Diagnostic Reasoning and Argumentation

Analysis and Facilitation Using Simulation-Based Learning Environments in Medical Education and Teacher Education

---

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

am Munich Center of the Learning Sciences

der Ludwig-Maximilians-Universität

München

vorgelegt von

Anna Elisabeth Bauer

München, Dezember 2021

1st Supervisor / Erstgutachter: Prof. Dr. Frank Fischer, LMU München

2nd Supervisor / Zweitgutachter: Prof. Dr. Martin Fischer, LMU Klinikum

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Stefan Ufer, LMU München

Prof. Dr. Michael Sailer, FAU Erlangen-Nürnberg

Tag der mündlichen Prüfung: 23.02.2022

Acknowledgments

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 my supervisors. In particular, I would like to thank my first supervisor, Frank Fischer: You inspired, challenged, and shaped my thinking, not only concerning the research that has been included in this thesis, but also beyond, in my professional and personal development throughout the recent years. Likewise, I am grateful for having Martin Fischer as my second supervisor: Thanks for provided critical comments, further directions, and encouraging words, whenever they were needed. I am also thankful for my third supervisor, David Shaffer; especially for welcoming me as a guest in his lab for several weeks. Thanks to him and his team, I will always remember this stay as a great experience. On a related note, and in general, I would also like to take the opportunity to express my appreciation for the funding that I have received by the Elite Network of Bavaria.

The day of my defense is yet to come; however, I would already like to thank Stefan Ufer and Michael Sailer, who agreed to take part in my examination committee. Beyond, Michael Sailer deserves a special thanks for taking over various other roles in my time as a PhD student, being my *daily supervisor*, role model, research buddy, and regular lifesaver, just to name a few; at all hours (of the day), he had a friendly ear for my manifold questions, thoughts, and concerns. I owe you so much more than a hard copy of this thesis. I am also thankful for the great experiences in FAMULUS, with my colleagues Jan Kiesewetter, Jan Zottmann, Jonas Pfeiffer, and Claudia Schulz. It was a joy to collaborate with you on this exciting research project. A big thank you is dedicated towards the many student assistants, who supported the work in FAMULUS: Hanna Mißbach, Simon Eichler, Maike Achtner, Julia Glas, Dorothee Pietsch, Aldin Alijagic, and Julia Schnagl. Besides FAMULUS, I always enjoyed and benefited from the connection to the research group COSIMA. I would like to thank especially my colleagues and peers Anika Radkowitsch, Amadeus Pickal, and Maximilian Fink: Thank you for your companionship in doing the PhD, your friendship, and the enjoyable time that we spent in the office, on conference trips, and in the beer garden. Moreover, thank you to all the colleagues from the REASON doctoral school for sharing the path towards doing a PhD. Special thanks is to Sarah Bichler for encouraging me to apply for a PhD position after all and for sharing wonderful conversations and memories, such as the late night practice session with Anika and me before my first conference presentation. Also special thanks is to Anna Horrer and Carolin Auner for your ongoing encouragement, companionship, and enjoyable conversations. Another thank you is dedicated to all the

colleagues from our research unit, who were supporting me and my doctoral research in various regards: to Matthias Stadler for your repeated help in all sorts of statistics problems; to Heidi Gesell and Antonia Gayed for sharing conversations and collaborating on side projects; to Georg Kuschel, who was always available for solving any technical concerns, such as preparing the technical setup for our laboratory studies; to Rosa Haas und Simone Steiger for handling all sorts of administrative issues and keeping track of the general chaos; to Alexa Krickel and Irina Ciobanu for taking over all sorts of PhD issues.

Not only in the work context but also outside from work, I have received great support and encouragement from numerous people, who are part of or crossed my path. Special thanks is to my longtime friend Katharina Winkler: You are not only one of the greatest supporters in my life, my *catalyzer*; but also one of my greatest critics, whenever I need to pull myself together. A true friend. Further, special thanks is to my dear friend Müge Arslantürk: Sometimes I feel like we are the same person, on two different sides of a mirror. Feeling that we are so much alike and understanding each other so well is a great source of energy at all times.

Last but not least, I am eternally grateful for the unconditional love and support that I receive by my family, my parents Birgit and Franz Bauer, and my siblings, Benedikt, Matthias, and Katharina Bauer. I feel like you are my safe base for all that I achieved so far and for all that is about to come.

Executive Summary

Diagnostic reasoning refers to the systematic collection and interpretation of problem-specific information with the goal of reducing uncertainty in solving a problem. Diagnostic reasoning is highly relevant in various professional contexts. Therefore, it is important to teach diagnostic reasoning in related areas of higher education. Diagnostic reasoning has been researched in medical education and teacher education in particular. However, only recently has research been started to systematically integrate the theoretical and empirical advancements of the two fields. The research presented in this thesis aims to contribute to the endeavor of developing a cross-disciplinary research perspective on diagnostic reasoning, to deepen the understanding of the relevant knowledge and skills, and ultimately, to further improve the teaching and learning of diagnostic reasoning in medical and teacher education. The thesis discusses the commonalities and differences in diagnostic reasoning in medical and teacher education with respect to diagnostic problems, epistemic processing, and cognitive structures and processes. In addition, further directions for researching and facilitating diagnostic reasoning are identified: Using simulation-based learning environments, cross-disciplinary comparisons of diagnostic activities and diagnostic practices seem to be a promising approach to identifying field-specific epistemic ideals, standards, and processes involved in diagnostic reasoning. A second research direction is the differentiation of diagnostic reasoning skills or diagnostic competences. In this thesis, diagnostic reasoning skills are suggested to be distinguished concerning the two subskills of diagnostic judgment and diagnostic argumentation. Diagnostic judgment aims to achieve diagnostic accuracy in solving diagnostic problems and has been investigated in prior research. In contrast, the newly defined concept of diagnostic argumentation, including the facets of justification, disconfirmation, and transparency, aims to achieve a common understanding with others by comprehensibly and persuasively explaining interpretations and conclusions made about a diagnostic problem. Differentiating between diagnostic reasoning skills includes a third research direction, that is, investigating whether facilitating students' learning of these skills requires specific learning opportunities. Using simulation-based learning has been found to facilitate students' learning of diagnostic reasoning, however, only if students receive sufficient support, such as adaptive feedback. Adaptive feedback on written task solutions can be automated by making use of recent technological advancements in Natural Language Processing (NLP). Moreover, when learning collaboratively, students can provide each other additional feedback and exchanging arguments while solving simulated diagnostic problems, which may not only foster diagnostic judgment but also diagnostic argumentation.

The three identified research directions are addressed in the three papers that are presented in the empirical part of this thesis. The first study compared diagnostic activities and diagnostic practices in medical education and teacher education. The second study explored the differentiation of diagnostic reasoning skills into diagnostic judgment and diagnostic argumentation in the context of teacher education. The third study investigated the effects of NLP-based automatic adaptive feedback and collaborative learning on preservice teachers' simulation-based learning of diagnostic judgment and diagnostic argumentation.

The first study investigated diagnostic activities and diagnostic practices in a cross-disciplinary comparison of students in a medical education program and students in a teacher education program. The students worked in a simulation-based learning environment in which they were confronted with eight simulated diagnostic problems from their respective fields. After making a diagnosis, they wrote a justificatory report for each simulated case in which they described their approach to solving the case. All justificatory reports were coded for four diagnostic activities: generating hypotheses, generating evidence, evaluating evidence, and drawing conclusions. Diagnostic practices were operationalized as the relative frequencies of co-occurring diagnostic activities by using the novel method of Epistemic Network Analysis. Significant differences were found between the medical students and preservice teachers with respect to both their diagnostic activities and their diagnostic practices. The medical students put relatively more emphasis on generating hypotheses and drawing conclusions, therefore applying a more hypothesis-driven approach. Preservice teachers focused on generating and evaluating evidence, indicating a more data-driven approach. These results may be explained by different epistemic ideals and standards taught in higher education in the two fields.

The second study explored the suggested differentiation between diagnostic judgment and diagnostic argumentation. In addition, the three facets of justification, disconfirmation, and transparency were investigated as potential subskills of diagnostic argumentation. Teacher education was selected as the context for an initial investigation. The justificatory reports collected from the preservice teachers during the first study were reanalyzed. In addition, the analyses included the preservice teachers' prior diagnostic knowledge and the accuracy of their diagnostic judgments. The correlational results supported the assumption that making accurate diagnostic judgments and formulating diagnostic argumentations may represent different diagnostic reasoning skills. Moreover, because of the disparities found in the underlying knowledge bases, the results supported the notion that justification, disconfirmation, and transparency may represent distinct subskills of diagnostic

argumentation. However, preservice teachers' diagnostic argumentations rarely involved all three facets, suggesting a need for more specific training.

The third study in this thesis is an experimental study of the effects of NLP-based automatic adaptive feedback and collaborative learning on preservice teachers' simulation-based learning of diagnostic reasoning skills, which were operationalized as the accuracy of diagnostic judgments and the quality of justifications in diagnostic argumentations. NLP-based automatic adaptive feedback was compared with an expert solution as a form of static feedback. Further, the social mode of learning was experimentally varied, in that students learned either individually or collaboratively as dyads. The results showed that, compared to static feedback, adaptive feedback facilitated the quality of preservice teachers' justifications in diagnostic argumentation. Moreover, adaptive feedback helped collaborative learners to achieve the same level of accuracy of diagnostic judgments as individual learners. Therefore, adaptive feedback may have helped collaborative learners to cope with the possibly higher demands induced by the collaborative learning situation.

The research presented in this thesis suggests that medical education and teacher education have developed specific diagnostic practices, which might relate to field-specific ideals and standards that are internalized throughout higher education. In addition to content-specific diagnostic knowledge, the knowledge of field-specific ideals and standards may be considered part of the knowledge base on which (future) professionals perform diagnostic reasoning. Further research is needed to advance the understanding of field-specific epistemic ideals and standards in diagnostic reasoning. Moreover, future research may address to what extent the knowledge of field-specific ideals and standards explains the performance and learning of diagnostic reasoning skills and, in particular, of diagnostic argumentation. The analyses suggested that the standardization of diagnostic reasoning in teacher education might be less advanced compared to that of medical education. Justification, disconfirmation, and transparency may provide a starting point for the further development of standards for diagnostic argumentation. Future research in medical education may address the replicability of distinguishing diagnostic judgment, indicated by diagnostic accuracy, from diagnostic argumentation, including the three facets of justification, disconfirmation, and transparency. Distinguishing between the two diagnostic reasoning skills seems to be practically relevant, in that adaptive feedback was especially important for preservice teachers' simulation-based learning of diagnostic argumentation. NLP methods provide particular benefits in automating support measures for facilitating the learning of complex reasoning skills, such as diagnostic argumentation, even in short-term interventions.

