between them.

In summary, the results of the process data analyses presented in this thesis provide support for the assumptions proposed in the CDR-M (see Table 2).

Table 2

Assumptions of the CDR-M Based on Presented Empirical Evidence

Assumptions …	**Empirical Evidence**
… from the CDR-M that are supported	
distinction between content and collaboration knowledge	• different relations of content and collaboration knowledge to collaborative diagnostic activities in Paper 2
collaborative diagnostic activities are relevant for successful collaborative diagnostic reasoning	• collaborative diagnostic activities explain variance beyond individual characteristics in Paper 2 • collaborative diagnostic activities constructed from log-file data are suitable for predicting collaborative diagnostic reasoning performance in Paper 3
sharing and elicitation of evidence are two distinct collaborative diagnostic activities	• differential relations of the collaborative diagnostic activities from the individual characteristics and to the diagnostic outcomes in Paper 2
… added to the CDR-M	
isolated analysis of collaborative diagnostic activities does not fully represent the complex interactions and relations among activities, individual characteristics, and diagnostic outcomes	• non-mediating effect of collaborative diagnostic activities in Paper 2 • different effects of the collaborative diagnostic activities in Paper 2 • importance of features depicting transitions between collaborative diagnostic activities in Paper 3
complex non-linear interactions between collaborative diagnostic activities	• predictive accuracy in Paper 3 • partial dependence plot for transition from evidence sharing to hypotheses sharing in additional analyses of Paper 3

First, the results provide evidence in support of the distinction between content and collaboration knowledge. Furthermore, the results provide support for the concept of illness scripts and internal collaboration scripts, which are used to store content and collaboration knowledge. The findings also underscore the importance of collaboration knowledge, in addition to content knowledge, potentially through reduced collaboration load. Secondly, the results support the assumption that collaborative diagnostic activities are a key factor in the success of collaborative diagnostic reasoning. Thirdly, the results support the differentiation between the processes of sharing and elicitation as two distinct collaborative diagnostic activities. However, the results also challenge assumptions made in the CDR-M, thereby introducing two new assumptions to the CDR-M. First, the results indicated that an isolated analysis of collaborative diagnostic activities does not fully represent the complex interactions and relations among activities, individual characteristics, and diagnostic outcomes. Second, the interpretation of the results highlights the need to investigate complex non-linear interactions between collaborative diagnostic activities.

These assumptions were tested in the context of collaborative diagnostic reasoning, specifically in the process of joint evidence generation between internists and radiologists in medicine. It seems reasonable to assume that the aforementioned relations can be generalized to other contexts in which two or more diagnosticians collaborate to solve diagnostic problems. Nevertheless, a systematic investigation of the generalizability of these findings to other contexts has yet to be conducted.

5.3 Implications for Supporting Collaborative Diagnostic Reasoning

The second sub-goal of this thesis was to inform learning and instruction. Through the analyses in papers 2 and 3, conclusions can be drawn about facilitating the development of collaborative diagnostic reasoning that can align assessment design with learning design and pave the way for adaptive instructional support (see 1.4.2).

Based on the implications of the previous chapter, it seems important to support the development of collaborative diagnostic reasoning skills in medical education by (1) facilitating the acquisition of collaboration knowledge and (2) facilitating the performance of collaborative diagnostic activities with high quality. This is indicated by the relation of collaboration knowledge to the quality of evidence sharing and the various, but not mediating, relations between the quality of collaborative diagnostic activities and diagnostic outcomes in Paper 2. Based on these findings and the study by Radkowitsch et al. (2021), which indicated that providing external collaboration scripts during simulation-based learning improves the quality of collaborative diagnostic activities, it seems beneficial to provide external collaboration

scripts. Incorporating collaboration scripts into simulation-based learning in medical education can enhance learners' experiences with authentic patient cases and collaborative settings, while providing additional instructional support to develop collaborative diagnostic reasoning skills.

Such additional instructional support is ideally adapted to the learner's needs to ensure that a task is in the learner's zone of proximal development (Plass & Pawar, 2020; Vygotsky, 1978). One way to implement this in medical education could be, following the results of Paper 3, to use learners' process data while working on the CoSiMed simulation (see 1.3.3) to predict their performance while they are still working on the task (Richters, Stadler, Radkowitsch et al., 2023). If the learner model predicts an inaccurate diagnosis and thus a failure to complete the task, the model could identify those learners who do not transition from evidence sharing to hypotheses sharing while filling out the radiological request form and provide them with additional instructional support (Basu et al., 2017). In particular, as a strong relation between collaboration knowledge and evidence sharing was found in Paper 2, the provision of external collaboration scripts could be an appropriate form of additional instructional support in this situation. This adaptive provision of instructional support could avoid an expertise-reversal effect, where scaffolds that are initially effective may hinder learning as the learner's expertise increases (Kalyuga, 2007).

In line with these considerations for a learner model that provides adaptive instructional support, it is important to initially design learning environments with respect to collaborative diagnostic activities so that the measurement of these processes is a design factor, rather than using log-file data as a by-product of the product data collection process (Goldhammer et al., 2021). It should be decided a priori how specific theory-based and thus high-level features will be constructed from log-file data. These considerations must then guide the design of the task. The use of high-level features not only allows generalization of findings to different tasks, but also provides actionable results. While the result *"it is important to click first in section A and then section B*" entails no relevant information for educational decisions, the result *"transitioning from evidence sharing to hypotheses sharing should be performed at least once"* is actionable allowing to design instructional support fostering this behavior.

Overall, the process data analyses presented in this thesis allow for three implications for supporting collaborative diagnostic reasoning. Firstly, the presented papers suggest that focusing on the acquisition of collaboration knowledge and the performance of high-quality collaborative diagnostic activities facilitates the development of collaborative diagnostic reasoning. Secondly, providing collaboration scripts in the absence of a transition from evidence sharing to hypotheses sharing during collaborative diagnostic reasoning could be identified as a strategy

for adaptive instructional support. Finally, a case was made for making the measurement of processes such as collaborative diagnostic activities a design factor of learning environments to support the development of collaborative diagnostic reasoning and to inform learning and instruction using process data.

5.4 Implications for Leveraging Process Data of Collaborative Problem-Solving

The overall goal of this thesis was to improve the use of process data to assess and support collaborative problem-solving. Therefore, two sub-goals have been identified and already discussed in light of papers 2 and 3, namely theoretical advancements and informing learning and instruction. What has not yet been addressed in this discussion is what implications can be drawn for the leverage of process data for collaborative problem-solving based on the three papers presented, with a focus on the three challenges identified (see 1.4.3): ethical considerations, dealing with complexity, and lack of theory.

Some of the ethical challenges during data collection can be effectively addressed through the use of high-level features, as demonstrated in papers 2 and 3. By focusing on collecting only relevant data, incorporated by the design of the task (Goldhammer et al., 2021), this approach minimizes unnecessary information collection. It also helps learners understand why certain data are needed and the implications of omitting them. For example, excluding such data could hinder the system's ability to provide adaptive instructional support, potentially reducing the effectiveness of the learning experience. This gives learners, at least in a higher education context, a choice about what they want to allow the researcher/educator to collect, and allows for informed consent based on why that data should be collected and what the consequences of not collecting the data might be. In addition, Paper 3 identifies learners in need of adaptive instructional support, which allows for enabling interventions, an ethical concern that is not currently systematically addressed in research (Cerratto Pargman & McGrath, 2021). By using partial dependence plots, the transparency issue could also be addressed to some extent, and additional insights into how the prediction is achieved could be gained (Yan et al., 2023).

Turning to the challenge of complexity in process data analyses. Research has long been interested in the cognitive (and collaborative) activities involved in (collaborative) problem-solving, but initially it was only possible to infer the outcome, for example through think aloud protocols (Ericsson & Simon, 1980). By using interactive tasks, process data can be collected unobtrusively without the need for additional measurements that could increase cognitive load, and thus the cognitive processes involved can be studied more easily (Matcha et al., 2019). Paper 3, meanwhile, showed that theory-based process indicators constructed from sequences of automatically coded log-file data corresponding to collaborative diagnostic activities

predicted task performance. Thus, it is possible to (partially) automate analyses using log-file data if learning environments are intentionally designed with theory-based indicators such as collaborative diagnostic activities in mind, making the measurement of these processes a design factor rather than using log-file data as a byproduct of the product data management process (Goldhammer et al., 2021).

Regarding the interpretation of machine learning algorithms (as used in paper 3), they come with the complexity of not explaining how the predictions were achieved in order to allow for complex nonlinear interactions, leading to less transparent models, also known as black boxes (Molnar et al., 2018; Yarkoni & Westfall, 2017). While such a black box model was used in Paper 3, additional analyses using partial dependence plots provided some additional insights, such as that the most important feature for prediction needs to be present at least once to increase the likelihood of successful performance. However, the use of machine learning algorithms in education is still a young and evolving field of research, with new models and approaches being developed all the time (Hilbert et al., 2021; Rane et al., 2024). For example, while knowledge tracing, a method for monitoring learners' skill mastery and predicting their performance, was developed thirty years ago (Corbett & Anderson, 1995), the recently developed interpretable knowledge tracing model outperforms known knowledge tracing models while allowing for more causal interpretations of the prediction (Minn et al., 2022). Such models could be used, for example, to predict collaborative diagnostic reasoning performance based on collaborative diagnostic activities constructed from log-file data, with a particular focus on the transition between collaborative diagnostic activities following the results of Paper 3.

Finally, the challenges of interpreting the results of machine learning on process data are addressed. While paper 1 argued for the need to link log-file data to theory-based constructs in process data analyses, papers 2 and 3 present two concrete examples of how this can facilitate theoretical advancements and derive actionable results from process data analyses to support collaborative diagnostic reasoning skills. Furthermore, theory-based constructs can go beyond idiosyncratic findings for one task and facilitate much-needed replication studies (Andres et al., 2017; Renkewitz & Heene, 2019; Zwaan et al., 2017). In particular, Paper 2 demonstrates the importance of empirically testing theoretical models in more than one dataset, as some of the hypotheses presented were present in one dataset but were not stable across datasets. Only by analyzing three data sets, the relevance of collaboration knowledge for collaborative diagnostic reasoning and the contribution of collaborative diagnostic activities alone do not mediate the effect of individual characteristics on diagnostic outcomes. These findings demonstrate the need to improve medical students' collaboration knowledge and collaborative diagnostic

activities beyond content knowledge, and serve as a resource for policy stakeholders in shaping curriculum decisions. In turn, through theoretical advancements, as detailed in 5.2, it is possible to align assessment design with instructional design to inform learning and instruction, as done in Paper 3 and subsequent analyses that identified the transition of the most important feature in prediction as a starting point for adaptive instructional support (see 5.3).

Collaborative diagnostic activities provide a promising foundation for the development of theory-based process indicators that enhance the generalizability and transferability of research findings across tasks and domains. A key advantage of collaborative diagnostic activities as theory-based process indicators is that they can be utilized as a common language in research, which is a prerequisite for research on generalizability and transfer of findings to other domains. In order to address the issue of the lack of generalizability of findings, given the strong task-specificity of findings, we suggested in Paper 1 the use of high-level features (Mislevy, 2019) representing meaningful process indicators of psychological constructs based on theory. Two types of high-level features were utilized as indicators of collaborative diagnostic activities. In Paper 2, the quality of collaborative diagnostic activities was evaluated through the calculation of metrics, including the precision of evidence elicitation from log-file data, with expert solutions serving as a reference point. In Paper 3, bi-grams of collaborative diagnostic activities per second were constructed from log-file data, incorporating the time stamps of clicks. In particular, Paper 3 illustrated that high-level features derived from log-file data based on theory are an effective means of predicting performance. In addition, Paper 2 revealed that the quality collaborative diagnostic activities offer a unique contribution to problem-solving performance, as indicated by the non-mediating effect of these activities. Therefore, collaborative diagnostic activities account for additional variance beyond that explained by individual characteristics, such as knowledge. One first analysis on generalizability and transfer of findings to other domains in the context of collaborative diagnostic reasoning is done by Oezsoy et al. (2024, August) where we applied parts of the CDR-M, specifically entailing evidence elicitation and sharing as collaborative diagnostic activities, to teacher education and compared it to data from medical education. In light of these findings, it seems reasonable to argue that the collaborative diagnostic activities entailed in the CDR-M represent suitable theory-based process indicators, facilitating the generalizability of findings across tasks and potentially even across domains. This is due to their potential to serve as a suitable starting point for a shared language among researchers.

In sum, the process data analyses presented in this thesis have six implications for the utilization of process data in the assessment and support of collaborative problem-solving. First of

all, the theory-based approaches presented in this thesis illustrate how challenges with respect to ethical considerations, e.g., lack of informed consent or a lack of transparency, can be minimized during the process of data collection. Furthermore, with regard to data analyses, the present findings enable the identification of adaptive instructional support strategies based on high-level features. Thirdly, the analyses of log-file data can be partially automated. Fourthly, with regard to the interpretation of process data analyses, this thesis demonstrates that learner models, such as knowledge tracing, are enhanced by a focus on the transitions between collaborative diagnostic activities. In addition, the findings demonstrate how high-level features facilitate theoretical advancements and enable the achievement of actionable results. Ultimately, the utilization of high-level features, such as collaborative diagnostic activities, to establish a common language in research enables the generalizability of findings across tasks and, potentially, even across domains.