Deutsche Zusammenfassung

Diagnostisches Denken bezeichnet das systematische Sammeln und Interpretieren von problembezogenen Informationen mit dem Ziel, Unsicherheit zu reduzieren und eine Problemlösung zu identifizieren. Diagnostisches Denken ist in verschiedenen beruflichen Kontexten und auch in der Hochschulbildung zukünftiger Fachkräfte von hoher Relevanz. Diagnostisches Denken wurde in der Forschung bisher insbesondere in den Bereichen der medizinischen Ausbildung und der Lehrerbildung berücksichtigt. Jedoch wurde die Integration theoretischer und empirischer Fortschritte der beiden Felder bisher weitestgehend vernachlässigt. Die in dieser Dissertation präsentierte Forschung soll einen Beitrag zu den jüngsten Bestrebungen einer disziplinübergreifenden Forschungsperspektive auf das diagnostische Denken zu entwickeln. In diesem Zusammenhang wird zudem angestrebt, das Verständnis von relevantem Wissen und Fertigkeiten vertiefen und letztendlich das Lehren und Lernen diagnostischen Denkens in der medizinischen Ausbildung und der Lehrerbildung weiter zu verbessern. In der Dissertation werde Gemeinsamkeiten und Unterschiede des diagnostischen Denkens in der der medizinischen Ausbildung und der Lehrerbildung in Bezug auf diagnostische Problemstellungen, epistemische Verarbeitung sowie kognitive Strukturen und Prozesse diskutiert. Dabei werden weitere Richtungen für die Erforschung und auch für die Förderung des diagnostischen Denkens aufgezeigt: Die Nutzung simulationsbasierter Lernumgebungen für interdisziplinäre Vergleiche von diagnostischen Aktivitäten und diagnostischen Praktiken erscheint vielversprechend, um feldspezifische epistemische Ideale, Standards und Prozesse im diagnostischen Denken zu identifizieren. Eine zweite Forschungsrichtung ist eine Unterscheidung der bisher undefinierten Mehrzahl diagnostischer Fertigkeiten oder Diagnosekompetenzen hinsichtlich zwei zu differenzierender Fertigkeiten: diagnostisches Urteilen und diagnostisches Argumentieren. Diagnostisches Urteilen zielt darauf ab, eine richtige Diagnosestellung bei der Lösung diagnostischer Problemstellungen zu erreichen und wurde bereits von vorangehender Forschung untersucht. Demgegenüber zielt das neu definierte Konstrukt des diagnostischen Argumentierens auf das Erreichen eines gemeinsamen Verständnisses mit anderen ab, mittels verständlicher und überzeugender Erklärung der Interpretationen und Schlussfolgerungen zu einer diagnostischen Problemstellung. Um das diagnostische Argumentieren genauer zu konzeptualisieren, werden drei Facetten eingeführt: Die Begründung diagnostischer Schlussfolgerungen, die Widerlegung alternativer Erklärungen und die Transparenz hinsichtlich des diagnostischen Vorgehens. Die Unterscheidung von diagnostischem Urteilen und diagnostischem Argumentieren impliziert eine dritte Forschungsrichtung: Diese adressiert die Frage, ob

Studierende zum Erlernen der beiden Fertigkeiten spezifische Lern- und Unterstützungsmaßnahmen benötigen. Es liegt bereits Evidenz dazu vor, dass der Einsatz von simulationsbasiertem Lernen förderlich für das Erlernen des diagnostischen Denkens ist – vorausgesetzt Studierende erhalten dabei ausreichende Unterstützung, beispielsweise in Form von adaptivem Feedback. Neue technologische Fortschritte im Bereich des Natural Language Processing (NLP) ermöglichen inzwischen die automatisierte Analyse und hierdurch das automatisierte Bereitstellen von adaptivem Feedback zu schriftlichen Aufgabenlösungen von Studierenden. Darüber hinaus können kollaborative Lernformate nützlich sein, in denen sich Studierenden gegenseitig zusätzliches Feedback geben. Kollaborative Lernformate haben zudem den Vorteil, dass Lernende bereits während des Lösungsprozesses Argumente formulieren und austauschen, was vor allem zur Förderung des diagnostischen Argumentierens, aber auch des diagnostischen Urteilens förderlich sein könnte.

Die drei aufgezeigten Forschungsrichtungen werden in drei Artikeln adressiert, die im empirischen Teil dieser Arbeit vorgestellt werden. Der erste Artikel vergleicht diagnostische Aktivitäten und diagnostische Praktiken in der medizinischen Ausbildung und der Lehrerbildung. Der zweite Artikel untersuchte die vorgeschlagene Differenzierung von diagnostischem Urteilen und diagnostischem Argumentieren im Kontext der Lehrerbildung. Der dritte Artikel untersuchte die Effekte von mittels NLP automatisiertem adaptivem Feedback und kollaborativem Lernen auf das simulationsbasierte Lernen des diagnostischen Urteilens und diagnostischen Argumentierens von Lehramtsstudierenden.

Im ersten Artikel wurde ein interdisziplinärer Vergleich diagnostischer Aktivitäten und diagnostischer Praktiken der medizinischen Ausbildung und der Lehrerbildung durchgeführt. Studierende beider Fächer bearbeiteten in einer simulationsbasierten Lernumgebung jeweils acht simulierte Fälle zu diagnostischen Problemstellungen aus ihrem jeweiligen Feld. Jeweils im Anschluss an die Diagnosestellung verfassten die Studierenden für jeden simulierten Fall eine Erklärung zu Ihrer Diagnosestellung, in der sie ihren Lösungsansatz beschreiben und begründen sollten. In allen Erklärungstexten wurden vier diagnostische Aktivitäten kodiert: Hypothesen generieren, Evidenz generieren, Evidenz evaluieren und Schlussfolgerungen ziehen. Die diagnostischen Praktiken beider Felder wurden mittels der neuen Methodik der epistemischen Netzwerkanalyse operationalisiert, welche die relativen Häufigkeiten des gemeinsamen Auftretens der diagnostischen Aktivitäten als Netzwerk abbildet. Es wurden signifikante Unterschiede zwischen Medizin- Und Lehramtsstudierenden festgestellt, sowohl hinsichtlich ihrer berichteten diagnostischen Aktivitäten als auch hinsichtlich der übergeordneten diagnostischen Praktiken. Im Vergleich

legten Medizinstudierende einen stärkeren Schwerpunkt auf das Generieren von Hypothesen und das Ziehen von Schlussfolgerungen, was im Gesamtbild einen eher hypothesen-gesteuerten Ansatz diagnostischer Praktiken ergibt. Lehramtsstudierende legten einen stärkeren Fokus auf die Generierung und Auswertung von Evidenz, was auf einen eher datengetriebenen Ansatz diagnostischer Praktiken hindeutet. Die Ergebnisse können durch unterschiedliche epistemische Ideale und Standards erklärt werden, welche im Rahmen der medizinischen Ausbildung und der Lehrerbildung gelehrt werden.

Der zweite Artikel untersuchte die vorgeschlagene Unterscheidung zwischen diagnostischem Urteilen und diagnostischem Argumentieren als zwei zu differenzierende diagnostische Fertigkeiten. Darüber hinaus wurden die drei Facetten Begründung, Widerlegung und Transparenz als potenzielle Teilfertigkeiten des diagnostischen Argumentierens untersucht. Zum Zweck einer initialen Analyse der Forschungsfragen wurde der Bereich der Lehrerbildung gewählt. Die für den ersten Artikel im Bereich der Lehrerbildung gesammelten Erklärungstexte wurden erneut analysiert. Die Analysen schlossen darüber hinaus das diagnostische Vorwissen der Lehramtsstudierenden und die Genauigkeit ihrer diagnostischen Urteile ein. Die korrelativen Ergebnisse stützten die Annahme, dass das Treffen genauer diagnostischer Urteile und das Formulieren diagnostischer Argumentationen unterschiedliche diagnostische Fertigkeiten darstellen. Darüber hinaus zeigten sich Unterschiede dahingehend, welche Wissensarten Varianz in den drei Facetten der Begründung, Widerlegung und Transparenz erklären konnten. Dies stützt die Annahme, dass Begründung, Widerlegung und Transparenz unterschiedliche Teilfertigkeiten des diagnostischen Argumentierens darstellen. Die diagnostischen Argumentationen der Lehramtsstudierenden beinhalteten jedoch nur selten alle drei Facetten. Dies könnte darauf hinweisen, dass bisher unzureichend spezifische Lerngelegenheiten in der Lehrerbildung zur Verfügung stehen.

Der dritte Artikel präsentiert eine experimentelle Studie zu den Effekten von mittels NLP automatisiertem adaptivem Feedback und kollaborativem Lernen auf das simulationsbasierte Lernen des diagnostischen Denkens von Lehramtsstudierenden. Hierbei wurden insbesondere die Genauigkeit diagnostischer Urteile und die Qualität von Begründungen im diagnostischen Argumentieren untersucht. Das automatische adaptive Feedback wurde mit einer Expertenlösung als statisches Feedback verglichen. Darüber hinaus wurde die Sozialform des Lernens experimentell variiert, indem die Studierenden entweder einzeln lernten oder kollaborativ als Dyaden. Die Ergebnisse zeigten, dass adaptives Feedback im Vergleich zu statischem Feedback zur Verbesserung der Qualität der Begründungen in den

diagnostischen Argumentationen von Lehramtsstudierenden beitragen konnte. Darüber hinaus half adaptives Feedback kollaborativen Lernenden, die gleiche Genauigkeit im diagnostischen Urteilen zu erreichen wie einzelne Lernende. Das adaptive Feedback könnte kollaborativen Lernenden geholfen haben, die möglicherweise durch die kollaborative Lernsituation höheren Anforderungen der Lernsituation zu bewältigen.

Die in dieser Dissertation präsentierten Forschungsergebnisse legen nahe, dass sich in der medizinischen Ausbildung und der Lehrerbildung spezifische diagnostische Praktiken entwickelt haben, welche möglicherweise auf fachspezifische Ideale und Standards zurückzuführen sind. Solche Ideale und Standards werden im Laude der hochschulischen Ausbildung von den Studierenden verinnerlicht. Neben inhaltsspezifischem diagnostischem Wissen sollte daher auch das Wissen um fachspezifische Ideale und Standards als Teil des professionellen Wissens angesehen werden, auf dessen Basis (zukünftige) Fachkräfte diagnostisches Denken anwenden. Weitere Forschung ist erforderlich, um das Verständnis zu fachspezifischen epistemischen Idealen und Standards im diagnostischen Denken voranzutreiben. Darüber hinaus sollte weitere Forschung untersuchen, welche Rolle das Wissen um fachspezifische Ideale und Standards in der Performanz und dem Erlernen von Fertigkeiten des diagnostischen Denkens und insbesondere des diagnostischen Argumentierens spielt. Die Befundmuster legten zudem nahe, dass die Standardisierung des diagnostischen Denkens in der Lehrerbildung im Vergleich zur medizinischen Ausbildung weniger weit fortgeschritten sein könnte. In Bezug auf Standards zum diagnostischen Argumentieren können die drei Facetten der Begründung, Widerlegung und Transparenz als Ansatzpunkt für die Weiterentwicklung von Standards dienen. Zukünftige Forschung im Bereich der medizinischen Ausbildung sollte insbesondere auch die Replizierbarkeit der Unterscheidung von diagnostischem Urteilen (gekennzeichnet durch diagnostische Genauigkeit) und diagnostischem Argumentieren, einschließlich der drei Facetten der Begründung, Widerlegung und Transparenz, adressieren. Die Unterscheidung der beiden vorgeschlagenen diagnostischen Fertigkeiten scheint zudem praktisch relevant zu sein, da sich das adaptives Feedback insbesondere für das simulationsbasierte Lernen des diagnostischen Argumentierens von Lehramtsstudierenden als effektiv zeigte. Zudem zeigten sich in den Ergebnissen die Potentiale NLP-basierter Methoden zur Automatisierung von Textanalysen für die Unterstützung des Lernens komplexer kognitiver Fertigkeiten, wie beispielsweise des diagnostischen Argumentierens, auch im Rahmen kurzfristiger Interventionen.

Table of Contents

Acknowledgments i
Executive Summary iii
Deutsche Zusammenfassung vii
Table of Contents xi

1 General Introduction 1

1.1 Aim and Structure of the Thesis 2
1.2 Diagnostic Reasoning across Different Fields: Solving Diagnostic Problems 5
1.3 Epistemic Processing in Diagnostic Reasoning: Aims, Activities, and Practices 8
1.3.1 Diagnostic Accuracy: The Central Epistemic Aim in Solving Diagnostic Problems 8
1.3.2 Diagnostic Activities: Conceptualizing Diagnostic Problem-Solving 9
1.3.3 Diagnostic Practices: Collective Patterns of Diagnostic Activities 10
1.4 Cognition in Diagnostic Reasoning: Knowledge, Processing, and Skills 12
1.4.1 Diagnostic Knowledge: A Content-Specific Basis for Diagnostic Reasoning 12
1.4.2 Cognitive Processing: Fundamental Commonalities in Diagnostic Reasoning 13
1.4.3 Diagnostic Reasoning Skills: Distinguishing Judgment and Argumentation 15
1.4.4 Conceptualizing Diagnostic Argumentation: Justification, Disconfirmation, and Transparency 16
1.5 Facilitating Diagnostic Reasoning Skills 20
1.5.1 Using Simulation-Based Learning in Higher Education 20
1.5.2 Automatic Adaptive Feedback in Simulation-Based Learning 21
1.5.3 Social Form of Learning in Simulation-Based Learning 23
1.6 General Research Questions, Methodological Considerations, and Outline of the Studies 25
1.6.1 Outline of Study 1 27
1.6.2 Outline of Study 2 28
1.6.3 Outline of Study 3 29

2 Study 1: Diagnostic Activities and Diagnostic Practices in Medical Education and Teacher Education: An Interdisciplinary Comparison 33

3 Study 2: Diagnostic Argumentation in Teacher Education: Making the Case for Justification, Disconfirmation, and Transparency 49

4 Study 3: Adaptive Feedback from Artificial Neural Networks Facilitates Pre-service Teachers' Diagnostic Reasoning in Simulation-Based Learning 67

5 General Discussion 79

5.1 Summary of the Results 80
5.1.1 Results of Study 1 80
5.1.2 Results of Study 2 81
5.1.3 Results of Study 3 83
5.2 Theoretical Implications 83
5.2.1 Advancing Cross-Disciplinary Research on Diagnostic Reasoning 83
5.2.2 Standards in Diagnostic Reasoning 86
5.2.3 Diagnostic Argumentation as a Distinct Diagnostic Reasoning Skill 88
5.2.4 Facilitating Diagnostic Reasoning Skills 91
5.3 Practical Implications 93
5.4 Limitations 96
5.5 Directions for Future Research 100

6 Conclusion 103

7 References 107

8 Appendices **127**

Appendix A Case Materials **128**
Case Materials from Teacher Education 128
Case Materials from Medical Education 129
Appendix B Coding Schemes **130**
Coding Scheme for Epistemic Diagnostic Activities 130
Coding Scheme for Differential Diagnoses 136
Coding Scheme for Diagnostic Accuracy 136
Coding Scheme for the Quality of Justification 137
Appendix C Knowledge Tests **139**
Conceptual Diagnostic Knowledge Test 139
Strategic Diagnostic Knowledge Test 139
Appendix D Feedback Intervention **142**
Adaptive Feedback 142
Static Feedback 143
Appendix E Ethical Approval **144**

Statement of Scientific Integrity **145**

1 General Introduction

1.1 Aim and Structure of the Thesis

Diagnostic reasoning is relevant in many fields (Heitzmann et al., 2019). Its learning should thus be considered an important part of the education of future professionals within these fields. A thorough understanding of it is necessary to define meaningful educational objectives and design and implement effective learning environments (Gagné & Merrill, 1990; Grossman et al., 2009). However, various research strands concerning diagnostic reasoning have been developed in different fields, using different terms and theoretical notions, which complicates the integration and (if possible) transfer of theoretical and empirical accomplishments across fields. This thesis describes research that aimed to contribute to (a) developing a cross-disciplinary research perspective on diagnostic reasoning, (b) integrating and refining the existing understanding of diagnostic reasoning skills, and (c) investigating approaches to facilitate the learning of diagnostic reasoning skills.

As the point of departure for developing a cross-disciplinary research perspective on diagnostic reasoning (see aim a), this work focused on medical education and teacher education. Future physicians, according to consensus, take over substantial responsibility for their patients' health, so they need extensive training in correctly assessing highly diverse symptomatologies. Similarly, teachers are significantly responsible for supporting their pupils' learning and development: they need to diagnose, for example, their pupils' performance, progress, and learning prerequisites (Praetorius et al., 2013). This has led to researchers and educators in teacher education increasingly acknowledging the role of diagnostic reasoning in (future) teachers' everyday practice (e.g., Herppich et al., 2018).

Medical education and teacher education share particularly interesting commonalities when it comes to diagnostic reasoning (Heitzmann et al., 2019). For instance, researchers from both fields have referred to *diagnostic competences* (e.g., Fink, Reitmeier et al., 2021; Hoth et al., 2016; Kramer, Förtsch, Boone et al., 2021; Papa, 2016) as an umbrella term for the knowledge, skills, and attitudes (see Blömeke et al., 2015) professionals need to perform competent diagnostic reasoning. *Diagnosing* "means 'recognizing exactly' or 'differentiating' and is associated with the activities and processes of classifying causes and forms of phenomena ('diagnosing', n.d.). These causes and forms are often not directly observable; they are latent or hidden and need to be identified" (Heitzmann et al., 2019, p. 3). In addition to medical education and teacher education, research in other fields, such as mechatronics, has referred to *diagnosing* (Abele, 2018). Therefore, to integrate research from different fields, this thesis investigated *diagnostic reasoning*, a "goal-oriented collection and interpretation of case-specific or problem-specific information to reduce uncertainty in order to make medical

or educational decisions" (Heitzmann et al., 2019, p. 4). Accordingly, diagnostic reasoning does not necessarily refer to a specific field or content area (e.g., reasoning in medicine) but rather denotes a certain type of reasoning that can address a wide range of problems and occur in diverse situations (e.g., Abele, 2018; Kron et al., 2021; Radkowitsch et al., 2020). Beyond diagnostic competence and diagnostic reasoning, research in medical education and teacher education, it is important to note, has, however, mostly referred to content-bound terms: *clinical reasoning* in medical education (e.g., Norman et al., 2017) and *assessment* (e.g., Herppich et al., 2018) or *judgment* (e.g., Praetorius et al., 2017) in teacher education. However, due to the broad scope of *diagnostic reasoning*, it seems reasonable to use it as a common term for elaborating on a cross-disciplinary research perspective, aiming to better understand not only the associated knowledge and skills but also the learning of diagnostic reasoning.

In addition to suggesting that the research strands from medical education and teacher education share terminological commonalities, Heitzmann et al. (2019) offered further arguments for integrating the diagnostic reasoning research of both fields (see Heitzmann et al., 2019, pp. 3–4). However, these arguments demand further elaboration and systematization, for which they can be broadly categorized as either epistemically grounded or cognitively grounded. Therefore, to (see aim a) develop a cross-disciplinary research perspective on diagnostic reasoning, this thesis further elaborated on these two main strands of epistemic and cognitive arguments.

In doing so, this thesis also aimed to (see aim b) integrate and advance the existing understanding of diagnostic reasoning skills. The medical education and teacher education literature does not clearly distinguish between different diagnostic reasoning skills; instead, it ascribes various indicators of diagnostic reasoning to a broad and not fully defined range of skills or competences (e.g., Heitzmann et al., 2019; Herppich et al., 2018). There are, however, both epistemic and cognitive arguments for distinguishing between at least two diagnostic reasoning skills, which this thesis refers to as diagnostic judgment (e.g., Loibl et al., 2020) and diagnostic argumentation. To distinguish between them demands not only elaboration of the two constructs, but also investigation, which this thesis will also undertake.