5.5 Transferability: Domain Specificity of Collaborative Problem-Solving

Collaborative problem-solving skills and collaborative diagnostic reasoning skills are frequently regarded as domain-specific skills, as they rely on domain-specific schemata and scripts to be performed with expertise (Sweller, 1988; van Lehn, 1989). For example, an internist's capacity to diagnose a medical condition is grounded in the structured knowledge inherent to illness scripts (Charlin et al., 2007). Similarly, a teacher's capacity to diagnose learning difficulties or identify potential misconceptions in students is rooted in their understanding of pedagogical frameworks (Heitzmann et al., 2019). However, the distinction between domain-specific and domain-general problem-solving is not a dichotomy; rather, it represents a continuum (Perkins & Salomon, 1989).

At one end of the continuum, domain-general strategies such as vary-one-thing-at-a-time (VOTAT) are essential in knowledge-lean tasks or when specific domain knowledge is absent (Greiff et al., 2014). These strategies are universally applicable across various domains because they rely on fundamental problem-solving principles rather than domain-specific content. Conversely, knowledge-rich tasks that necessitate the utilization of well-structured, domain-specific scripts, such as the aforementioned medical or educational diagnoses, demand more specialized approaches that are substantially influenced by the expert's experience and domain knowledge. Despite these differences, collaborative diagnostic activities, such as those involved in collaborative diagnostic reasoning, indicate a form of cross-domain applicability. For example, an internist who is engaged in a collaboration with a radiologist with the objective of sharing information for the purpose of further evidence generation, or two teachers who are exchanging observations about a student, both engage in similar collaborative diagnostic activities from a

conceptual perspective. These scenarios require the elicitation and sharing of evidence, processes that are applicable across domains. However, the domain-specific expertise determines which information is considered relevant and irrelevant, respectively. An internist would lack the pedagogical knowledge necessary to diagnose educational issues, and vice versa.

To effectively support research on cross-domain transferability of collaborative problem-solving processes, it is important to not only conceptualize them as done in several theoretical frameworks (e.g. Hesse et al., 2015; Radkowitsch et al., 2022; Sun et al., 2020), but also to measure these processes using high-level features. These frameworks offer a common language and structure for the description of collaborative activities, such as evidence sharing and planning, which can be applied across different domains.

In conclusion, while collaborative problem-solving and collaborative diagnostic reasoning are dependent on domain-specific expertise, the conceptualization of collaborative problem-solving activities demonstrates cross-domain applicability. By employing theory-based frameworks to describe these activities, we can establish a common foundation across different domains, thereby enhancing the utilization of process data for the assessment and support of collaborative problem-solving skills.

5.6 Limitations

This thesis is not without limitation that should be kept in mind and lower in some aspects the potential to generalize the findings.

A limitation with respect to the collaborative diagnostic activities is the operationalization of quality in Paper 2. The quality indicators could always only shed light on one perspective of each activity, while possibly obscuring others. For instance, it is possible that content knowledge is not related to the precision of hypotheses sharing. However, this may be different when examining other quality indicators, such as sensitivity or specificity. Accordingly, a more detailed examination of the measurement of expert performance in the context of collaborative diagnostic activities seems to be warranted. Moreover, while collaborative problem-solving and collaborative diagnostic reasoning encompass a range of essential collaborative activities, the CoSiMed simulation concentrated on sharing-related skills and did not examine activities such as negotiation. This decision was made based on findings from prior research (e.g., Tschan et al., 2009), which indicated that diagnosticians face challenges in sharing information. Additionally, interviews with practitioners revealed that the primary obstacle in the collaboration between internists and radiologists (a common and complex situation requiring collaborative diagnostic reasoning) is the lack of precise justification for the test (e.g., the absence of relevant information) and the lack of patient information clustering (Radkowitsch et al., 2020).

Nevertheless, the findings and implications of Papers 2 and 3 are limited to sharing-related skills of collaborative diagnostic reasoning. Consequently, further research on additional aspects, such as negotiation skills, is required.

Another limitation is the use of an agent-based simulation of collaborative diagnostic reasoning, which allows for a standardized and controlled setting that is difficult to establish in human collaborations while ensuring all necessary activities are performed (Rosen, 2015). Nevertheless, the restricted nature of conversational interactions permitted by agent-based collaboration has been identified as a potential limitation, with the extent to which natural collaboration can unfold being constrained as a result (Graesser et al., 2017). This may account for the limited impact of social skills in Paper 2, as the context may not necessitate the application of a broad range of social skills (Hesse et al., 2015; Radkowitsch et al., 2020). In a real-life collaboration, the effects of social skills might be more pronounced. Nevertheless, research demonstrated that the human-to-agent approach yielded comparable outcomes in collaborative problem-solving to the human-to-human approach in the PISA 2015 study, and correlations with other measures of collaborative skills have been identified (Herborn et al., 2020; Stadler et al., 2020). Moreover, the CoSiMed simulation has been thoroughly validated and is perceived as authentic, given that this specific collaborative situation also occurs in practice via distance communication (Radkowitsch et al., 2020). However, the advent of generative artificial intelligence and the utilization of large language models in educational settings has opened the possibility of developing a more flexible and authentic generative agent for collaboration. This agent could adapt its behavior in response to the learner's process and queries, facilitating a more natural and authentic interaction (Kasneci et al., 2023; Yan et al., 2024).

A further potential limitation is the appropriateness of using diagnostic accuracy as a reliable and valid measure of (collaborative) diagnostic reasoning skills. In addition to diagnostic accuracy, which serves as an indicator of diagnostic quality, diagnostic reasoning skills also encompass professional knowledge and diagnostic activities (Heitzmann et al., 2019). Nevertheless, achieving diagnostic accuracy represents the primary objective of diagnostic reasoning, not only within the medical domain (Pickal et al., 2023). The importance of accurate diagnoses cannot be overstated, given the potentially grave consequences for patients when a diagnosis is inaccurate (Balogh et al., 2015). Furthermore, a patient's diagnosis has a substantial impact on subsequent steps, including the formulation of treatment plans (Cook et al., 2019).

A final limitation concerns the sample analyzed in Papers 2 and 3. It consisted exclusively of medical students with an intermediate level of expertise, which may have prevented the observation of behaviors characteristic of both experts and novices. In addition, results in Paper 2

showed that participants overall had on average a high level of collaboration knowledge. This could be an indicator that the collaboration knowledge tests might be too easy or to narrowly focusing on evidence sharing. Thus, in future studies, revising the measurement of collaboration knowledge is important. While the current assessment focuses on meta-knowledge about the collaboration partner, future operationalizations should aim to measure internal collaboration scripts, given the relevance of this construct found in this thesis.

5.7 Directions for Future Research

The overarching goal of this thesis was to enhance the use of process data for assessing and supporting collaborative problem-solving skills. This thesis examined the potential of process data analyses to (1) facilitate theoretical advancements and (2) inform learning and instruction of collaborative problem-solving in the context of collaborative diagnostic reasoning. In addition, the results also highlight several promising avenues for future research.

A first direction for future research could be to investigate non-linear relations between the collaborative diagnostic activities. In light of the findings presented in Papers 2 and 3, it becomes evident that an isolated examination of collaborative diagnostic activities is insufficient to fully capture the complex interactions among these activities. The use of bigrams in Paper 3 represented a preliminary investigation into this direction; however, further research is necessary to examine longer sequences, such as sequence clustering (Piccarreta, 2017). As the exploratory use of longer sequences can rapidly increase the required sample size, it is recommended that potentially relevant sequences be defined a priori based on theory or previous studies. Hypotheses about such strategic behavior in collaborative problem-solving tasks should then be explicitly tested. This would automatically account for the relevance of theory outlined before. For instance, in light of the findings presented in Paper 3, it may be reasonable to exclude the analysis of time spent on collaborative diagnostic activities. Instead, it would be more fruitful to examine the actions undertaken directly before and after the first transition from evidence sharing to hypotheses sharing.

A second direction of future research is to investigate the extent to which the proposed adaptive instructional support is suitable for facilitating the learning of collaborative diagnostic reasoning skills. In light of the found relation between collaboration knowledge and evidence sharing in Paper 2, as well as the finding from additional analyses in Paper 3 that transitioning from evidence sharing to hypotheses sharing at least once decreases the likelihood of concluding an inaccurate diagnosis, it is proposed that an adaptive instructional support strategy be implemented, whereby learners who are not transitioning from evidence sharing to hypotheses sharing are provided with a collaboration script. Prior research has already indicated that the

performance of evidence sharing is facilitated with an adaptive collaboration script (Radkowitsch et al., 2021). However, in this study, adaptivity was realized according to the quality of the performed activities. It remains unclear whether adapting to the sequence of activities is more beneficial.

A third direction of future research is the utilization of a generative AI-enhanced agent. Although agent-based collaboration offers significant advantages (see 1.2.3), it also entails certain limitations (see 5.4). Consequently, it may be beneficial to investigate whether the findings of Papers 2 and 3 can be replicated in a human-to-human setting. However, as it would be more challenging to obtain comprehensive process data on the required activities in such a setting, a preliminary approach could be to utilize a generative AI-enhanced agent to bridge the gap between human-to-agent and human-to-human approaches while maintaining a controlled environment, individual measurement, and the necessity of all required processes.

A fourth direction for future research could be to examine the transferability of the CDR-M and the findings of the presented Papers 2 and 3 to other diagnostic situations and to other domains. While Papers 2 and 3 concentrated on information-sharing skills during collaborative problem-solving using the CoSiMed Simulation, the CDR-M also entailed other collaborative activities. This decision was made during the design of the simulation, as prior research has indicated difficulties in sharing and eliciting information with and from collaboration partners (Tschan et al., 2009) and interviews with medical doctors identified the task of requesting a radiologic examination as of crucial importance (Radkowitsch et al., 2020). Therefore, future research may examine the role of, for instance, negotiation skills in collaborative diagnostic reasoning. To enhance the generalizability of the current findings, it would be further interesting to investigate other tasks within the medical domain, as well as in other domains requiring collaborative diagnostic reasoning skills. These could include diagnosing learning difficulties or students' possible misconceptions in teacher education (Heitzmann et al., 2019). In comparison to other domains, medicine is characterized by highly standardized procedures, which could have an impact on collaborative diagnostic activities and, as a result, limit the extent to which the findings can be transferred to other domains. Therefore, further research is required to investigate the transferability of the conceptualization of collaborative problem-solving processes.

On a more general note, future research is encouraged to challenge and advance existing theories with respect to the details they imply for relevant activities. While theories of collaborative problem-solving describe an idealized representation of the problem-solving process through the coordination of externalized individual activities into a coherent sequence of events

(Hesse et al., 2015), research on this sequence of events during collaborative problem-solving using process data is still rare. Therefore, a promising direction for future research is the identification of beneficial sequences of processes and the examination of how these sequences are influenced by individual and task characteristics. For example, Paper 3 could identify the transition from evidence sharing to hypotheses sharing as a relevant process to be performed at least once, or the finding that transitions between collaborative diagnostic activities seem to be more important than the time spent on these activities. These findings can then be used to advance existing theories on collaborative problem-solving (e.g., Hesse et al., 2015; OECD, 2017; Radkowitsch et al., 2022; Sun et al., 2020) in order to facilitate the sustainable use of process data, thereby enhancing the theory and practice of collaborative problem-solving. The use of a joint language in this line of research would also have the potential to facilitate the generalization of findings across domains.

CONCLUSION

6

Laura Brandl

The ability to engage in collaborative diagnostic reasoning is a fundamental skill for numerous professionals on a daily basis (Graesser et al., 2018; Radkowitsch et al., 2022). A lack of skills and errors can have severe negative consequences, including serious adverse events and suboptimal patient care, particularly in the context of medicine (Hooftman et al., 2024). To gain a deeper understanding of collaborative diagnostic reasoning from a theoretical perspective and thereby inform learning and instruction, the present thesis focused on how process data derived from interactive collaborative problem-solving tasks, particularly within the context of collaborative diagnostic reasoning in agent-based simulations, can be utilized to enhance theoretical models and learning and instruction of collaborative problem-solving skills. The overarching goal of this thesis was to demonstrate how process data can be employed in a sustainable and meaningful manner to enhance theoretical models and instructional support for collaborative problem-solving skills. Recent advancements in the context of process data analyses permit researchers to collect process data in an unobtrusive manner, thereby facilitating the investigation of the cognitive processes involved in collaborative problem-solving and collaborative diagnostic reasoning (Matcha et al., 2019). However, the use of process data presents both opportunities and challenges. Process data analyses allow for more nuanced assessments of learners' problem-solving skills, thereby providing a basis for targeted interventions and a deeper theoretical understanding. To leverage the full potential of process data analyses of collaborative problem-solving it is crucial to use standardized data collection, complex analyses methods, and robust theoretical frameworks.