Moreover, trying to distinguish between diagnostic judgment and diagnostic argumentation also raises several questions. Especially the matter of facilitating the two diagnostic reasoning skills is important. Aiming to (see aim c) identify and investigate approaches to facilitate the learning of diagnostic reasoning skills, this thesis elaborated on the benefits of using simulation-based learning environments (Chernikova, Heitzmann,

Stadler et al., 2020; see Fink, Radkowitsch et al., 2021; Heitzmann et al., 2019) for researching and facilitating diagnostic reasoning skills. To specifically foster diagnostic judgment and diagnostic argumentation, two specific means that might support simulation-based learning of diagnostic reasoning skills are suggested: adaptive feedback and collaborative learning. However, the effects of adaptive feedback and collaborative learning on the simulation-based learning of diagnostic reasoning skills also required investigation. In particular, whether the proposed interventions have differing effects on diagnostic argumentation and diagnostic judgment must be investigated, to determine the practical impact of differentiating diagnostic reasoning skills for teaching, learning, and measurement purposes.

The thesis has three main parts. The first part discusses diagnostic reasoning in medical education and teacher education, introducing several task-related, epistemically grounded, and cognitively related reasons that explain why and in which regard different fields – medical education and teacher education, in particular – may be comparable or limited in their comparability with respect to diagnostic reasoning (see aim a). First, the object of reasoning – diagnostic problems – is characterized (see section 1.2). Diagnostic problems can be compared in terms of several characteristics – content area, exemplarity, complexity, and required activities – that can affect their processing in terms of diagnostic reasoning. These characteristics should thus be considered when researching diagnostic reasoning across different diagnostic problems. Further, an epistemically grounded view of diagnostic reasoning (see section 1.3) is elaborated on; in particular, epistemic aims (see Chinn et al., 2011), diagnostic activities (see Fischer et al., 2014), and diagnostic practices (see Bauer et al., 2020), and how comparing different fields with respect to their epistemic processing may offer specific insights into their diagnostic reasoning. This is followed by detailing the differences and similarities in the cognitive aspects of diagnostic reasoning and offering a rationale as to why and how the existing concepts may be further integrated and refined (see aim b; see section 1.4). In doing so, a differentiation of diagnostic reasoning skills into diagnostic judgment (e.g., Loibl et al., 2020) and the novel conceptualization of diagnostic argumentation is suggested, followed by how both may be facilitated (see aim c; see section 1.5) using simulation-based learning approaches (e.g., Chernikova, Heitzmann, Fink et al., 2020), adaptive feedback (e.g., Bimba et al., 2017), and collaborative learning (e.g., Csanadi et al., 2021). The first part concludes by describing the general research questions (see section 1.6) investigated in the subsequent empirical part.

The second part of the thesis describes the three empirical studies that addressed several questions within the outlined research agenda. The first study exemplified the comparison of diagnostic activities and diagnostic practices in medical education and teacher education (see section 2). The second study addressed the differentiation of diagnostic reasoning skills into diagnostic judgment and diagnostic argumentation (see section 3). The third study presented the results of the different and interacting effects that adaptive feedback and collaborative learning in simulations have on preservice teachers' learning of diagnostic judgment and diagnostic argumentation (see section 4).

The third and final part of the thesis (see section 5) summarizes and integrates the findings into the initially outlined theoretical assumptions and research agenda. Based on the summary and integrations, the thesis concludes by discussing the implications for research and practice.

1.2 Diagnostic Reasoning across Different Fields: Solving Diagnostic Problems

The differences in terms of *diagnostic problems* are one of the major challenges in researching diagnostic reasoning across different fields. Researchers from medical education, teacher education, and other fields, such as mechatronics, have conceptualized diagnostic reasoning as a problem-solving process (e.g., Abele, 2018; Barrows & Pickell, 1991; Csanadi et al., 2021; Heitzmann et al., 2019; Kiesewetter et al., 2013). As such, diagnostic problems can be described in terms of the characteristics of a problem-solving task: The task's content area, exemplarity, complexity, and the required activities to solve the problem can vary across problems in different fields but also problems within a field.

Diagnostic problems clearly differ from each other in terms of their *content area*. Physicians usually diagnose a patient's health and symptomatology. For example, a general practitioner may try to determine the causes of a patient's symptoms, such as fever and nausea, which may be caused by one of the various virus infections (e.g., hepatitis A virus infection). Teachers are mainly concerned with determining their pupils' performance, progress, and learning prerequisites, such as reading difficulties, which may be caused, for example, by lack of practice, visual impairment, or dyslexia (Westwood, 2008). Clearly, the content area of diagnostic problems is not comparable across different fields or even across diagnostic problems within the same field (Schwartz & Elstein, 2008; Wimmers et al., 2007).

However, different content areas are associated with another characteristic of diagnostic problems, which is still content-related yet more abstract: Diagnostic problems can be characterized regarding what may be referred to as *exemplarity* of a problem for a specific professional context. This characteristic relates to the prevalence of a diagnostic problem's

exemplars within a specific field's professional environment, which determines the likelihood that a professional will gain or has already gained experience regarding the respective problem (see Kolodner, 1992; Renkl, 2014; Thomassen & Stentoft, 2020). Accordingly, exemplarity increases the practical need of a professional to have knowledge about the respective diagnostic problem and how to solve it (see Winch, 2004). In medical education, the exemplarity of diagnostic problems, such as specific symptomatological patterns, may be derived from their prevalence estimate in the population or from expert interviews (see Charlin et al., 2000; Papa et al., 1996). However, owing to the various specializations of professionals and the specifics of their professional environments, determining the exemplarity of diagnostic problems in rather broad fields is, to some degree, difficult. For example, the diagnostic problems oncologists and cardiologists typically encounter in their professional environments will be considerably diverse. Likewise, the diagnostic problems teachers typically encounter in their professional environments will vary depending on factors such as the school track, pupils' age, and the teaching subjects. However, diagnostic problems can be ascribed to broader classes of diagnostic problems that relate to more or less exemplary tasks. For example, teachers will consider the assessment of pupils' misconceptions or level of skill belonging to the typical classes of diagnostic problems they encounter in their professional environments. Therefore, in rather broadly defined fields such as medical education and teacher education, diagnostic problems can still, to some degree, be classified according to their exemplarity.

In addition to content area and exemplarity, diagnostic problems can be characterized using other, less content-related but more structural means, and one among them is *complexity*, which has been defined as the demands imposed by the structure of a problem (Robinson, 2001). A problem's complexity is mainly determined by the amount and connectivity of information that needs to be processed (see Campbell, 1988; Stadler et al., 2019; Sweller, 2010). For example, the amount and connectivity of problem information, which is available as potential evidence, may vary (e.g., information about a patient's symptoms); further, the number and connectivity of potential problem solutions (e.g., relevant differential diagnoses) can differ across problems. In medical education, since the interconnections between problem information and relevant differential diagnoses easily accumulate to a high number and often form ambiguous patterns, diagnostic problems have been described as highly complex (Mamede et al., 2007; Papa, 2016). In teacher education, a similarly structured and complex area of diagnostic problems is the assessment of pupils' individual learning prerequisites: If pupils exhibit extensive performance or behavioral

problems, many different and interconnected factors must be considered, such as the pupils' prior academic progress and achievements, their home environment, cognitive abilities, social behavior, emotional and motivational states and traits, and so on. When researching diagnostic reasoning across different fields or even across different problems, the amount and connectivity of problem information must be considered, because it can affect, for example, the cognitive processing of information (see section 1.4.2).

Collecting and processing information requires specific activities that are necessary to solve the problem; however, the *required activities* can systematically differ across diagnostic problems and situations and, thus, across different fields. Therefore, as another structural characteristic of diagnostic problems, the required activities refer to the observable interactions with the problem (e.g., Eichmann et al., 2020). These interactions can be conceptualized in terms of epistemic activities (Fischer et al., 2014), such as problem identification, hypothesis generation, evidence generation, evidence evaluation, or drawing conclusions. Epistemic activities, since they are particularly relevant for solving diagnostic problems, have also been referred to as *diagnostic activities* (see Heitzmann et al., 2019). For example, when confronted with a patient's health problem, physicians are trained to generate hypotheses about the nature of the problem, generate and evaluate evidence accordingly, and finally draw conclusions about the diagnosis and appropriate treatment. Teachers, identifying a potential problem in one of their pupils, may as well generate hypotheses about the nature of the problem, generate and evaluate evidence, and draw conclusions about the diagnosis and appropriate interventions. Beyond identifying the potential similarities across different fields, diagnostic activities can also offer a systematic view of the differences between typical diagnostic problems. For example, a diagnostic problem may either be given already (e.g., introduced by the patient or a colleague) or may need to be identified in the first place (e.g., during classroom observations).

The similarities and differences in diagnostic problems across (and within) different fields may be specifically addressed by research. However, there are other research interests, such as studying the similarities and differences in epistemic processing or cognitive processing (see sections 1.3 and 1.4) across different fields such as medical education and teacher education. Accordingly, research can be affected by the choice of the researched diagnostic problems, since variations in the characteristics of the diagnostic problems can systematically influence the diagnostic reasoning (e.g., systematic differences in the diagnostic activities required to solve the different diagnostic problems). Cross-disciplinary research in diagnostic reasoning should, therefore, consider the variations in the

characteristics of the diagnostic problems across different fields and, if possible, try to match the characteristics of the researched diagnostic problems.

1.3 Epistemic Processing in Diagnostic Reasoning: Aims, Activities, and Practices

1.3.1 Diagnostic Accuracy: The Central Epistemic Aim in Solving Diagnostic Problems

As a problem-solving process, diagnostic reasoning is also a process of generating knowledge about a diagnostic problem. Diagnostic reasoning is thus linked to several aspects of epistemic cognition (e.g., Chinn et al., 2011; Greene et al., 2008) and scientific reasoning (e.g., Fischer et al., 2018). Epistemic cognitions are "cognitions directed at epistemic aims and their achievement" (Chinn et al., 2011, p. 147), with e*pistemic aims* being the "intended objectives of cognition and action" (Chinn et al., 2014, p. 428), such as acquiring accurate knowledge. In diagnostic reasoning, the central epistemic aim is to achieve accurate knowledge of a diagnosis, also referred to as *diagnostic accuracy* (e.g., Kolovou et al., 2021; Kramer, Förtsch, Boone et al., 2021). The high relevance of diagnostic accuracy is grounded by the fact that inaccurate diagnoses can lead to inadequate decisions, ultimately resulting in insufficient improvement or even harm. If a physician arrives at an inaccurate diagnosis, a patient may receive no or inappropriate treatment, which fails to improve or even harms the patient's condition (Norman et al., 2017). Similarly, teachers' erroneous diagnoses can critically harm their pupils' educational success and individual development, which can ultimately derail their future careers and impair their societal participation (e.g., Elhoweris, 2008). This is especially true if a diagnostic problem is associated with a critical, high-impact content area; for example, if a pupil has persisting learning difficulties caused by a developmental disorder (e.g., Reinke et al., 2011; Volpe et al., 2006). A high degree of accountability and responsibility is considered as increasing the situational epistemic value of achieving an epistemic aim (see Chinn et al., 2011), such as diagnostic accuracy, and therefore needs to be considered another factor of influence on diagnostic reasoning (see Blömeke et al., 2015; Loibl et al., 2020).