The findings of the presented papers demonstrated how process data can be utilized to advance theoretical models, as illustrated by the CDR-M, to facilitate the learning of collaborative diagnostic reasoning skills and, consequently, enhance the utilization of process data in collaborative problem-solving scenarios. This is exemplified by the application of collaborative diagnostic reasoning in agent-based simulations. With respect to theoretical advancements, the thesis provided support for four assumptions proposed in the CDR-M. In addition, two new assumptions were added to the model: Firstly, the unique contribution of collaborative diagnostic activities to collaborative diagnostic reasoning and secondly, the need to investigate complex non-linear interactions between collaborative diagnostic activities. In terms of supporting the development of collaborative diagnostic reasoning skills, there are several practical implications: Firstly, it is important to focus on collaboration knowledge and collaborative diagnostic activities. Secondly, it is necessary to consider how to measure processes such as collaborative diagnostic activities and make this a key design factor. Moreover, a strategy for providing adaptive instructional support is presented. Lastly, the findings of this thesis also provide insights

into the potentials of enhancing the usage of process data analyses in the assessment and support of collaborative problem-solving. It is of crucial importance to employ theory-based frameworks to describe collaborative problem-solving processes, such as collaborative diagnostic activities of collaborative diagnostic reasoning, in order to establish a common ground for the assessment and support of collaborative problem-solving skills across different domains. This will, in turn, facilitate further improvements in the use of process data analyses. Which will lead to more proficient collaborators in the future, not only in the medical domain.

REFERENCES

7

Laura Brandl

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

Anderson, J. R. (1983). *The architecture of cognition*. Harvard University Press.

Anderson, J. R. (1987). Skill Acquisition: Compilation of Weak-Method Problem Solutions. *Psychological Review*, 94(2), 192-210.

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist*, *48*(1), 35–44. https://doi.org/10.1037//0003-066x.48.1.35

Andres, J. M. L., Baker, R. S., Siemens, G., Gašević, D., & Spann, C. A. (2017). Replicating 21 Findings on Student Success in Online Learning. *Technology, Instruction, Cognition and Learning*, *10*(4), 313–333.

Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, *104*, 105759. https://doi.org/10.1016/j.chb.2018.10.025

Andrews-Todd, J., Jiang, Y., Steinberg, J., Pugh, S. L., & D'Mello, S. K. (2023). Investigating collaborative problem solving skills and outcomes across computer-based tasks. *Computers & Education*, *207*, 104928. https://doi.org/10.1016/j.compedu.2023.104928

Anghel, E., Khorramdel, L., & Davier, M. von (2024). The use of process data in large-scale assessments: a literature review. *Large-Scale Assessments in Education*, *12*(1). https://doi.org/10.1186/s40536-024-00202-1

Atkinson, R. K., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. In K. Spence & J. Spence (Eds.), *The psychology of learning and motivation (Vol. 2, pp).* (2nd ed., pp. 89–195). Academic Press.

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology*, *29*(3), 344–370. https://doi.org/10.1016/j.cedpsych.2003.09.002

Azevedo, R., Cromley, J. G., Winters, F. I., Moos, D. C., & Greene, J. A. (2005). Adaptive Human Scaffolding Facilitates Adolescents' Self-regulated Learning with Hypermedia. *Instructional Science*, *33*(5-6), 381–412. https://doi.org/10.1007/s11251-005-1273-8

Azevedo, R., & Gašević, D. (2019). Analyzing Multimodal Multichannel Data about Self-Regulated Learning with Advanced Learning Technologies: Issues and Challenges. *Computers in Human Behavior*, *96*, 207–210. https://doi.org/10.1016/j.chb.2019.03.025

Baker, E. L., Chung, G. K. W. K., & Cai, L. (2016). Assessment Gaze, Refraction, and Blur. *Review of Research in Education*, *40*(1), 94–142. https://doi.org/10.3102/0091732X16679806

Baker, R., Di Xu, Park, J., Yu, R., Li, Q., Cung, B., Fischer, C., Rodriguez, F., Warschauer, M., & Smyth, P. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: opening the black box of learning processes. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-00187-1

Balogh, E. P., Miller, B. T., & Ball, J. R. (Eds.). (2015). *Improving Diagnosis in Health Care*. https://doi.org/10.17226/21794

Bandura, A. (1977). *Social learning theory*. Prentice Hall.

Bandura, A. (2008). Observational Learning. In W. Donsbach (Ed.), *The International Encyclopedia of Communication*. Wiley. https://doi.org/10.1002/9781405186407.wbieco004

Basu, S., Biswas, G., & Kinnebrew, J. S. (2017). Learner modeling for adaptive scaffolding in a Computational Thinking-based science learning environment. *User Modeling and User-Adapted Interaction*, *27*(1), 5–53. https://doi.org/10.1007/s11257-017-9187-0

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*, Article 977631. https://doi.org/10.3389/feduc.2022.977631

Bellezza, F. S., & Bower, G. H. (1981). The representational and processing characteristics of scripts. *Bulletin of the Psychonomic Society*, *18*(1), 1–4.

Bond, M., Khosravi, H., Laat, M. de, Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21(1). https://doi.org/10.1186/s41239-023-00436-z

Boshuizen, H. P., 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., & Schmidt, H. G (1992). On the Role of Biomedical Knowledge in Clinical Reasoning by Experts, Intermediates and Novices. *Cognitive Science*, *16*, 153–184.

Boshuizen, H. P., Schmidt, H. G, Custers, E. J. F. M., & van de Wiel, M. W. (1995). Knowledge Development And Restructuring in The Domain of Medicine: The Role of Theory and Practice. *Learning and Instruction*, *5*, 269–289. https://doi.org/10.1016/0959-4752(95)00019-4

Bowen, J. L. (2006). Educational Strategies to Promote Clinical Diagnostic Reasoning. *The New England Journal of Medicine*, *355*(21), 2217–2225.

Brady, A. P. (2017). Error and discrepancy in radiology: Inevitable or avoidable? *Insights into Imaging*, *8*(1), 171–182. https://doi.org/10.1007/s13244-016-0534-1

Brandl, L., Richters, C., Radkowitsch, A., Obersteiner, A., Fischer, M. R., Schmidmaier, R., Fischer, F., & Stadler, M. (2021). Simulation-Based Learning of Complex Skills: Predicting Performance With Theoretically Derived Process Features. *Psychological Test and Assessment Modeling*, *63*(4), 542–560. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2021-4/PTAM__4-2021_6_kor.pdf

Brandl, L., Richters, C., Radkowitsch, A., Obersteiner, A., Fischer, M. R., Schmidmaier, R., Fischer, F., & Stadler, M. (2022, August/September). *Complex Skills in Simulations: Predicting Performance with Theoretically Derived Process Features [Paper Presentation]*. EARLI SIG 27 Conference, Southampton, England.

Brandl, L., Stadler, M., Richters, C., Radkowitsch, A., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2024). Collaborative Problem-Solving in Knowledge-Rich Domains: A Multi-Study Structural Equation Model. *International Journal of Computer-Supported Collaborative Learning.* Advance online publication. https://doi.org/10.1007/s11412-024-09425-4

Braun, L. T., Zottmann, J. M., Adolf, C., Lottspeich, C., Then, C., Wirth, S., Fischer, M. R., & Schmidmaier, R. (2017). Representation scaffolds improve diagnostic efficiency in medical students. *Medical Education*, 51(11), 1118–1126. https://doi. org/10. 1111/medu. 13355

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The Four Generations of Computerized Educational Measurement. *ETS Research Report Series*, *1988*(1). https://doi.org/10.1002/j.2330-8516.1988.tb00291.x

Cannon-Bowers, J. A., & Bowers, C. (2010). Synthetic learning environments: On developing a science of simulation, games, and virtual worlds for training. In S. W. J. Kozlowski & E. Salas (Eds.), *Learning, training, and development in organizations* (pp. 229–261). Routledge/Taylor & Francis Group.

Care, E., Griffin, P., & McGaw B. (2012). *Assessment and teaching of 21st century skills.* Springer.

Cerratto Pargman, T., & McGrath, C. (2021). Mapping the Ethics of Learning Analytics in Higher Education: A Systematic Literature Review of Empirical Research. *Journal of Learning Analytics*, *8*(2), 123–139.

Chai, H., Hu, T., & Wu, L. (2024). Computer-based assessment of collaborative problem solving skills: A systematic review of empirical research. *Educational Research Review*, *43*, 100591. https://doi.org/10.1016/j.edurev.2023.100591

Charlin, B., Boshuizen, H. P., Custers, E. J. F. M., & 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. *Medical Education*, *46*(5), 454–463. https://doi.org/10.1111/j.1365-2923.2012.04242.x

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Chatti, M. A., Dyckhoff, A. L., Schroder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5/6), 318–331.

Chaudhry, M. A., Cukurova, M., & Luckin, R. (2022). A Transparency Index Framework for AI in Education. *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium*, 195–198. https://doi.org/10.1007/978-3-031-11647-6_33

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical Analysis of Complex Problem-Solving Process Data: An Event History Analysis Approach. *Frontiers in Psychology*, *10*, Article 486. https://doi.org/10.3389/fpsyg.2019.00486

Chernikova, O., Heitzmann, N., Opitz, A., Seidel, T., & Fischer, F. (2022). A theoretical framework for fostering diagnostic competences with simulations in higher education. In F. Fischer & A. Opitz (Eds.), *Springer Briefs in Education. Learning to diagnose with simulations - Examples from teacher education and medical education.* (pp. 3–14). Springer.

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

Chetverikov, A., & Upravitelev, P. (2016). Online versus offline: The Web as a medium for response time data collection. *Behavior Research Methods*, *48*(3), 1086–1099. https://doi.org/10.3758/s13428-015-0632-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

Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator*, *15*(3), 6–11. http://www.psy.lmu.de/isls-naples/intro/all-webinars/collins/cognitive-apprenticeship.pdf

Cook, D. A. (2014). How much evidence does it take? A cumulative meta-analysis of outcomes of simulation-based education. *Medical Education*, *48*(8), 750–760. https://doi.org/10.1111/medu.12473

Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. M. (2013). Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. *Academic Medicine: Journal of the Association of American Medical Colleges*, *88*(6), 872–883. https://doi.org/10.1097/ACM.0b013e31828ffdcf

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 : Journal of the Association of American Medical Colleges*, *94*(9), 1310–1316. https://doi.org/10.1097/ACM.0000000000002768

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

Corbalan, G., Kester, L., & van Merriënboer, J. J. G. (2006). Towards a personalized task selection model with shared instructional control. *Instructional Science*, *34*(5), 399–422. https://doi.org/10.1007/s11251-005-5774-2

Corbett, A. T., & Anderson, J. R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253–278.

Costa, E. B., Fonseca, B., Santana, M. A., Araújo, F. F. de, & 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

Coughlin, L. D., & Patel, V. L. (1987). Processing of critical information by physicians and medical students. *Journal of Medical Education*, *62*(10), 818–828.

Croskerry, P. (2009). Clinical cognition and diagnostic error: Applications of a dual process model of reasoning. *Advances in Health Sciences Education : Theory and Practice*, *14 Suppl 1*, 27–35. https://doi.org/10.1007/s10459-009-9182-2

Csapó, B., & Funke, J. (2017). *The Nature of Problem Solving*. OECD. https://doi.org/10.1787/9789264273955-en

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

Cutrer, W. B., Sullivan, W. M., & Fleming, A. E. (2013). Educational strategies for improving clinical reasoning. *Current Problems in Pediatric and Adolescent Health Care*, *43*(9), 248–257. https://doi.org/10.1016/j.cppeds.2013.07.005

Damashek, M. (1995). Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science (New York, N.Y.)*, *267*(5199), 843–848. https://doi.org/10.1126/science.267.5199.843

Daniel, M.], Rencic, J., Durning, S. J., Holmboe, E. S., Santen, S. A., Lang, V., Ratcliffe, T., Gordon, D., Heist, B., Lubarsky, S., Estrada, C. A., Ballard, T., Artino, A. R., Da Sergio Silva, A., Cleary, T., Stojan, J., & Gruppen, L. D. (2019). Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. *Academic Medicine: Journal of the Association of American Medical Colleges*, *94*(6), 902–912. https://doi.org/10.1097/ACM.0000000000002618

Davier, M. von, Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in Psychometric Population Models for Technology-Based Large-Scale Assessments: An Overview of Challenges and Opportunities. *Journal of Educational and Behavioral Statistics*, *44*(6), 671–705. https://doi.org/10.3102/1076998619881789

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 (London, England : 1995)*, *24*(3), 252–256. https://doi.org/10.1016/j.radi.2018.03.005

De Boeck, P., & Scalise, K. (2019). Collaborative Problem Solving: Processing Actions, Time, and Performance. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.01280

Dörner, D. (1975). Wie Menschen eine Welt verbessern wollten [How people wanted to improve a world]. *Bild Der Wissenschaft*, *12*, 48–53.

Drachsler, H., & Greller, W. (2016). Privacy and analytics. In D. Gašević, G. Lynch, S. Dawson, H. Drachsler, & C. Penstein Rosé (Eds.), *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16* (pp. 89–98). ACM Press. https://doi.org/10.1145/2883851.2883893

Dumas, D., Torre, D. M., & Durning, S. J. (2018). Using Relational Reasoning Strategies to Help Improve Clinical Reasoning Practice. *Academic Medicine: Journal of the Association of American Medical Colleges*, *93*(5), 709–714. https://doi.org/10.1097/ACM.0000000000002114

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, *6*(5), 178–190. https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8

Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2020). Using process data to explain group differences in complex problem solving. *Journal of Educational Psychology*, *112*(8), 1546–1562. https://doi.org/10.1037/edu0000446

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

Ercikan, K., Guo, H., & He, Q. (2020). Use of Response Process Data to Inform Group Comparisons and Fairness Research. *Educational Assessment*, *25*(3), 179–197. https://doi.org/10.1080/10627197.2020.1804353

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning using examinee response processes for the next generation of assessments. Routledge.* Routledge.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine : Journal of the Association of American Medical Colleges*, *79*(10 Suppl), S70-81. https://doi.org/10.1097/00001888-200410001-00022

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and Exceptional Performance : Evidence of Maximal Adaptation to Task Constraints: Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*(1), 273–305.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. https://doi.org/10.1037//0033-295x.87.3.215

European Parliament. (2023, August 6). *EU AI Act: first regulation on artificial intelligence*. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Accessed: 28.11.2024

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

Ferguson, R., Hoel, T., Scheffel, M., & Drachsler, H. (2016). Guest Editorial: Ethics and Privacy in Learning Analytics. *Journal of Learning Analytics*, *3*(1), 5–15.

Fink, M. C., Heitzmann, N., Reitmeier, V., Siebeck, M., Fischer, F., & Fischer, M. R. (2023). Diagnosing virtual patients: The interplay between knowledge and diagnostic activities. *Advances in Health Sciences Education : Theory and Practice*, 1–20. https://doi.org/10.1007/s10459-023-10211-4

Fiore, S. M., Graesser, A. C., & Greiff, S. (2018). Collaborative problem-solving education for the twenty-first-century workforce. *Nature Human Behaviour*, *2*(6), 367–369. https://doi.org/10.1038/s41562-018-0363-y

Fischer, A., Greiff, S., & Funke, J. (2011). The Process of Solving Complex Problems. *The Journal of Problem Solving*, *4*(1), 19–42.

Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction*, *12*(2), 213–232. https://doi.org/10.1016/S0959-4752(01)00005-6

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. J., 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

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M. R., Girwidz, R., Obersteiner, A., Reiss, K., Stürmer, K., Siebeck, M., Schmidmaier, R., Seidel, T., Ufer, S., Wecker, C., & Neuhaus, B. J. (2018). Systematizing Professional Knowledge of Medical Doctors and Teachers: Development of an Interdisciplinary Framework in the Context of Diagnostic Competences. *Education Sciences*, *8*(4), 207. https://doi.org/10.3390/educsci8040207

Francis, M., Avoseh, M. B. M., Card, K., Newland, L., & Streff, K. (2023). Student Privacy and Learning Analytics: Investigating the Application of Privacy Within a Student Success Information System in Higher Education. *Journal of Learning Analytics*, *10*(3), 102–114.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Funke, J., Fischer, A., & Holt, D. V. (2018). Competencies for Complexity: Problem Solving in the Twenty-First Century. In E. Care, P. Griffin, & M. Wilson (Eds.), *Educational Assessment in an Information Age. Assessment and Teaching of 21st Century Skills* (pp. 41–53). Springer International Publishing. https://doi.org/10.1007/978-3-319-65368-6_3

Funke, J., & Frensch, P. A. (2007). Complex Problem Solving: The European Perspective—10 Years After. In D. H. Jonassen (Ed.), *Learning to Solve Complex Scientific Problems.* Routledge.

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

Geary, D. C., Nicholas, A., Li, Y., & Sun, J. (2017). Developmental Change in the Influence of Domain-General Abilities and Domain-Specific Knowledge on Mathematics Achievement: An Eight-Year Longitudinal Study. *Journal of Educational Psychology*, *109*(5), 680–693. https://doi.org/10.1037/edu0000159

Goldhammer, F., Hahnel, C., & Kroehne, U. (2020). Analysing Log File Data from PIAAC. In D. B. Maehler & B. Rammstedt (Eds.), *Methodology of Educational Measurement and Assessment. Large-Scale Cognitive Assessment* (pp. 239–269). Springer International Publishing. https://doi.org/10.1007/978-3-030-47515-4_10

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: on validating the interpretation of process indicators based on log data. *Large-Scale Assessments in Education*, *9*(1). https://doi.org/10.1186/s40536-021-00113-5

Goldhammer, F., Naumann, J., & Keßel, J. (2013). Assessing individual differences in basic computer skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment*, *29*(4), 263–275.

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.), *Methodology of Educational Measurement and Assessment. Competence Assessment in Education* (pp. 407–425). Springer International Publishing. https://doi.org/10.1007/978-3-319-50030-0_24

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 : A Journal of the American Psychological Society*, *19*(2), 59–92. https://doi.org/10.1177/1529100618808244

Graesser, A. C., Kuo, B.-C., & Liao, C.-H. (2017). Complex Problem Solving in Assessments of Collaborative Problem Solving. *Journal of Intelligence*, *5*(2). https://doi.org/10.3390/jintelligence5020010

Graesser, A. C., Sabatini, J. P., & Li, H. (2022). Educational Psychology Is Evolving to Accommodate Technology, Multiple Disciplines, and Twenty-First-Century Skills. *Annual Review of Psychology*, *73*, 547–574. https://doi.org/10.1146/annurev-psych-020821-113042

Graesser, A. C., Woll, S. B., Kowalski, D. J., & Smith, D. A. (1980). Memory for typical and atypical actions in scripted activities. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(5), 503–515. https://doi.org/10.1037//0278-7393.6.5.503

Greiff, S., Holt, D. V., & Funke, J. (2013). Perspectives on Problem Solving in Educational Assessment: Analytical, Interactive, and Collaborative Problem Solving. *The Journal of Problem Solving*, *5*(2). https://doi.org/10.7771/1932-6246.1153

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, *126*, 248–263. https://doi.org/10.1016/j.compedu.2018.07.013

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., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review*, *13*, 74–83. https://doi.org/10.1016/j.edurev.2014.10.002

Griffin, P., & Care, E. (Eds.). (2015). *Assessment and Teaching of 21st Century Skills*. 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: The Voice of Scholarship in Education*, *111*(9), 2055–2100. https://doi.org/10.1177/016146810911100905

Han, A., Krieger, F., Borgonovi, F., & Greiff, S. (2023). Behavioral patterns in collaborative problem solving: a latent profile analysis based on response times and actions in PISA 2015. *Large-Scale Assessments in Education*, *11*(1). https://doi.org/10.1186/s40536-023-00185-5

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, *166*, Article 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., & Davier, M. von. (2015). Identifying Feature Sequences from Process Data in Problem-Solving Items with N-Grams. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Springer Proceedings in Mathematics & Statistics. Quantitative Psychology Research* (Vol. 140, pp. 173–190). Springer International Publishing. https://doi.org/10.1007/978-3-319-19977-1_13

Hegland, P. A., Aarlie, H., Strømme, H., & Jamtvedt, G. (2017). Simulation-based training for nurses: Systematic review and meta-analysis. *Nurse Education Today*, *54*, 6–20. https://doi.org/10.1016/j.nedt.2017.04.004

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., Schmidmaier, R., Neuhaus, B. J., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019).

Facilitating Diagnostic Competences in Simulations in Higher Education A Framework and a Research Agenda. *Frontline Learning Research*, *7*(4), 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

Hesse, F. W., 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* (pp. 37–56). Springer Netherlands.

Hetmanek, A., Engelmann, K., Opitz, A., & Fischer, F. (2018). Beyond Intelligence and Domain Knowledge. In F. Frank, C. Clark A., E. Katharina, O. Jonathan, F. Fischer, C. A. Chinn, K. Engelmann, & J. Osborne (Eds.), *Scientific Reasoning and Argumentation* (pp. 203–226). Routledge. https://doi.org/10.4324/9780203731826-12

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.* https://doi.org/10.31234/osf.io/3hnr6

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin*, *121*(1), 43–64. https://doi.org/10.1037/0033-2909.121.1.43

Hooftman, J., Dijkstra, A. C., Suurmeijer, I., van der Bij, A., Paap, E., & Zwaan, L. (2024). Common contributing factors of diagnostic error: A retrospective analysis of 109 serious adverse event reports from Dutch hospitals. *BMJ Quality & Safety*, *33*(10), 642–651. https://doi.org/10.1136/bmjqs-2022-015876

Hruska, P., Krigolson, O., Coderre, S., McLaughlin, K., Cortese, F., Doig, C., Beran, T., Wright, B., & Hecker, K. G. (2016). Working memory, reasoning, and expertise in medicine-insights into their relationship using functional neuroimaging. *Advances in Health Sciences Education : Theory and Practice*, *21*(5), 935–952. https://doi.org/10.1007/s10459-015-9649-2

Ifenthaler, D., & Greiff, S. (2021). Leveraging Learning Analytics for Assessment and Feedback. In J. Liebowitz (Ed.), *Online Learning Analytics* (pp. 1–18). Auerbach Publications.

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

Järvelä, S., & Hadwin, A. F. (2013). New Frontiers: Regulating Learning in CSCL. *Educational Psychologist*, *48*(1), 25–39. https://doi.org/10.1080/00461520.2012.748006

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

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Khalil, M., Prinsloo, P., & Slade, S. (2023a). Fairness, Trust, Transparency, Equity, and Responsibility in Learning Analytics. *Journal of Learning Analytics*, *10*(1), 1–7. https://doi.org/10.18608/jla.2023.7983

Khalil, M., Prinsloo, P., & Slade, S. (2023b). The use and application of learning theory in learning analytics: a scoping review. *Journal of Computing in Higher Education*, *35*(3), 573–594. https://doi.org/10.1007/s12528-022-09340-3

Kiesewetter, J., Ebersbach, R., Tsalas, N., Holzer, M., Schmidmaier, R., & Fischer, M. R. (2016). Knowledge is not enough to solve the problems - The role of diagnostic knowledge in clinical reasoning activities. *BMC Medical Education*, *16*(1), 303. https://doi.org/10.1186/s12909-016-0821-z

Kiesewetter, J., Fischer, F., & Fischer, M. R. (2017). Collaborative Clinical Reasoning-A Systematic Review of Empirical Studies. *The Journal of Continuing Education in the Health Professions*, *37*(2), 123–128. https://doi.org/10.1097/CEH.0000000000000158

Kiesewetter, J., Sailer, M., Jung, V. M., Schönberger, R., Bauer, E., Zottmann, J. M., Hege, I., Zimmermann, H., Fischer, F., & Fischer, M. R. (2020). Learning clinical reasoning: How virtual patient case format and prior knowledge interact. *BMC Medical Education*, *20*(1), 73–83. https://doi.org/10.1186/s12909-020-1987-y

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, *92*(1), 109–129. https://doi.org/10.1037/0033-295X.92.1.109

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). A Cognitive Load Approach to Collaborative Learning: United Brains for Complex Tasks. *Educational Psychology Review*, *21*(1), 31–42. https://doi.org/10.1007/s10648-008-9095-2

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

Klahr, D., & Dunbar, K. (1988). Dual Space Search During Scientific Reasoning. *Cognitive Science*, *12*(1), 1–48. https://doi.org/10.1207/s15516709cog1201_1

Klerk, S. de, Veldkamp, B. P., & Eggen, T. J. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education*, *85*, 23–34. https://doi.org/10.1016/j.compedu.2014.12.020

Knight, S., & Buckingham Shum, S. (2017). Theory and Learning Analytics. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of Learning Analytics* (pp. 17–22). Society for Learning Analytics Research (SoLAR). https://doi.org/10.18608/hla17.001

Koedinger, K. R., & Anderson, J. R. (1990). Abstract Planning and Perceptual Chunks: Elements of Expertise in Geometry. *Cognitive Science*, *14*(4), 511–550. https://doi.org/10.1207/s15516709cog1404_2

Kollar, I., Fischer, F., & Hesse, F. W. (2006). Collaboration Scripts – A Conceptual Analysis. *Educational Psychology Review*, *18*(2), 159–185. https://doi.org/10.1007/s10648-006-9007-2

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* (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*(1), 3–34.