Past research on diagnostic reasoning in medical education (e.g., Norman et al., 2017) and teacher education (e.g., Urhahne & Wijnia, 2021) strongly focused on achieving diagnostic accuracy, with several studies investigating how to avoid diagnostic inaccuracy in terms of diagnostic errors and cognitive biases in diagnostic reasoning (e.g., Norman et al., 2017). Moreover, achieving diagnostic accuracy was investigated in relation to cognitive aspects such as knowledge structures (e.g., Charlin et al., 2012; see section 1.4.1) or types of cognitive processing (e.g., Croskerry, 2009; Loibl et al., 2020; see section 1.4.2) and also in

relation to the application of epistemic activities while solving diagnostic problems (e.g., Kramer, Förtsch, Boone et al., 2021).

1.3.2 Diagnostic Activities: Conceptualizing Diagnostic Problem-Solving

To achieve epistemic aims (e.g., diagnostic accuracy), individuals engage in *epistemic activities* for generating and justifying scientific knowledge (such as generating hypotheses, generating evidence, evaluating evidence, and drawing conclusions; see section 1.2; Fischer et al., 2014). The research on epistemic activities primarily focused on how individuals within different fields – teacher education (Csanadi et al., 2021), social work education (Ghanem et al., 2018), or medical education (Lenzer et al., 2017) – perform epistemic activities during reasoning or problem-solving. However, while focusing on how individuals in these fields perform epistemic activities, this research disregarded to compare the different fields with respect to their approaches in performing diagnostic activities.

From a socio-cultural perspective, epistemic activities are established by and within epistemic communities (see Kelly, 2008). Consequently, individuals within an epistemic community feature a shared understanding of when and how epistemic activities need to be performed; hence, epistemic activities must be considered at an individual as well as a collective level (see Kelly, 2008; Leont'ev, 1978; Roth & Lee, 2006). Therefore, comparing epistemic activities across different fields may reveal observable variations regarding the specific ways and preferences of when and how epistemic activities are performed within fields.

When physicians or teachers solve diagnostic problems and aim to achieve diagnostic accuracy, they engage in epistemic activities, referred to as diagnostic activities in the context of diagnostic reasoning (Heitzmann et al., 2019). Though prior research concerned with diagnostic reasoning investigated diagnostic activities both in medical education (e.g., Fink, Reitmeier et al., 2021; Lenzer et al., 2017) and teacher education (e.g., Kramer, Förtsch, Boone et al., 2021; Wildgans-Lang et al., 2020), it did not compare how the diagnostic activities were applied in the two fields. This research gap might stem from concerns regarding the comparability of diagnostic activities across fields. For instance, since the content of physicians' diagnostic problems and teachers' diagnostic problems is different (see section 1.2), the specific content of their concrete hypotheses, evidence, and conclusions also varies: a physician's hypothesis about the causes for a patient's back pain varies from a teacher's hypothesis about the causes for a pupil's writing difficulties. However, if the intended epistemic aims are the same (i.e., achieving diagnostic accuracy), the purpose of each diagnostic activity is conceptually transferable across different fields (see Hetmanek et

al., 2018): Irrespective of the specific content of a hypothesis, the activity of generating hypotheses holds the purpose of identifying potential explanations, which may require further investigation. Therefore, the variations across different fields regarding when and how diagnostic activities are performed may provide specific insights into the individual fields' epistemic approaches toward diagnostic reasoning. Thus, to gain insights into the epistemic approaches involved in diagnostic reasoning in medical education and teacher education, the first study presented in this thesis investigated the differences in medical students' and preservice teachers' diagnostic activities (see study 1 in section 2, research question 1).

1.3.3 Diagnostic Practices: Collective Patterns of Diagnostic Activities

Research can investigate not only the individual epistemic activities with respect to when and how they are performed but also the collective patterns of activities observable across the individuals of an epistemic community. These patterns can indicate a community's *epistemic practices* (see Kelly, 2008; Roth & Lee, 2006): "the specific ways members of a community propose, justify, evaluate, and legitimize knowledge claims within a disciplinary framework" (Kelly, 2008, p. 99). Epistemic practices are considered to exist at a *collective* level, but "collectives do not act: *individuals* always realize these practices in concrete ways" (Roth & Lee, 2006, p. 32). The individuals of an epistemic community might, therefore, vary in their application of epistemic activities and yet contribute to the overall pattern of their community's epistemic practices. The specifics in a community's epistemic practices relate to shared epistemic ideals (standards and criteria to assess the achievement of aims; e.g., a diagnosis is grounded on valid and convincing evidence) and a shared understanding of which processes are accepted as being reliable (e.g., how to generate valid and convincing evidence; see Duncan & Chinn, 2016). Therefore, comparing communities' epistemic practices and identifying the specifics in their epistemic practices can offer insights into their epistemic ideals, standards, and processes (see Duncan & Chinn, 2016).

Similar to transferring epistemic activities into the context of diagnostic reasoning, the idea of epistemic practices may be integrated into researching diagnostic reasoning as well. Relating the definitions of diagnostic reasoning and epistemic practices, this thesis defined *diagnostic practices* as the systematic approaches applied to collect and integrate information to reduce uncertainty and make and communicate informed and justifiable decisions in professional situations (see Heitzmann et al., 2019; Kelly, 2008). The diagnostic practices within different fields may involve specifics concerning their epistemic ideals, standards, and processes (Duncan & Chinn, 2016). Therefore, comparing diagnostic practices across medical

education and teacher education might improve both fields' conceptual understanding of diagnostic reasoning, which might facilitate future research.

The research in both medical education and teacher education has identified and conceptualized their approaches toward diagnostic reasoning. However, both fields have developed rather separate strands of research, involving different terms and theoretical notions, and this complicates the integration of the theoretical and empirical accomplishments into a cross-disciplinary research perspective on diagnostic reasoning. To investigate diagnostic practices in medical education and teacher education, the theoretical conceptualizations from the two fields can be initially interpreted and integrated by referring to their diagnostic activities (see Bauer et al., 2020; Kramer, Förtsch, Seidel et al., 2021).

In medical education, studies have found that medical students follow a diagnostic practice denoted as a *hypothesis-driven approach*: Students generate several hypotheses and evaluate the corresponding evidence to draw conclusions about the initially generated hypotheses (e.g., Coderre et al., 2010; Kiesewetter et al., 2013). The hypothesis-driven approach conforms to an epistemic ideal of differential diagnosing, which is regarded as a reliable process in medicine and thus systematically taught to students in medical education (see Duncan & Chinn, 2016; Kassirer, 2010). In addition, research found that some medical students comply with a *data-driven approach*, which involves generating and evaluating evidence while simultaneously neglecting to generate specific hypotheses and integrate evidence into conclusions (e.g., Gräsel & Mandl, 1993; Kiesewetter et al., 2013; Norman et al., 2007).

In teacher education, the research on diagnostic practices has primarily referred to the framework of professional vision (Goodwin, 1994). This framework distinguishes between two components: *noticing* and *reasoning*. The former includes identifying problems and generating hypotheses, and the latter includes three further subcomponents: describing, explaining, and predicting (e.g., Seidel & Stürmer, 2014). *Describing* corresponds to reporting the generated evidence. *Explaining* means evaluating the evidence in reference to diagnostic knowledge. Accordingly, these two subcomponents focus on evidence, whereas *predicting* the consequences of observations is indicated by generating hypotheses or drawing conclusions. Research found that expert teachers' diagnostic practices comprise describing, explaining, and predicting (Seidel & Prenzel, 2008). However, describing was found to be a prevailing aspect compared to predicting, which was found to be more diverse (Stürmer et al., 2016).

Diagnostic activities provide a point of departure to conceptually integrate the theoretical approaches from medical education and teacher education. However, the idea of

considering the collective patterns of diagnostic activities as diagnostic practices and comparing the different fields with respect to their diagnostic practices is a novel approach in diagnostic reasoning research. The observable variations regarding the diagnostic practices in medical education and teacher education may provide further insights into the fields' specific approaches toward diagnostic reasoning concerning, for example, epistemic ideals, standards, and processes (see Duncan & Chinn, 2016). Therefore, the first study presented in this thesis additionally aimed to explore how students' diagnostic practices differ between the two fields of medical education and teacher education (see study 1 in section 2, research question 2).

Overall, the presented ideas and approaches with respect to conceptualizing and researching epistemic processing in diagnostic reasoning across medical education and teacher education may critically advance a shared research perspective on diagnostic reasoning. In particular, researching diagnostic activities and diagnostic practices may result in new insights about similarities and differences in terms of field-related specifics concerning epistemic ideals, standards, and processes (see Duncan & Chinn, 2016).

1.4 Cognition in Diagnostic Reasoning: Knowledge, Processing, and Skills

1.4.1 Diagnostic Knowledge: A Content-Specific Basis for Diagnostic Reasoning

Beyond advancing research on epistemic processing in diagnostic reasoning in different fields as a collective perspective on diagnostic reasoning, it is critical to proceed and further advance research focusing on individuals' diagnostic reasoning. In this regard, especially prior research on the cognitive aspects in diagnostic reasoning must be considered to identify the potential and limitations concerning the transferability of concepts and findings across different fields.

Independent of the field, *diagnostic knowledge* is considered a crucial prerequisite for diagnostic reasoning (e.g., Blömeke et al., 2015). However, both medical education and teacher education have developed their own theoretical conceptualizations of diagnostic knowledge (e.g., Croskerry, 2009; Shulman, 1986). In accordance with the content areas of diagnostic problems, several models conceptualize diagnostic knowledge in terms of content. In medical education, the models differentiate content, for example, in the area of biomedical knowledge (e.g., pathophysiology or biochemistry) and clinical knowledge (e.g., symptomatology of specific diseases; Boshuizen, 1992), and partially also distinguish other content areas, such as psychological or sociological knowledge (Charlin et al., 2012). In teacher education, diagnostic knowledge has been most commonly conceptualized in reference to Shulman (1986, 1987), who differentiated between content knowledge (knowledge about the connections between the contents of one subject), pedagogical

knowledge (knowledge about the more general aspects of teaching, such as learning processes or classroom management), and pedagogical content knowledge (such as knowledge about instructional strategies and knowledge about typical errors made by students in a subject).