Larson, J. R., Christensen, C., Franz, T. M., & Abbott, A. S. (1998). Diagnosing groups: The pooling, management, and impact of shared and unshared case information in team-based medical decision making. *Journal of Personality and Social Psychology*, *75*(1), 93–108. https://doi.org/10.1037//0022-3514.75.1.93

Lee, Y.-H., Hao, J., Man, K., & Ou, L. (2019). How Do Test Takers Interact With Simulation-Based Tasks? A Response-Time Perspective. *Frontiers in Psychology*, *10*, Article 906. https://doi.org/10.3389/fpsyg.2019.00906

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning Analytics in Higher Education—A Literature Review. In A. Peña-Ayala (Ed.), *Studies in Systems, Decision and Control. Learning Analytics: Fundaments, Applications, and Trends* (Vol. 94, pp. 1–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-52977-6_1

Leutner, D. (1993). Guided discovery learning with computer-based simulation games: Effects of adaptive and non-adaptive instructional support. *Learning and Instruction*, *3*(2), 113–132. https://doi.org/10.1016/0959-4752(93)90011-N

Levy, R. (2020). Implications of considering Response Process Data for Greater and Lesser Psychometrics. *Educational Assessment*, *25*(3), 218–235. https://doi.org/10.1080/10627197.2020.1804352

Li, M., Liu, H., Cai, M., & Yuan, J. (2024). Estimation of individuals' collaborative problem solving ability in computer-based assessment. *Education and Information Technologies*, *29*(1), 483–515. https://doi.org/10.1007/s10639-023-12271-w

Lindner, M. A., & Greiff, S. (2023). Process Data in Computer-Based Assessment. *European Journal of Psychological Assessment*, *39*(4), 241–251. https://doi.org/10.1027/1015-5759/a000790

Liu, L., Hao, J., Davier, A. A. von, Kyllonen, P., & Zapata-Rivera, J.-D. (2016). A Tough Nut to Crack: Measuring Collaborative Problem Solving. In J. Keengwe, 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

Lubarsky, S., Dory, V., Audétat, M.-C., Custers, E. J. F. M., & Charlin, B. (2015). Using script theory to cultivate illness script formation and clinical reasoning in health professions education. *Canadian Medical Education Journal*, *6*(2), e61-e70. https://doi.org/10.36834/cmej.36631

Lucas, H. C., Upperman, J. S., & Robinson, J. R. (2024). A systematic review of large language models and their implications in medical education. *Medical Education*. https://doi.org/10.1111/medu.15402

Ma, Y., Zhang, H., Ni, L., & Da Zhou (2023). Identifying collaborative problem-solver profiles based on collaborative processing time, actions and skills on a computer-based task. *International Journal of Computer-Supported Collaborative Learning.* Advance online publication. https://doi.org/10.1007/s11412-023-09400-5

Maddox, B. (2023). *The uses of process data in large-scale educational assessments*. OECD Education Working Papers (No. 286). OECD Publishing. https://doi.org/10.1787/5d9009ff-en

Makary, M. A., & Daniel, M. (2016). Medical error-the third leading cause of death in the US. *BMJ (Clinical Research Ed.)*, *353*, i2139. https://doi.org/10.1136/bmj.i2139

Matcha, W., Gašević, D., Uzir, N. A., Jovanović, J., & Pardo, A. (2019). Analytics of Learning Strategies. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 461–470). ACM. https://doi.org/10.1145/3303772.3303787

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *The Journal of Applied Psychology*, *85*(2), 273–283. https://doi.org/10.1037/0021-9010.85.2.273

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). W H Freeman/Times Books/ Henry Holt & Co.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. *Handbook of Educational Psychology*, 47–62.

Mayer, R. E., & Wittrock, M. C. (2006). Problem Solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology*. Macmillian.

Michaelides, M. P., Militsa, G. I., & Demetris, A. (2024). The impact of filtering out rapid-guessing examinees on PISA 2015 country rankings. *Psychological Test and Assessment Modeling*, *66*, 50–62. https://doi.org/10.2440/001-0012

Minn, S., Vie, J.-J., Takeuchi, K., Kashima, H., & Zhu, F. (2022). Interpretable Knowledge Tracing: Simple and Efficient Student Modeling with Causal Relations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12810–12818. https://doi.org/10.1609/aaai.v36i11.21560

Mislevy, R. J. (2019). Advances in Measurement and Cognition. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 164–182. https://doi.org/10.1177/0002716219843816

Mislevy, R. J., Haertel, G. D., Riconsente, M. M., Wise Rutstein, D., & Ziker, C. (2017). *Assessing Model-Based Reasoning using Evidence- Centered Design*. Springer International Publishing. https://doi.org/10.1007/978-3-319-52246-3

Molnar, C., Casalicchio, G., & Bischl, B. (2018). iml: An R package for Interpretable Machine Learning. *Journal of Open Source Software*, *3*(26), 786. https://doi.org/10.21105/joss.00786

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. In I. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Guidotti, Ö. Özgöbek, R. P. Ribeiro, R. Gavaldà, J. Gama, L. Adilova, Y. Krishnamurthy, P. M. Ferreira, D. Malerba, I. Medeiros, M. Ceci, G. Manco, E. Masciari, . . . J. A. Gulla (Eds.), *Communications in Computer and Information Science. ECML PKDD 2020 Workshops* (Vol. 1323, pp. 417–431). Springer International Publishing. https://doi.org/10.1007/978-3-030-65965-3_28

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program. *IFIP Congress*, *Vol. 256*, 64. http://bitsavers.trailing-edge.com/pdf/rand/ipl/p-1584_report_on_a_general_problem-solving_program_feb59.pdf

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Prentice-Hall.

Nicolay, B., Krieger, F., Stadler, M., Gobert, J., & Greiff, S. (2021). Lost in transition – Learning analytics on the transfer from knowledge acquisition to knowledge application in complex problem solving. *Computers in Human Behavior*, *115*, Article 106594. https://doi.org/10.1016/j.chb.2020.106594

Noroozi, O., Biemans, H. J., 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

Novick, L. R., & Bassok, M. (2005). *Problem Solving*. Cambridge University Press.

OECD. (2010). *Pisa 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*. *PISA*. OECD Publishing. https://doi.org/10.1787/9789264062658-en

OECD. (2013). *Pisa 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. *PISA*. Organisation for Economic Co-operation and Development. http://gbv.eblib.com/patron/FullRecord.aspx?p=1188961

OECD. (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, revised edition*. PISA, OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 Results (Volume V): Collaborative Problem Solving*. *PISA*. OECD. https://doi.org/10.1787/9789264285521-en

Oezsoy, M., Pickal, A. J., Brandl, L., Richters, C., Engelmann, K., Neuhaus, B., Stadler, M., Schmidmaier, R., Fischer, M. R., Fischer, F., Wecker, C., & Sailer, M. (2024, August). *Collaborative Diagnostic Reasoning in Simulations: Insights into Teacher and Medical Education*. EARLI Sig 6 & 7 Conference, Tübingen, Germany.

Oliveri, M. E., Lawless René, & Molloy Hillary. (2017). *A Literature Review on Collaborative Problem Solving for College and Workforce Readiness:* (ETS GRE®).

O'Neil, H. F. (1999). Perspectives on computer-based performance assessment of problem solving. *Computers in Human Behavior*, *15*, 255–268.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive Load Theory: Instructional Implications of the Interaction between Information Structures and Cognitive Architecture. *Instructional Science*, *32*(1/2), 1–8.

Paas, F., van Gog, T., & Sweller, J. (2010). Cognitive Load Theory: New Conceptualizations, Specifications, and Integrated Research Perspectives. *Educational Psychology Review*, *22*(2), 115–121. https://doi.org/10.1007/s10648-010-9133-8

Patel, V. L., Arocha, J. F., & Kaufman, D. R. (1994). Diagnostic Reasoning and Medical Expertise. In *Psychology of Learning and Motivation. Advances in Research and Theory* (Vol. 31, pp. 187–252). Elsevier. https://doi.org/10.1016/S0079-7421(08)60411-9

Patel, V. L., Arocha, J. F., & Zhang, J. (2005). Thinking and reasoning in medicine. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 727–750). Cambridge University Press.

Patel, V. L., & Groen, G. J. (1986). Knowledge Based Solution Strategies in Medical Reasoning. *Cognitive Science*, *10*, 91–116.

Patel, V. L., Groen, G. J., & Arocha, J. F. (1990). Medical expertise as a function of task difficulty. *Memory & Cognition*, *18*(4), 394–406.

Patel, V. L., Kaufman, D. R., & Arocha, J. F. (2002). Emerging paradigms of cognition in medical decision-making. *Journal of Biomedical Informatics*, *35*(1), 52–75. https://doi.org/10.1016/S1532-0464(02)00009-6

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing What Students Know: The Science and Design of Educational Assessment.* National Academy Press.

Perkins, D. N., & Salomon, G. (1989). Are Cognitive Skills Context-Bound? *Educational Researcher*, 16–25.

Piccarreta, R. (2017). Joint Sequence Analysis. *Sociological Methods & Research*, *46*(2), 252–287. https://doi.org/10.1177/0049124115591013

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 *Proceedings of the International Conference on Computer-supported for Collaborative Learning, 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/cscl2023.638463

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

Provasnik, S. (2021). Process data, the new frontier for assessment development: rich new soil or a quixotic quest? *Large-Scale Assessments in Education*, *9*(1). https://doi.org/10.1186/s40536-020-00092-z

Puntambekar, S., & Hubscher, R. (2005). Tools for Scaffolding Students in a Complex Learning Environment: What Have We Gained and What Have We Missed? *Educational Psychologist*, *40*(1), 1–12. https://doi.org/10.1207/s15326985ep4001_1

Radkowitsch, A., Fischer, M. R., Schmidmaier, R., & Fischer, F. (2020). Learning to diagnose collaboratively: validating a simulation for medical students. *GMS Journal for Medical Education*, *37*(5).

Radkowitsch, 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.), *Springer Briefs in Education. Learning to diagnose with simulations - Examples from teacher education and medical education.* (pp. 116–134). Springer.

Radkowitsch, 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

Radkowitsch, A., Sommerhoff, D., Fischer, F., & Ufer, S. (2023, September). *Diagnosekompetenzen und ihre Entwicklung – ein theoretisches Modell und seine (wissenschafts)praktischen Implikationen [Diagnostic competences and their development - a theoretical model and its implications for science and practice.]*. Presentation at the 19th Section Meeting on Educational Psychology (PAEPS) 2023,, Kiel, Germany.

Rane, N. L., Mallick, S. K., Kaya, Ö., & Rane, J. (2024). Techniques and optimization algorithms in machine learning: A review. In N. L. Rane, S. K. Mallick, Ö. Kaya, & J. Rane (Eds.), *Applied Machine Learning and Deep Learning: Architectures and Techniques*. Deep Science Publishing. https://doi.org/10.70593/978-81-981271-4-3_2

Renkewitz, F., & Heene, M. (2019). The Replication Crisis and Open Science in Psychology: Methodological Challenges and Developments. *Zeitschrift Für Psychologie*, *227*(4), 233–236. https://doi.org/10.1027/2151-2604/a000389

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 *Proceedings of the International Conference on Computer-supported for Collaborative Learning, 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/cscl2023.596913

Richters, C., Stadler, M., Radkowitsch, 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 (Chair), *International Society of the Learning Sciences*, Hiroshima, Japan.

Richters, C., Stadler, M., Radkowitsch, 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). https://doi.org/10.1186/s40536-023-00151-1

Rikers, R. M., Schmidt, H. G., & Boshuizen, H. P. (2000). Knowledge Encapsulation and the Intermediate Effect. *Contemporary Educational Psychology*, *25*(2), 150–166. https://doi.org/10.1006/ceps.1998.1000

Rochelle, J., & Teasley, S. (1995). The Construction of Shared Knowledge in Collaborative Problem Solving. In C. O'Malley (Ed.), *Computer-Supported Collaborative Learning* (pp. 66–97). Springer.

Rosen, Y. (2014). Comparability of Conflict Opportunities in Human-to-Human and Human-to-Agent Online Collaborative Problem Solving. *Technology, Knowledge and Learning*, *19*(1-2), 147–164. https://doi.org/10.1007/s10758-014-9229-1

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

Rupp, A. A., Levy, R., Dicerbo, K. E., Sweet, S. J., Calico, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into Practice: The Interplay of Theory and Data in Evidence Models within a Digital Learning Environment. *Journal of Educational Data Mining Special Issue*, *4*(2), 49–102.

Schäfer, J., Reuter, T., Karbach, J., & Leuchter, M. (2024). Domain-specific knowledge and domain-general abilities in children's science problem-solving. *The British Journal of Educational Psychology*, *94*(2), 346–366. https://doi.org/10.1111/bjep.12649

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

Schmidt, H. G., & Boshuizen, H. P. (1993a). On acquiring expertise in medicine. *Educational Psychology Review*, *5*(3), 205–221.

Schmidt, H. G., & Boshuizen, H. P. (1993b). On the origin of intermediate effects in clinical case recall. *Memory & Cognition*, *21*(3), 338–351.

Schmidt, H. G., Boshuizen, H. P., & Hobus, P. P. M. (1988). Transitory Stages In the Development of Medical Expertise The "Intermediate Effect" In Clinical Case Representation Studies. *Roceedings of the Annual Meeting of the Cognitive Science Society*, *10(0)*, 139–145.

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

Sclater, N. (2016). Developing a Code of Practice for Learning Analytics. *Journal of Learning Analytics*, *3*(1), 16–42. https://doi.org/10.18608/jla.2016.31.3

Siewiorek, A., & Gegenfurtner, A. (2010). Leading to Win: The Influence of Leadership Styles on Team Performance during a Computer Game Training. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010)* (pp. 524–531). International Society of the Learning Sciences.

Simbeck, K. (2024). They shall be fair, transparent, and robust: auditing learning analytics systems. *AI and Ethics*, 4(2), 555–571. https://doi.org/10.1007/s43681-023-00292-7

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

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*, Article jcal.12847. Advance online publication. 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*, Article 777. https://doi.org/10.3389/fpsyg.2019.00777

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education*, *157*, 103964. https://doi.org/10.1016/j.compedu.2020.103964

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. *Computers in Human Behavior*, *111*, 106442. https://doi.org/10.1016/j.chb.2020.106442

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

Stasser, G., & Stewart, D. (1992). Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *Journal of Personality and Social Psychology*, *63*(3), 426–434. https://doi.org/10.1037//0022-3514.63.3.426

Stegmann, K., Pilz, F., Siebeck, M., & Fischer, F. (2012). Vicarious learning during simulations: Is it more effective than hands-on training? *Medical Education*, *46*(10), 1001–1008. https://doi.org/10.1111/j.1365-2923.2012.04344.x

Sugrue, B. (1995). A Theory-Based Framework for Assessing Domain-Specific Problem-Solving Ability. *Educational Measurement: Issues and Practice*, *14*(3), 29–35. https://doi.org/10.1111/j.1745-3992.1995.tb00865.x

Sun, C., Shute, V. J., Stewart, A. E., Beck-White, Q., Reinhardt, C. R., Zhou, G., Duran, N., & D'Mello, S. K. (2022). The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Computers in Human Behavior*, *128*, 107120. https://doi.org/10.1016/j.chb.2021.107120

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. K. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, Article 103672. https://doi.org/10.1016/j.compedu.2019.103672

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive Load Theory. Advance online publication. https://doi.org/10.1007/978-1-4419-8126-4

Swiecki, Z., Ruis, A. R., Farrell, C., & Shaffer, D. W. (2020). Assessing individual contributions to Collaborative Problem Solving: A network analysis approach. *Computers in Human Behavior*, *104*, 105876. https://doi.org/10.1016/j.chb.2019.01.009

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* (pp. 191–200). Routledge. https://doi.org/10.4324/9781315617572-19

Tang, S., Siju, S., & Zhen, L. (2023). Detecting Atypical Test-Taking Behavior with Behavior Prediction Using LSTM. *Psychological Test and Assessment Modeling*, *65*(1), 76–124.

Teasley, S. D. (1997). Talking About Reasoning: How Important Is the Peer in Peer Collaboration? In L. B. Resnick, R. Säljö, C. Pontecorvo, & B. Burge (Eds.), *Discourse, Tools and Reasoning* (pp. 361–384). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-03362-3_16

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

Thille, C., Kizilee, R. F., Piech, Christopher, Halawa, Scherif A., & Greene, D. K. (2014). The Future of Data–Enriched Assessment. *Research & Practice in Assessment*, *9*, 5–16.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, *143*, 103676. https://doi.org/10.1016/j.compedu.2019.103676

Tsai, Y.-S., Rates, D., Moreno-Marcos, P. M., Muñoz-Merino, P. J., Jivet, I., Scheffel, M., Drachsler, H., Delgado Kloos, C., & Gašević, D. (2020). Learning analytics in European higher education—Trends and barriers. *Computers & Education*, *155*, 103933. https://doi.org/10.1016/j.compedu.2020.103933

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. (2023). 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 der Linden, W. J. (2008). Using Response Times for Item Selection in Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5–20. https://doi.org/10.3102/1076998607302626

van Joolingen, W. R., & Jong, T. de (1997). An extended dual search space model of scientific discovery learning. *Instructional Science*, *25*(5), 307–346. https://doi.org/10.1023/A:1002993406499

van Lehn, K. (1989). Problem solving and cognitive skill acquisition. In M. I. Posner (Ed.), *Foundations of Cognitive Science* (pp. 526–579). M. I. T. Press.

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98–110.

Vicente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, *105*(1), 33–57. https://doi.org/10.1037//0033-295x.105.1.33

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., Radkowitsch, A., Schmidmaier, R., . . . Noroozi, O. (2023). Transactivity and Knowledge Co-Construction in Collaborative Problem Solving. In *Proceedings of the International Conference on Computer-supported for Collaborative Learning, 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 and society: The development of higher psychological processes.* Harvard University Press.

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 New York. https://doi.org/10.1007/978-1-4612-4634-3_9

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, *46*(1), 71–95. https://doi.org/10.1016/j.compedu.2005.04.003

Whitelock-Wainwright, A., Gašević, D., Tejeiro, R., Tsai, Y.-S., & Bennett, K. (2019). The Student Expectations of Learning Analytics Questionnaire. *Journal of Computer Assisted Learning*, *35*(5), 633–666. https://doi.org/10.1111/jcal.12366

Wise, A. F., & Shaffer, D. W. (2015). Why Theory Matters More than Ever in the Age of Big Data. *Journal of Learning Analytics*, *2*(2), 5–13. https://doi.org/10.18608/jla.2015.22.2

Wise, S. L. (2017). Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wong, J., Baars, M., Koning, B. B. de, van der Zee, T., Davis, D., Khalil, M., Houben, G.-J., & Paas, F. (2019). Educational Theories and Learning Analytics: From Data to Knowledge. In D. Ifenthaler, D.-K. Mah, & J. Y.-K. Yau (Eds.), *Utilizing Learning Analytics to Support Study Success* (pp. 3–25). Springer International Publishing. https://doi.org/10.1007/978-3-319-64792-0_1

Wong, L., Meyer, G., Timson, E., Perfect, P., & White, M. (2012). Objective and subjective evaluations of flight simulator fidelity. *Seeing and Perceiving*, *25*(0), 91. https://doi.org/10.1163/187847612X647108

Wood, D., Bruner, J. S., & Ross, G. (1976). The Role of Tutoring in Problem Solving. *J. Child Psychol. Psychiat*, *17*, 89–100.

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, *8*(10), 1839–1850. https://doi.org/10.1038/s41562-024-02004-5

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and Ethical Challenges of Large Language Models in Education: A Systematic Literature Review. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.13379

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *12*(6), 1–23. https://doi.org/10.1177/1745691617693393

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 : Journal of the Association of American Medical Colleges*, *90*(11 Suppl), S56-62. https://doi.org/10.1097/ACM.0000000000000900

Zhan, P., & Qiao, X. (2022). Diagnostic Classification Analysis of Problem-Solving Competence using Process Data: An Item Expansion Method. *Psychometrika*. Advance online publication. https://doi.org/10.1007/s11336-022-09855-9

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), Doc35. https://doi.org/10.3205/zma001181

Zumbo, B. D., Maddox, B., & Care, N. M. (2023). Process and Product in Computer-Based Assessments. *European Journal of Psychological Assessment*, *39*(4), 252–262. https://doi.org/10.1027/1015-5759/a000748

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *The Behavioral and Brain Sciences*, 1–50. https://doi.org/10.1017/S0140525X17001972

APPENDIX

8

Laura Brandl

8.1 Patient Cases Used in Paper 2 and 3

For Paper 2 and 3, fictitious but authentic patient cases were developed with medical experts in the project team. Table 3 shows an overview of which patient cases were used in Paper 2 and 3.

Table 3

Overview About all Patient Cases Used in Paper 2 and 3

Patient	Diagnosis	Usage	
		Paper 2	Paper 3
Marianne Freundorf	Acute pancreatitis	Study A	
Herma Goettlich	Aspiration pneumonia	Study B Study C	Case 1
Sabine Winkler	Community acquired pneumonia (CAP)		Case 2
Anton Fomin	Acute tuberculosis		Case 3
Mark Binder	Pneumocystis jirovecii Pneumonia (PJP)		Case 4
Maria Schenker	Hospital acquired pneumonia		Case 5

Afterwards, the health record of one patient case is presented exemplarily (Table 4) along a screenshot how the patient case is presented in the CoSiMed simulation (Figure 4).

Table 4

Example Case: Herma Goettlich

Health Record Section	Content
Introduction	You have been working in a medium-sized regional hospital for a few months and are currently working on a general internal medicine ward. Today you are also in charge of the emergency department. In the late hours of Monday morning, 78-year-old Herma Goettlich is brought in by the ambulance, accompanied by her worried husband. Mrs Goettlich is suffering from severe shortness of breath, so her husband answers most of her questions. You have taken blood samples and ‘hastily’ sent them to the laboratory, Mrs Goettlich has her medical history taken and examined as far as possible. By the time you have finished, part of the lab is also ready and you can use the file to decide what the next diagnostic steps should be.

Ambulance 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 does not consider it worse than usual. There are no B symptoms. *Pre-existing Conditions* • History of media infarction (middle cerebral artery infarction) in 12/2017 • 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/m²). *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.

Physical Examination

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

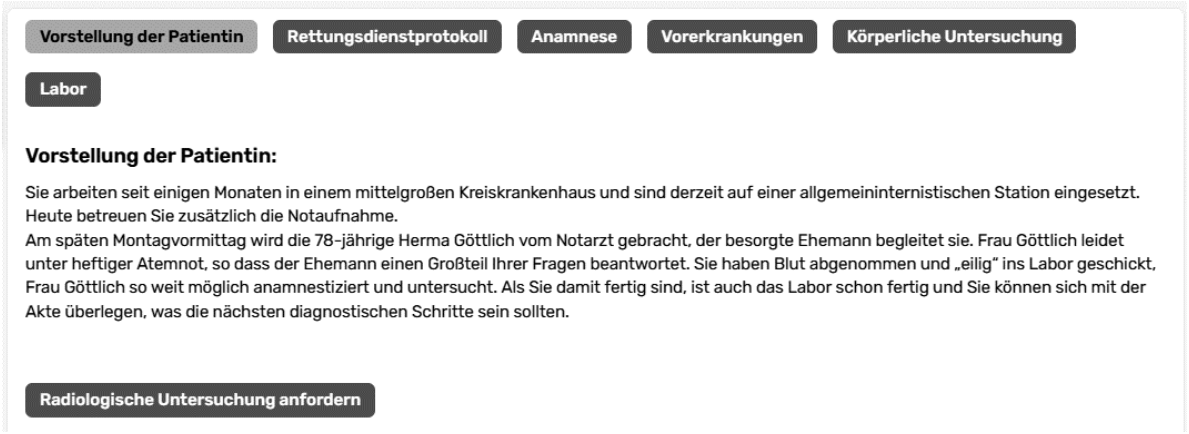
Parameter	**Value**	**Reference range (f)**
Blood Count		
Erythrocytes	3.8 X 10^6 /µl	3.5 - 5 X 10^6 /µl

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 103 /µl	4 - 11 X 10^3 /µl
Platelets	182,000 /µl	150,000 - 400,000 /µl
Reticulocytes	1%	0.5 - 2 %
Differential Blood Count		
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 %
Coagulation		
Quick	100%	70 - 120%
INR	1	1
PTT	38 sec.	28 - 40 sec.
Serum		
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/1.73m^2	>60 ml/min/1.73 m^2
Urea	>60 ml/min/1.73 m^2	>60 ml/min/1.73 m^2
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
y-GT	8 U/l	< 17 U/l
a-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 µg/l	15 - 250 µg/l
Erythrocyte Sedimentation Rate	1.8 µU/ml	0.2 - 3.1 µU/ml
Urine-Stick		

pH	5	5-7
Protein	-	-
Bilirubin	-	-
Urobilinogen	-	-
Nitrite	-	-
Glucose	-	-
Acetone	-	-
Blood	-	-

Figure 4

Screenshot of the Introduction to the Health Record of Herma Goettlich



8.2 Measures for Individual Characteristics Used in Paper 2

Content Knowledge

Content knowledge was assessed in Paper 2 by conceptual (Boshuizen & Schmidt, 1992) and strategic knowledge (Stark et al., 2011) of radiology and internal medicine, respectively. The items in each construct were presented in a randomized way in each study. However, the items for study C were shortened due to the embedding of the data collection in the curriculum (Table 5).

Table 5

Overview of Number of Questions in the Content Knowledge Test

	Conceptual knowledge		Strategic knowledge	
Study	internal medicine	radiology	internal medicine	radiology
Study A	20	15	24 8 cases 3 questions per case	16 8 cases 2 questions per case
Study B	20	15	24 8 cases 3 questions per case	16 8 cases 2 questions per case
Study C	13	12	24 8 cases 3 questions per case	12 6 cases 2 questions per case

Conceptual Knowledge

Conceptual knowledge was measured using single-choice questions including 5 options adapted from a database of examination questions from the Medical Faculty of the LMU Munich, focusing pathophysiology, disease triggers, and radiologic interpretation of relevant and closely related diagnoses of the patient cases used in the simulation. A mean score of 0-1 was calculated, representing the percentage of correct answers and indicating the average conceptual

knowledge of the participant per medical knowledge domain. Below is an example item from internal medicine, the correct answer is in bold:

The treatment of an acute, febrile respiratory infection is based on the pathogen spectrum (viral/bacterial). From this point of view, the use of antibiotics is indicated in approximately

- less than 10% of cases
- 10-20% of cases
- 21-30% of cases
- 31-50% of cases
- over 50% of cases

Strategic Knowledge

Strategic content knowledge was measured contextually using key features questions (Fischer et al., 2005). Short cases were introduced followed by two to three follow up questions (e.g., What is your most likely suspected diagnosis?, What is your next examination?, What treatment do you choose?). Each question had eight possible answers, from which the learners were asked to choose one. A mean score of 0-1 was calculated, representing the percentage of correct responses, indicating the average strategic content knowledge of the participant per domain. This is an example item from radiology, correct answers to the question are written in bold:

It's Monday morning in your radiology practice. 24-year-old Karin Ungenau comes to see you. She is referred by her neurologist with suspected multiple sclerosis.