Beyond distinguishing the content areas, the theoretical models in both fields have conceptualized diagnostic knowledge in terms of the different cognitive types of knowledge, such as conceptual knowledge and strategic knowledge (e.g., Kopp et al., 2009; Mayer, 2010; Shulman, 1986). Conceptual knowledge consists of categories, such as diagnoses and relevant cues, as well as their relations with each other (e.g., Kopp et al., 2009). Strategic knowledge indicates the knowledge about how to proceed in diagnosing a specific problem (e.g., which differential diagnoses are relevant, how they can be rejected, which informational sources can deliver critical evidence; e.g., Kopp et al., 2009).

A review of diagnostic knowledge integrated the conceptualizations from medical education and teacher education into a cross-disciplinary model with two dimensions: one dimension relating to *content-related facets of knowledge*; the second relating to *cognitive types of knowledge* (Förtsch et al., 2018). The two-dimensional model of professional diagnostic knowledge (Förtsch et al., 2018) is a valuable starting point to integrate prior research from medical education and teacher education into a cross-disciplinary research perspective on diagnostic knowledge. However, the integrated model acknowledges the issue of content-specificity in diagnostic knowledge across different fields or even across different sets of problems within a field (e.g., Kolovou et al., 2021; Schwartz & Elstein, 2008; Wimmers et al., 2007). The content-specificity, in turn, also limits the comparability of abstract conceptualizations of knowledge (in terms of the cognitive types of knowledge). The content-specificity of diagnostic knowledge implies that the potential to empirically compare different fields with respect to diagnostic knowledge is limited.

1.4.2 Cognitive Processing: Fundamental Commonalities in Diagnostic Reasoning

Despite the limited comparability of content-specific diagnostic knowledge, various cognitive processes can be assumed to be theoretically and conceptually comparable across different fields. In particular, the processes of *expertise development* in terms of encapsulation and script formation (e.g., Charlin et al., 2007; Lachner et al., 2016; Schmidt & Rikers, 2007) as well as *information processing* (e.g., Croskerry, 2009; Norman, 2009) may be relevant to be considered with regard to diagnostic reasoning.

In the course of expertise development, (future) professionals extensively practice diagnostic reasoning by applying diagnostic knowledge, which is initially organized in detailed causal networks (Schmidt & Rikers, 2007). By means of repeated knowledge

application and exposure to various diagnostic problems, the diagnostic knowledge becomes increasingly encapsulated into higher-level concepts (Schmidt & Rikers, 2007). With further practice and experience, diagnostic knowledge becomes significantly associated with episodic knowledge from previously diagnosed problems. This integration of diagnostic knowledge into episodic representations of diagnostic problems is called script formation (Barrows & Feltovich, 1987; Charlin et al., 2007; Lachner et al., 2016; Putnam, 1987).

The encapsulated concepts and formed scripts enable the recognition of patterns of information as a whole, without demanding the working memory to consciously process detailed relations between information (see Evans, 2008; Norman et al., 2007). The subconscious recognition of patterns of information has been characterized as *intuitive information processing* (e.g., Kahneman, 2003, 2011; Morewedge & Kahneman, 2010). It is associated with a fast processing speed and minimized cognitive load (see Evans, 2008; Kahneman, 2011; Kalyuga, 2011), which facilitates resource-efficient acting in professional situations. In contrast, *controlled information processing* denotes evaluating isolated pieces of information and consciously analyzing their causal relations (e.g., Kahneman, 2003, 2011; Morewedge & Kahneman, 2010). The controlled type of information processing is associated with increased cognitive load and is thus functionally limited by the working memory capacity (see Evans, 2008; Kahneman, 2003; Kalyuga, 2011). Controlled information processing can be found in (future) professionals with lower levels of expertise, who have lower chances of recognizing familiar patterns of information in a diagnostic problem (see Schmidt & Rikers, 2007). However, controlled information processing can also be specifically evoked by certain characteristics of the diagnostic problem, such as increased complexity due to ambiguous information (Mamede et al., 2007), or certain characteristics of the situation, such as collaborative diagnostic reasoning (Kiesewetter et al., 2017; Radkowitsch et al., 2020), which requires explicating reasons toward a collaborating professional (or future professional).

In summary, the cognitive processes involved in diagnostic reasoning may be considered non-specific to a field (Heitzmann et al., 2019; see Kirschner et al., 2017). They may be influenced by the characteristics of the individual (e.g., the knowledge and level of expertise of the professional or future professional), the characteristics of the diagnostic problem (e.g., exemplarity or complexity as described in section 1.2), and the characteristics of the situation (e.g., collaborative diagnostic reasoning). Assuming that cognitive processing is non-specific to any field, the respective research findings may be considered transferable

across different fields, while considering the limitations associated with the characteristics of the researched individuals as well as diagnostic problems and situations.

1.4.3 Diagnostic Reasoning Skills: Distinguishing Judgment and Argumentation

The situational application of diagnostic knowledge and cognitive processing to solve a diagnostic problem requires abilities, which can be subsumed as *diagnostic reasoning skills* (e.g., Ilgen et al., 2012). Considering the field-specifics of diagnostic knowledge in relation to the non-specific nature of cognitive processing in diagnostic reasoning, diagnostic reasoning skills can be deemed to comprise shares of both field-specific and non-specific aspects (see Hetmanek et al., 2018), which research may further investigate and disentangle.

However, one key obstacle in doing so is that the literature does not clearly distinguish between different diagnostic reasoning skills. Instead, it is common to ascribe various indicators to a broad and not fully defined plural of skills or competences (e.g., Heitzmann et al., 2019; Herppich et al., 2018; Ilgen et al., 2012). One reason for this lack of conceptual clarity in diagnostic reasoning skills might be the implicit nature of the cognitive processes, which limits research to the assessment of indicators of observable processes and products (Loibl et al., 2020). The processes of diagnostic reasoning are observed, for example, regarding the performance of diagnostic activities (e.g., Wildgans-Lang et al., 2020; see section 1.3.2), while the products of diagnostic reasoning can be assessed regarding the achievement of aims, such as diagnostic accuracy (e.g., Fink, Heitzmann et al., 2021; see section 1.3.1). Another typical approach to assess diagnostic reasoning is through the use of verbalization (see Loibl et al., 2020); for example, in terms of thinking aloud or dialogue (e.g., Csanadi et al., 2021), which can be recorded during the process, or in terms of assessing the post-hoc explanations of a diagnosis as a product of diagnostic reasoning (e.g., Braun et al., 2018; Rapanta & Felton, 2021).

However, ascribing non-verbal and verbal indicators to the same diagnostic reasoning skills may be considered problematic in terms of the assumed distinction of the underlying cognitive processing types (see section 1.4.2): While controlled information processing is considered conscious and thus explicable, intuitive information processing is considered unconscious (e.g., Evans, 2008). Accordingly, if diagnostic accuracy is achieved through intuitive information processing, it is not necessarily a given that the resulting diagnosis will be well explained in terms of verbalizing the previously processed information.

A second reason for non-verbal and verbal indicators not being ascribed to the same diagnostic reasoning skills concerns epistemic aims (see Chinn et al., 2011; see also section 1.3.1), which go beyond achieving diagnostic accuracy in situations requiring verbalization.

For example, in situations of collaborative diagnostic reasoning (Kiesewetter et al., 2017; Radkowitsch et al., 2020), providing verbalized reasons is associated with aiming to achieve a common *understanding* with others (Chinn et al., 2011; Mercier & Heintz, 2014) by means of verbal sense-making, articulation, and persuasion (Berland & Reiser, 2009; Rapanta & Felton, 2021). Additionally, there are nonimmediate dialogical situations (see Walton, 1990), such as documenting diagnostic reasoning, which, nonetheless, aim to explain one's own understanding and evoke understanding in others at a later point in time.

Therefore, this thesis differentiated diagnostic reasoning skills into the well-established concept of diagnostic judgment (e.g., Loibl et al., 2020) and the novel concept of diagnostic argumentation. Within this distinction, *diagnostic judgment* refers to interpreting information about a diagnostic problem and integrating it into a diagnostic conclusion, pursuing the aim to achieve diagnostic accuracy (see Heitzmann et al., 2019; Loibl et al., 2020; Victor-Chmil, 2013). On the other hand, *diagnostic argumentation* refers to explaining the interpretations about a diagnostic problem as well as the resulting diagnostic conclusions comprehensibly and persuasively (see Berland & Reiser, 2009; Walton, 1990).

Drawing on the reasons related to cognitive processing and epistemic aims that are considered non-specific to any field, distinguishing between diagnostic judgment and diagnostic argumentation as different diagnostic reasoning skills may be relevant to both medical education and teacher education. However, the assumed distinctiveness of diagnostic judgment and diagnostic argumentation demanded empirical investigation, which was addressed in the second study presented in this thesis (see study 2 in section 3, research question 3).

1.4.4 Conceptualizing Diagnostic Argumentation: Justification, Disconfirmation, and Transparency

Investigating diagnostic argumentation as a distinct diagnostic reasoning skill initially requires a detailed and field-unspecific approach to conceptualizing diagnostic argumentation. Considering the epistemic aim of achieving a common understanding (e.g., by means of verbal sense-making, articulation, and persuasion; see section 1.4.3), diagnostic argumentation should be conceptualized with respect to facets that contribute to facilitating a potential recipient's understanding and evaluation of diagnostic reasoning. In this regard, it needs to be acknowledged that both the individual, who is engaging in diagnostic reasoning, and the potential recipients are embedded in an epistemic community, one that might have its own specific practices and standards of diagnostic reasoning and argumentation (see section 1.3.3). The different epistemic communities engaging in diagnostic reasoning are, however,

embedded in a broader context of science, which is characterized by most widely generalizable scientific practices and standards (see Jiménez-Aleixandre & Crujeiras, 2017; Mercier & Heintz, 2014; Osborne, 2014). Therefore, field-specific standards in diagnostic argumentation should comply with (or otherwise be reconsidered in reference to) the fundamental norms and standards that have developed in the broader context of *scientific argumentation* (e.g., Bricker & Bell, 2008; Mercier & Heintz, 2014; Sampson & Clark, 2008).