Which of the following imaging procedures do you carry out to confirm the diagnosis?

- CT and MRI with cranial contrast agent
- CT and MRI with spinal contrast agent
- CT and MRI with spinal and cranial contrast agent
- MRI with cranial contrast agent
- MRI with contrast agent spinal
- MRI native cranial
- MRI native spinal
- **MRI native and with contrast agent spinal and cranial**

Assume that Mrs Ungenau has been deaf since birth and received cochlear implants (CI) on both sides as a child. *What do you need to consider for the MRI that is now required?*

- The implant must be switched off while the examination is running
- The implant must be reprogrammed before the examination

- The patient needs special headphones for the examination
- The patient must be monitored for 24 hours after the examination
- The examination must generally be refused
- The examination must be performed under anaesthesia
- No special precautions need to be taken
- **Before the examination, the exact fabric of the CI must be known and compatibility must be clarified**

Collaboration Knowledge

Collaboration knowledge measured specific to the simulated task and consistent across the three studies used in Paper 2 as meta-cognitive knowledge about the collaboration partner (Engelmann & Hesse, 2011). Collaboration knowledge was measured 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 process. Participants were asked to select all relevant information for seven different patient cases with the cardinal symptom fever (internal medicine). The patient cases were presented in a randomized order and always included twelve pieces of information regarding the chief complaints, medical history and physical examination of the patient cases. We then assessed whether each piece of information was (not) shared correctly (i.e. whether relevant information was shared and irrelevant information was not shared) and assigned one point and divided it by the maximum of 12 points to standardized the range of measure to 0-1. This is an example case:

> 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.

From the information provided below, please select the details that you would communicate to a radiologist for the Emergency CCT (correct answers are in bold):

Condition after fall from ladder	**Impaired vigilance**
Multiple episodes of vomiting	Reduced left eye aperture, right eye slim
Shallow breathing Additional information	**Contusion on the left forehead**

Abrasions on the left side	Respiratory rate 20/min
Pulse 90/min Physical examination	Blood pressure 100/65 mmHg
Extremities inspection unremarkable	**Anisocoria**

Social Skills

Social skills were measured consistently across the three studies in paper 2 based on self-report on a 6-point Likert scale ranging from total disagreement to total agreement. The construct was measured using 23 questions divided into five subscales. Five questions aimed to measure the overall construct, the other four subscales were identified using the complex problem solving-frameworks of Liu et al. (2016) and Hesse et al. (2015): perspective taking (four questions), information sharing (five questions), negotiation (four questions) and coordination (five questions). For the final score, the mean of all subcategories is calculated, ranging from 1 to 6, representing general social skills. Table 6 shows the example items per subscale.

Table 6

Example Items for Each Subscale for Measuring Social Skills

Subscale	Item
Direct Measurement	I enjoy working with others.
Perspective Taking	It is easy for me to put myself in the position of my collaboration partners.
Information Sharing	When I collaborate with others, I purposefully share relevant information.
Negotiating	I can negotiate compromises when working with others.
Coordination	When I work with others, we have a clear common goal in mind.

References:

Boshuizen, H. P., & Schmidt, H. G. (1992). On the Role of Biomedical Knowledge in Clinical Reasoning by Experts, Intermediates and Novices. *Cognitive Science*, 16, 153–184.

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. 10.1016/j.chb.2011.06.002

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. 10.1080/01421590500078471

Hesse, F. W., 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* (pp. 37–56). Springer Netherlands.

Liu, L., Hao, J., Davier, A. A. von, Kyllonen, P., & Zapata-Rivera, J.-D. (2016). A Tough Nut to Crack: Measuring Collaborative Problem Solving. In J. Keengwe, 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

8.3 Measures for Collaborative Diagnostic Activities Used in Paper 2

After reading the health record, the learners' task was to collaborate with an agent-based radiologist through requesting radiological examinations using a request form (Figure 5). For better readability the request form is translated into English in Figure 6 and the collaborative diagnostic activities derived from the CDR-M are mapped to the respective part of the request form.

Figure 5

Screenshot of the Radiological Request Form of Herma Goettlich

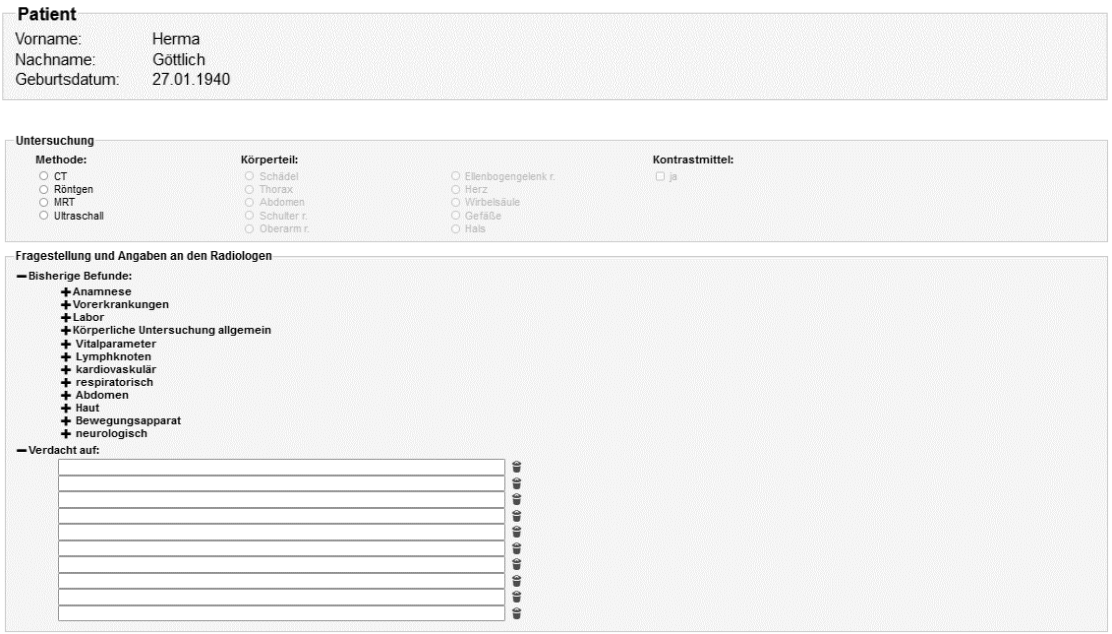
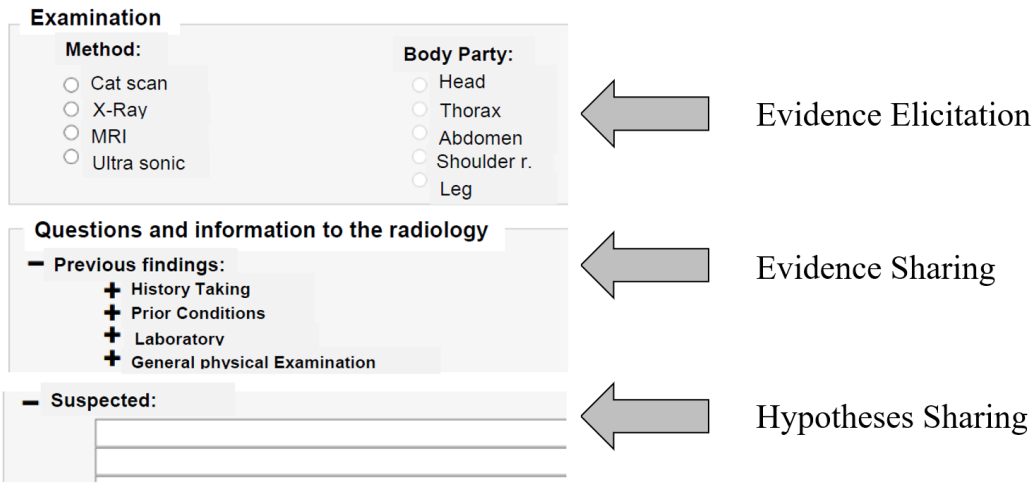


Figure 6

Collaborative Diagnostic Activities Mapped to the Request Form



Quality of Evidence Elicitation

The quality of evidence elicitation was measured by assessing the appropriateness of the requested radiological examination for the indicated diagnosis. Therefor an expert solution showed which radiological examinations were appropriate for each of the possible diagnoses. If participants requested an appropriate radiological examination for the indicated diagnoses, they received 1 point for that request attempt. Finally, a mean score across all request attempts (maximum of 3) was calculated and scored. The final mean score was transformed into a binary indicator, due to the categorical nature of the original data and its skewed distribution, with a majority of responses concentrated in a single category. Thus, 1 indicates that all requested radiological examinations were appropriated and 0 indicates that also inappropriate radiological examinations were requested. For example, the appropriate radiological examination for diagnosing aspiration pneumonia, the accurate diagnosis of the example case Herma Goettlich are: X-ray thorax, CT thorax native, ultrasound abdomen.

Quality of Evidence Sharing

The quality of evidence sharing was measured using a precision indicator. This was calculated as the proportion of shared relevant evidence out of all shared evidence. Relevant evidence is defined per case and per diagnosis and indicated by the expert solution. The precision indicator was first calculated per radiological request. We then calculated the mean score, summarizing all attempts in that patient case. This resulted in a range from 0 points, indicating that only irrelevant evidence was shared, to 1 point, indicating that only relevant evidence was shared. For example, the relevant evidence for requesting a radiological exam for the accurate diagnosis of the example case Herma Goettlich are:

- Shortness of breath
- Rapid onset
- Started this morning
- History of nicotine abuse, 10 pack-years
- Severe sweating
- History of tonsillectomy, 1962
- Fever since this morning
- Sudden onset of shortness of breath
- Dysphagia (difficulty swallowing)
- Initial pO_2 92%
- Leukocytes 13.6 x $10^3/\mu l$

- CRP 53 mg/dl
- Erythrocyte sedimentation rate 10/23
- Reduced general condition (AZ)
- Temperature 37.9°C
- Respiratory rate 27/min
- pO_2 96% with 2 liters of O_2
- Lungs with coarse crackles on the right side
- Cough with sputum
- Foul-smelling sputum

Quality of Hypotheses Sharing

The quality of hypotheses sharing was measured using a precision score indicating how many of the diagnoses that the participants shared with the radiologist were actually relevant to the case. Therefore, the participant could choose out of a long menu of 249 diagnoses. For example, this are the 36 relevant diagnoses for the example case Herma Goettlich in the original German language along the English translation:

- Alveolitis
- Alveolitis, exogen allergisch (EAA) – Extrinsic Allergic Alveolitis (EAA)
- Autoimmunes Geschehen – Autoimmune Disorder
- Bronchitis – Bronchitis
- Bronchitis, bakteriell akut – Acute Bacterial Bronchitis
- Bronchitis, viral akut – Acute Viral Bronchitis
- COPD – Chronic Obstructive Pulmonary Disease (COPD)
- COPD, akut exazerbiert – Acute Exacerbation of COPD
- COPD, chronisch – Chronic COPD
- Degeneratives Geschehen – Degenerative Disorder
- Entzündliches Geschehen – Inflammatory Disorder
- Grippaler Infekt – Upper Respiratory Tract Infection (Common Cold)
- Herzinsuffizienz – Heart Failure
- Herzinsuffizienz, akut bei Myokardinfarkt/Herzinfarkt – Acute Heart Failure Due to Myocardial Infarction
- Herzinsuffizienz, akut bei Myokarditis – Acute Heart Failure Due to Myocarditis
- Herzinsuffizienz, chronisch, akut dekompensiert – Chronic Heart Failure, Acutely Decompensated

- Infekt – Infection
- Infekt, bakteriell – Bacterial Infection
- Infekt, viral – Viral Infection
- Influenza/Grippe – Influenza
- Ischämie, Lungenarterienembolie, Lungenembolie – Ischemia, Pulmonary Artery Embolism, Pulmonary Embolism
- Mykobakteriose, atypisch – Atypical Mycobacteriosis
- Pneumonie/Lungenentzündung – Pneumonia
- Pneumonie/Lungenentzündung, Aspirationspneumonie – Aspiration Pneumonia
- Pneumonie/Lungenentzündung, atypisch – Atypical Pneumonia
- Pneumonie/Lungenentzündung, bakteriell – Bacterial Pneumonia
- Pneumonie/Lungenentzündung, begleitend bei systemischem Wurmbefall – Pneumonia with Systemic Parasitic Infestation
- Pneumonie/Lungenentzündung, CAP – Community-Acquired Pneumonia (CAP)
- Pneumonie/Lungenentzündung, Pilzpneumonie – Fungal Pneumonia
- Pneumonie/Lungenentzündung, Pneumocystis jirovecii Pneumonie (PCP) – Pneumocystis jirovecii Pneumonia (PCP)
- Pneumonie/Lungenentzündung, viral – Viral Pneumonia
- Pneumothorax – Pneumothorax
- Pneumothorax, spontan – Spontaneous Pneumothorax
- Pneumothorax, traumatisch – Traumatic Pneumothorax
- Rheumatisches Fieber – Rheumatic Fever
- Sarkoidose – Sarcoidosis
- Sepsis/Blutvergiftung – Sepsis
- Thrombose, tiefe Beinvenenthrombose (TVT) – Deep Vein Thrombosis (DVT)

8.4 Measures for Diagnostic Outcomes Used in Paper 2

In principle, the same coding schemes for the diagnostic outcomes (diagnostic accuracy and diagnostic justification) were used for across the studies used in paper 2. However, during the course of the studies, the coding schemes were revised in collaboration with the medical experts in the project. Therefore, the coding schemes for the same cases differ slightly between the studies. The most recent coding scheme is provided below.