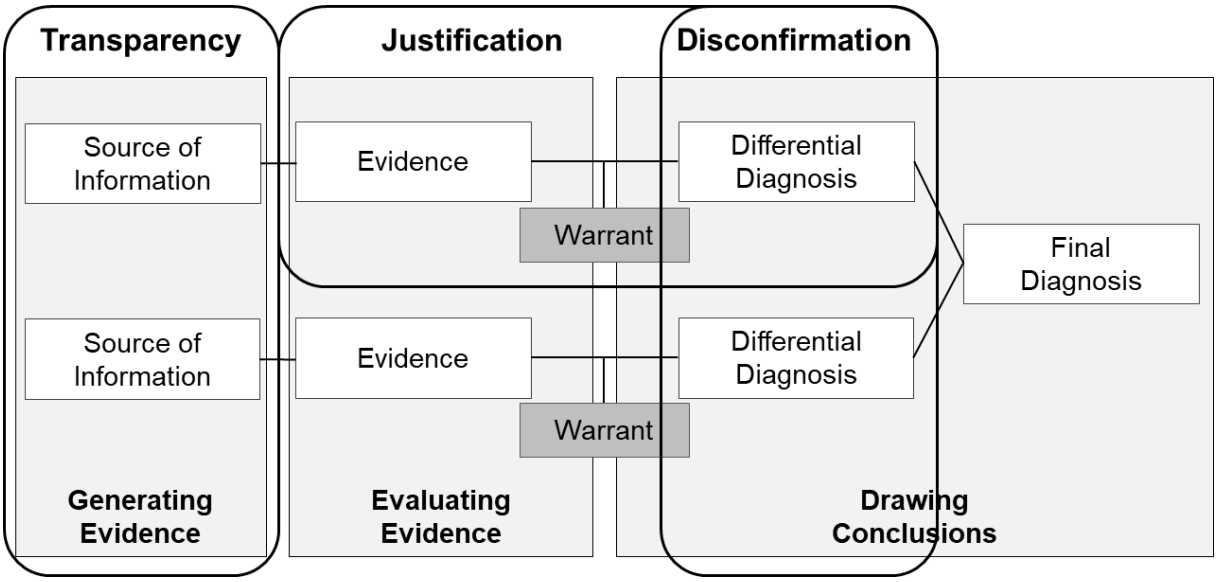
Scientific argumentation has been evaluated in reference to various conceptualizations and frameworks (e.g., Kelly & Takao, 2002; Lawson, 2003; Sandoval & Millwood, 2005; Schwarz et al., 2003; van Eemeren & Grootendorst, 2004; Zohar & Nemet, 2002), among which especially the Toulmin argumentation pattern (Toulmin, 1958) gained wide acceptance and foundational influence. The most common theme suggested by this framework is that claims need to be justified by evidence to be comprehensible to any recipient (Hitchcock, 2005; Toulmin, 1958). Moreover, justifying a claim by providing evidence allows the recipients to discuss the presented evidence and raise potential issues about the line of reasoning. However, there are two further salient themes emerging from the literature of scientific reasoning and argumentation: One of these themes emphasizes the relevance of considering and disconfirming alternative explanations (e.g., Lawson, 2003). Similar to the scientific approach of disconfirming alternative hypotheses (e.g., Gorman et al., 1984), the argumentative approach of disconfirming alternative explanations offers additional support for the explanation aimed to be presented as accurate or preferable in any way (see Lawson, 2003; Toulmin, 1958). The second theme emphasizes the role of methodological transparency regarding the approaches and informational sources used to generate evidence (e.g., Fischer et al., 2014). Explicating the methods and informational sources facilitates their critical evaluation and, thus, the evaluation of the quality and persuasiveness of the presented evidence and conclusions (see Bromme et al., 2018; Chinn & Rinehart, 2016).

In reference to the three identified themes in scientific argumentation, diagnostic argumentation can be conceptualized and evaluated with respect to the three facets of justification, disconfirmation, and transparency. Because these three facets resemble some standards and practices involved in scientific argumentation, they should be well-suited to contribute to a recipients' understanding and persuasion in the context of diagnostic argumentation. *Justification* in diagnostic argumentation refers to providing evidence for a diagnosis (see Figure 1). Diagnoses resemble claims that need to be justified by making a warranted connection to evidence generated about the diagnostic problem (see Hitchcock, 2005; Toulmin, 1958). Therefore, justifications in diagnostic argumentation present and

evaluate evidence as a basis from which to draw conclusions about a diagnosis (see Fischer et al., 2014; Heitzmann et al., 2019). *Disconfirmation* in diagnostic argumentation emphasizes the role of discussing differential diagnoses. Attempting to reduce uncertainty, professionals who engage in diagnostic reasoning might generate and evaluate hypotheses; that is, differential diagnoses (Fischer et al., 2014; Heitzmann et al., 2019; Klahr & Dunbar, 1988). Considering the full picture of available evidence, there might be clear differences in the likelihood of an accurate diagnosis and the relevant differential diagnoses. However, the differential diagnoses may still be considered alternative explanations in solving the given diagnostic problem. Therefore, to demonstrate that alternative explanations have been considered, the differential diagnoses should be explicated and discussed in diagnostic argumentation (see Figure 1). Explicitly considering and disconfirming the differential diagnoses as competing explanations facilitates the comprehensibility and persuasiveness of the final diagnosis. In addition, the recipients can build on this information to evaluate and criticize whether the relevant differential diagnoses have been missed or mistakenly rejected. *Transparency* in diagnostic argumentation refers to describing the processes of evidence generation (see Figure 1). Explicating how the evidence was generated offers information about the reliability of the methodology and informational sources (Chinn et al., 2014; Fischer et al., 2014). Therefore, transparency in diagnostic argumentation facilitates a recipient's understanding and evaluation of the quality of the evidence and, ultimately, the validity of the diagnostic conclusions (see Bromme et al., 2018; Chinn & Rinehart, 2016; Vazire, 2017).

Figure 1

Underlying Framework of the Proposed Conception of Diagnostic Argumentation



There are, however, several research questions that emerge with respect to the presented conceptualization of diagnostic argumentation, all of which require further investigation. First, considering the assumption that diagnostic practices of argumentation may vary across different fields, research may explore how students from different fields (such as teacher education and medical education) make use of justification, disconfirmation, and transparency in their diagnostic argumentation (see study 2 in section 3, research question 1 for an exploration of justification, disconfirmation, and transparency in teacher education).

Second, it is unclear whether justification, disconfirmation, and transparency represent distinct subskills or one joint underlying skill of diagnostic argumentation. One part of this question concerns to what extent justification, disconfirmation, and transparency are based on the same or different knowledge. As a diagnostic reasoning skill, diagnostic argumentation involves the situational application of diagnostic knowledge. In particular, it involves explicating diagnostic reasoning and problem-solving, for which conceptual and strategic diagnostic knowledge are needed (e.g., for generating evidence from informational sources and for making a warranted connection between the evidence and a diagnosis or several differential diagnoses; Heitzmann et al., 2019). However, diagnostic argumentation additionally aims to provide a comprehensible and persuasive presentation of diagnostic reasoning, for which further knowledge and skills beyond conceptual and strategic diagnostic knowledge may be necessary (see Hetmanek et al., 2018). Thus, research may explore the interrelations between justification, disconfirmation, and transparency in diagnostic argumentation and determine to what extent the three facets are explained by conceptual and strategic diagnostic knowledge (see study 2 in section 3, research questions 1 and 2 for an exploration of justification, disconfirmation, and transparency in teacher education).

Third, if the findings support the proposition of differentiating between judgment and argumentation as two distinct diagnostic reasoning skills, there is a need to further investigate which educational interventions are specifically suitable to support the learning of diagnostic argumentation. Moreover, to determine the practical impact of differentiating diagnostic reasoning skills for teaching, learning, and assessment purposes, it has to be investigated whether learning interventions have differing effects on diagnostic argumentation and diagnostic judgments. The matter of facilitating diagnostic reasoning skills in higher education is addressed in the following section.

1.5 Facilitating Diagnostic Reasoning Skills

1.5.1 Using Simulation-Based Learning in Higher Education

Research that aims to understand the variables and processes involved in diagnostic reasoning ultimately aims to identify the potential for facilitating the learning and, thus, the performance of diagnostic reasoning skills. To prepare future professionals to engage in diagnostic reasoning in real professional settings (see Grossman et al., 2009), higher education must foster knowledge encapsulation, script formation (see section 1.4.2), and the situational application of diagnostic reasoning skills (see section 1.4.3). Research in medical education and teacher education has addressed the question of how to learn diagnostic reasoning skills in higher education: The findings suggest that it is highly beneficial to confront future professionals with authentic diagnostic problems in the course of higher education (Chernikova, Heitzmann, Fink et al., 2020). However, students' access to real-life practice and, consequently, the number and diversity of real-life encounters with diagnostic problems are limited (Grossman et al., 2009; Heitzmann et al., 2019). Involving inexperienced students in real-life practice is also associated with increased resource requirements (e.g., educators' time investment in supporting students) and potential risks (e.g., concerning patients' health; Ziv et al., 2003). In both medical education and teacher education, *simulation-based learning* was suggested as a promising approach for approximating practice in higher education to familiarize students with everyday professional situations (Bradley, 2006; Grossman et al., 2009; Kaufman & Ireland, 2016).

Simulations are simplified but valid representations of professional situations that have a set of features, which can be manipulated by learners (Heitzmann et al., 2019; Sauvé et al., 2007). In higher education, digital simulations such as virtual patients (Cook et al., 2010) have become increasingly popular (Gegenfurtner et al., 2014) because of their resource-effectiveness and accessibility to large groups of learners. Designing a simulation allows the presentation of specifically selected diagnostic problems and the functional adaptation of their characteristics, such as reducing a problem's complexity (see section 1.2). The option to adapt diagnostic problems makes simulations a valuable tool not only for educational purposes but also for researching diagnostic reasoning across and within different fields (Fink, Radkowitsch et al., 2021).