Diagnostic Accuracy

A main diagnosis was assigned to each patient case as expert solution. Participants typed in the first three letters of their desired diagnosis and then received suggestions from a list of 249 possible diagnoses. Diagnostic accuracy was then calculated by coding the agreement between the final diagnosis given and the expert solution. Accurate diagnoses (e.g., hospital-acquired pneumonia) were coded as 1, correct but inaccurate diagnoses (e.g., pneumonia) were coded as 0.5, and incorrect diagnoses were coded as 0. A binary indicator was used for the final diagnostic accuracy score, with 0 indicating an incorrect diagnosis and 1 indicating an at least inaccurate diagnosis, due to the categorical nature of the original data and its skewed distribution, with a majority of responses concentrated in a single category.

The exact coding instructions were as follows:

- The learner receives 1 point for recognizing the accurate diagnosis.
- If the accurate diagnosis is stated in the justification, this is scored as usual, but only if no diagnosis was previously stated in the diagnosis field.
- If a diagnosis is first stated inaccurately (i.e. 0.5 points) or incorrectly (i.e. 0 points) and then specified more precisely in the justification, the score for diagnostic accuracy is not changed.
- If the main diagnosis is missing, -99 is entered as the missing value (do not enter 0).

Referring to the example case Herma Goettlich, the accurate diagnosis (1 point) was Aspiration Pneumonia, however these diagnoses were still coded as correct but inaccurate (0.5 points): Pneumonia, Bacterial Pneumonia, Community-Acquired Pneumonia (CAP), and Atypical Pneumonia. All other diagnoses were coded as incorrect (0 points)

Diagnostic Justification

A prerequisite for diagnostic justification is the provision of at least an inaccurate diagnosis (diagnostic accuracy coded with at least 0.5). If a participant provided an incorrect diagnosis (coded as 0), diagnostic justification was immediately scored as 0. After choosing a final

diagnosis, participants were asked to justify their decision in an open text field. Diagnostic justification was then calculated as the proportion of relevant reported information out of all relevant information that would have fully justified the final accurate diagnosis. Again, medical experts agreed on an expert solution that included all relevant information to justify the correct diagnosis. The participants' solution was coded by two independent coders, each coding the full data, and differences in coding were discussed until the coders agreed. The exact coding instructions were as follows:

- The learner receives 1 point for each aspect of justification mentioned by the learner that also appears in the expert solution (including synonyms).
- The points are then divided by the maximum number of points that can be achieved, so that the learner can receive a score of between 0 and 1 point. Only the raw scores are entered into the coding table. The percentage score is then computed.
- The justification is only coded if the diagnosis is correctly identified (diagnostic accuracy coded with at least 0.5). If the justification is not coded, -66 is entered. Important: But always check whether a diagnosis was given in the justification.
- If the justification is missing when the diagnosis is correct (diagnostic accuracy coded with at least 0.5), -99 is entered. If the justification is missing when the diagnosis is incorrect (diagnostic accuracy coded with 0)., -66 is entered. If the diagnosis is missing, -66 is entered.
- Expressions such as 'no high fever' count as 'fever' unless 'no fever' is specifically written.
- Where signs of infection are mentioned, the reference to the laboratory should be noted.

Referring to the example case Herma Goettlich there were nine relevant aspects of justification in the expert solution (without synonyms): (1) Dyspnea, (2) tachypnoea, (3) fever, (4) reduced SpO2, (5) cough with foul-smelling sputum, (6) dysphagia, (7) coarse-bubbling rales, (8) inflammation values increased/ infection parameters increased, (9) X-ray/CT chest: compressions or consolidations or shadows

Diagnostic Efficiency

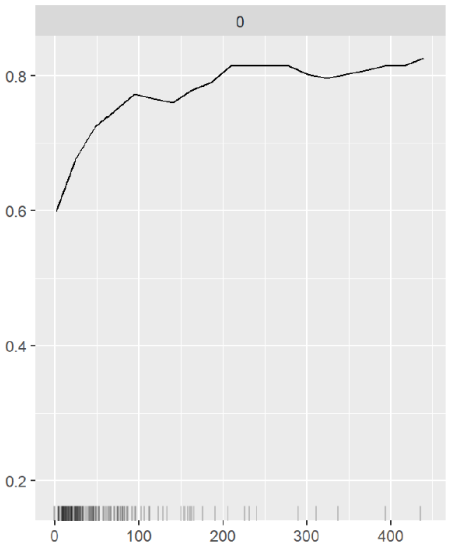
A prerequisite for diagnostic efficiency is the provision of at least an inaccurate diagnosis (diagnostic accuracy coded with at least 0.5). Diagnostic efficiency was then calculated by dividing the non-binary version of diagnostic accuracy by the minutes required to solve the patient case. If a participant provided an incorrect diagnosis (coded as 0), diagnostic efficiency was immediately scored as 0.

8.5 Partial Dependence Plots Used in Additional Analyses of Paper 3

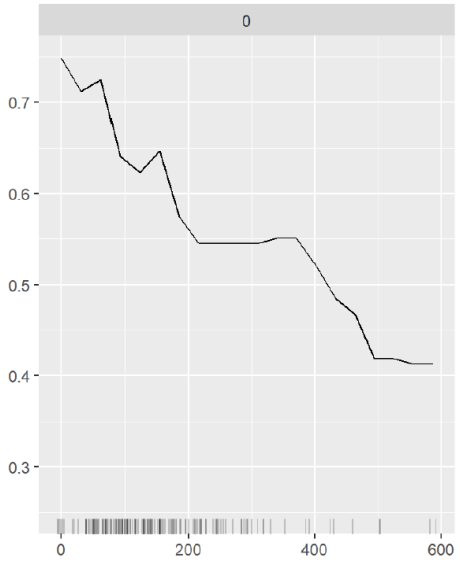
Partial dependence plots were used in additionally analyses of Paper 3, Table 7 depicts them for each bigrams of collaborative diagnostic activities used as a feature to predict diagnostic accuracy. On the x-axis the discrete frequency of the feature can be seen, while the y-axis displays the likelihood of the model to predict 0, indicating an inaccurate diagnosis. For example, looking at the first feature, the time spent with evidence elicitation, the partial dependence plot indicates that the likelihood of the model to predict an inaccurate diagnosis is around 0.6 for no time spent with evidence elicitation. When the time increases also the likelihood of the model to predict an inaccurate diagnosis increases up to 0.8 when this feature occurred more than 200 times in a process. The lines on the x-axis further indicate how often each value of the feature occurred across processes (i.e. learners working on a patient case). When looking, again on the first feature this indicates that most often learners had between 0 and 100 bigrams of evidence elicitation followed by evidence elicitation in their process per case.

Table 7

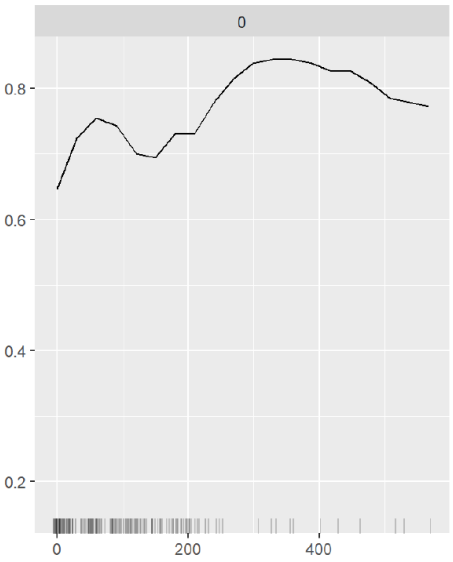
Partial Dependence Plots for Bigrams of Collaborative Diagnostic Activities Predicting Diagnostic Accuracy

Feature	Partial Dependence Plot
Time spent with evidence elicitation	

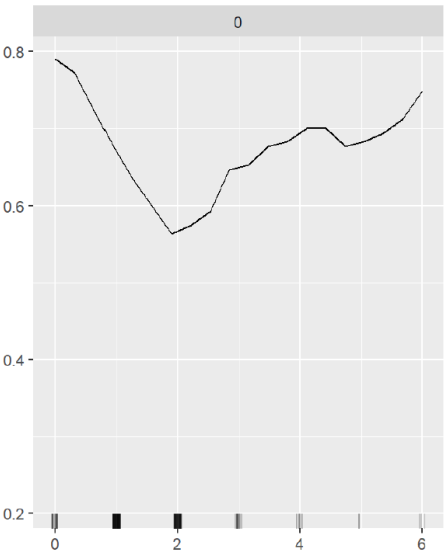
Time spent with evidence sharing



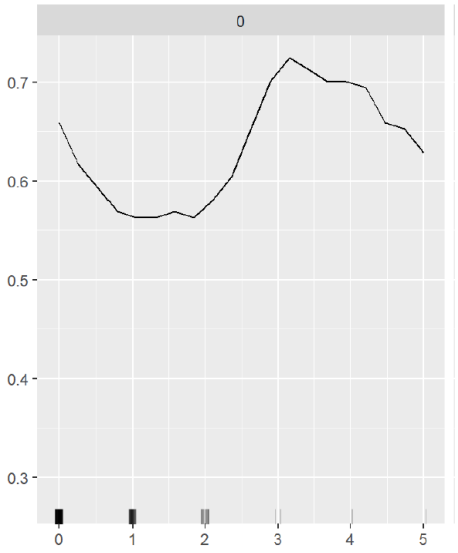
Time spent with hypotheses sharing



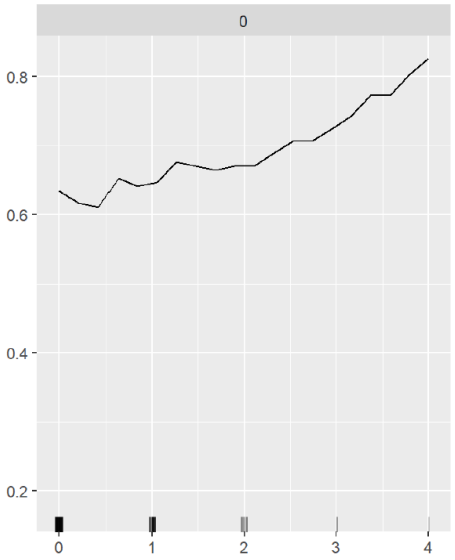
Transition from evidence elicitation to evidence sharing



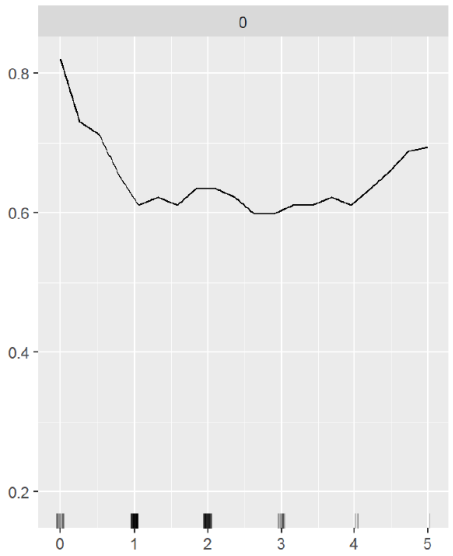
Transition from evidence elicitation to hypotheses sharing



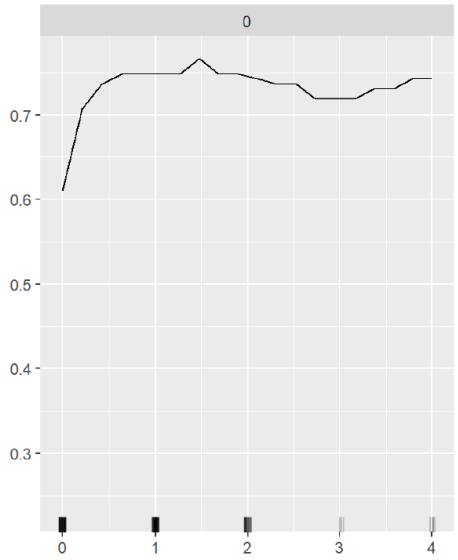
Transition from evidence sharing to evidence elicitation



Transition from evidence sharing to hypotheses sharing



Transition from hypotheses sharing to evidence elicitation



Transition from hypotheses sharing to evidence sharing