Simulations are particularly beneficial not only for practicing everyday professional situations but also for specifically focusing on infrequent or high-risk situations and repeatedly practicing specific subsets of skills (De Coninck et al., 2019; Grossman et al., 2009; Kaufman & Ireland, 2016; Ziv et al., 2003). Simulation-based learning is thus

considered well-suitable to foster diagnostic reasoning skills (Heitzmann et al., 2019) and may be also regarded as well-suited to specifically practice diagnostic judgment and diagnostic argumentation (see section 1.4.3). As meta-analytical findings indicate, simulation-based learning exerts a large positive effect on diagnostic reasoning when compared to no intervention (Chernikova, Heitzmann, Stadler et al., 2020; Cook et al., 2010; Cook et al., 2011) and small to medium positive effects when compared to other types of instruction (Cook et al., 2010; Cook et al., 2012). However, despite presenting simplified representations of diagnostic problems and situations, simulation-based learning remains challenging for learners. In particular, novice learners require support and feedback to cope with the complexity of the simulated problems and effectively learn diagnostic reasoning skills (Chernikova, Heitzmann, Fink et al., 2020; Cook et al., 2010; Cook et al., 2013; Wisniewski et al., 2020).

Research has not investigated whether facilitating diagnostic judgment and diagnostic argumentation in simulation-based learning requires different support and feedback to be effective. For example, supporting learners to provide high-quality justifications in diagnostic argumentation might require particularly elaborated feedback. Moreover, if research identifies the differing effects of feedback or other support measures on the learning of diagnostic judgment and diagnostic argumentation, these findings would emphasize the practical relevance of differentiating diagnostic reasoning skills for teaching and learning purposes. Whether simulation-based learning of diagnostic judgment and diagnostic argumentation requires different feedback will be explored in the third study presented in this thesis (see study 3 in section 4).

1.5.2 Automatic Adaptive Feedback in Simulation-Based Learning

Receiving feedback is considered crucial for maximizing the benefits of simulations in learning complex skills such as diagnostic reasoning (Cook et al., 2010; Cook et al., 2013; Scheuer et al., 2012). Feedback provides information about the aspects of a learner's performance or understanding, such as a correct task solution, corrective information about the gap between the learner's performance and the correct task solution, clarifying and complementary information about a task, information about alternative strategies, or encouragement (Hattie & Timperley, 2007). However, different types of feedback are associated with varying benefits for different types of learning objectives. For example, task-related feedback focuses solely on the information about correct task solutions and is considered effective for facilitating simple and familiar tasks (i.e., they can be performed using mainly intuitive information processing; see section 1.4.2; Hattie & Timperley, 2007;

Narciss et al., 2014). However, to facilitate the performance of tasks that demand the identification of causal relations through controlled information processing (see section 1.4.2), elaborated feedback is needed, which offers information on how to appropriately process a task (Cook et al., 2013; Hattie & Timperley, 2007; Narciss et al., 2014; Scheuer et al., 2012; Wisniewski et al., 2020). Considering the learning of diagnostic judgment or diagnostic argumentation, both diagnostic reasoning skills, since they can involve controlled information processing, should benefit from elaborated feedback. However, diagnostic argumentation is assumed to involve a higher necessity for controlled information processing, whereas diagnostic judgment can partly rely on intuitive information processing (see sections 1.4.2 and 1.4.3). Therefore, the effects of elaborated feedback on diagnostic judgment and diagnostic argumentation may vary.

Elaborated feedback on the appropriate or optimal processing of a task can be implemented using different means. Often, elaborated feedback is implemented as *static feedback* (i.e., non-adaptive feedback), for example, by presenting an expert solution, which exemplifies the processing of the task and can be made available to all learners after they submit their own attempts at solving the problem (e.g., Renkl, 2014). In doing so, an expert solution can serve as a generic and thus a resource-efficient form of feedback in higher education. Moreover, providing expert solutions as static feedback can be easily automated in digital learning environments (e.g., in digital simulations). However, a disadvantage of any form of static feedback is that learners need to compare their own task processing and solution with the expert solution and identify areas for improvement by themselves. Since this comparison confronts learners with large amounts of information, it can exceed learners' working memory capacity and restrain their learning (Sweller et al., 2019).

In contrast to static feedback, *adaptive feedback* adjusts to a learner's current task processing and solution (see Plass & Pawar, 2020) by, for example, highlighting the gaps between the current and desired performance or by providing additional explanation if significant mistakes are detected (Bimba et al., 2017; Narciss et al., 2014; Plass & Pawar, 2020). Adaptive feedback can thus facilitate learners' understanding of their current state of knowledge and options for improvement and thus free the working memory capacity for the actual learning processes (see Moreno, 2004; Narciss, 2008; Sweller et al., 2019). Therefore, elaborated adaptive feedback may be particularly beneficial for learning skills, such as diagnostic argumentation, that require identifying causal relations through controlled information processing and are associated with high demands on working memory.

However, providing adaptive feedback manually by analyzing learners' task solutions and writing detailed and elaborated feedback is an extremely resource-intensive task for higher education teachers. Automating the analysis of learners' task processing in digital learning environments (e.g., in digital simulations) to provide *automatic adaptive feedback* to numerous learners seems to be a potential solution. Prior research primarily explored the use of closed format questions or log data, which can be assessed automatically by cognitive tutors and intelligent tutoring systems (Graesser et al., 2018). However, providing automatic adaptive feedback on diagnostic argumentation requires automating the analysis of verbal data and, to this end, Natural Language Processing (NLP) can be employed, which uses methods of artificial intelligence and machine learning to parse, analyze, and understand human language (Manning & Schuetze, 2005). The recent advancements in the methods of artificial intelligence and machine learning, namely artificial neural networks, offer new technical capabilities that enable the analysis of complex verbal data (Li, 2018), such as diagnostic argumentation, to be automated. NLP methods may thus be considered useful in automating the detailed, real-time analyses of learners' diagnostic argumentation to offer automatic adaptive feedback without involving a human corrector (see Plass & Pawar, 2020). Some initial evidence suggests, that using NLP to automate adaptive feedback on written explanations can facilitate learners' revision of their explanations, which was found to enhance the quality of the learners' justifications (Zhu et al., 2017; Zhu et al., 2020).

NLP-based automatic adaptive feedback may also increase the benefits of students' simulation-based learning of diagnostic reasoning skills, especially regarding the quality of justification in diagnostic argumentation. Therefore, the third study presented in this thesis compared the effects of NLP-based automatic adaptive feedback and static feedback (i.e., an expert solution) on the accuracy of diagnostic judgments and the quality of justifications in diagnostic argumentation in the context of simulation-based learning (see study 3 in section 4, research questions 1 and 3).

1.5.3 Social Form of Learning in Simulation-Based Learning

Simulation-based learning can be implemented in different learning settings. In particular, the *social form of learning* can be varied: learners can solve diagnostic problems in simulation-based learning *individually* or *collaboratively* (see Chi, 2009). Collaborative learning refers to two or more individuals' coordinated and synchronous engagement in learning activities, thereby exerting mutual influence on each other's learning (see O'Donnell & Hmelo-Silver, 2013; Roschelle & Teasley, 1995). Collaborative learning often involves collaborative problem-solving, which is characterized by the shared goal of collectively

finding a problem solution (see Dillenbourg et al., 2009; Hmelo-Silver & DeSimone, 2013; Roschelle & Teasley, 1995). Research on collaborative learning found that using simulations as a learning environment is highly effective when compared to other technologically supported learning environments (Jeong et al., 2019).

Collaborative simulation-based learning of diagnostic reasoning skills involves collaborative learning and collaborative problem-solving by letting learners solve diagnostic problems together, thus making them engage in collaborative diagnostic reasoning (see Kiesewetter et al., 2017; Radkowitsch et al., 2020). Collaborative diagnostic reasoning requires learners to explicate and exchange reasons while they are engaged in solving a diagnostic problem. This in-process diagnostic argumentation might facilitate post-hoc diagnostic argumentation, that is, the comprehensible and persuasive explanation of the diagnostic conclusions (see section 1.4.3). Evidence concerning collaborative problem-solving suggests that collaborative learners reflect more on hypotheses and evidence when solving diagnostic problems, whereas individual problem-solvers were found to be rather solution-oriented (Csanadi et al., 2021; Okada, 1997). Learning partners can benefit from being challenged by each other's questions, which motivates reflection about unexplored perspectives (Asterhan & Schwarz, 2009; Roscoe & Chi, 2008). In addition, research has found that learners can critically evaluate others' arguments better than their own arguments (Mercier & Sperber, 2017). Critically evaluating and reconciling different arguments within collaborative diagnostic reasoning might facilitate not only diagnostic argumentation but also accurate diagnostic judgments. Collaborative simulation-based learning thus seems to offer further potential for facilitating the learning of diagnostic reasoning skills (Scheuer et al., 2012).

Another assumed benefit of collaborative learning and problem solving is that learners receive additional feedback from their learning partners (see Weinberger et al., 2010). Collaborating learners evaluate each other's arguments, adaptively correct each other, and fill each other's knowledge gaps. Therefore, the need for adaptive feedback in simulation-based learning of diagnostic reasoning skills might interact with the social form of learning, in that the collaborative learners' need for adaptive feedback in diagnostic reasoning might be lower compared to that of individual learners.

However, two meta-analyses suggested that research findings concerning collaborative simulation-based learning are mixed (Cook et al., 2012; Cook et al., 2013). Moreover, there is evidence that collaborative learners also significantly benefit from receiving adaptive feedback (Chuang & O'Neil, 2013; Hsieh & O'Neil, 2002). Besides the potential benefits for

learning, collaboration involves transaction costs associated with interacting, communicating, and coordinating (Kirschner et al., 2009). The collaborative activities can demand additional working memory capacity, characterized as collaboration load (Janssen & Kirschner, 2020), so that less working memory capacity is available for the actual task performance and learning processes (see Sweller et al., 2019). The collaboration load might restrain the learning of complex skills that specifically demand controlled information processing and are associated with high demands on the working memory, such as diagnostic argumentation (see section 1.4.3).

The need for feedback and thus the effects of the different kinds of feedback might differ depending on whether learners learn individually or collaboratively. However, considering the contrasting findings and theoretical assumptions regarding the benefits and costs of collaboration, the direction of a potential interaction effect between the social form of learning and the need for feedback in the simulation-based learning of diagnostic reasoning skills is unclear. Moreover, whether the effects differ with respect to facilitating diagnostic judgment or diagnostic argumentation is an open question. These questions are addressed in the third study presented in this thesis (see study 3 in section 4, research questions 2 and 4).

1.6 General Research Questions, Methodological Considerations, and Outline of the Studies

This thesis describes research that aimed to contribute to (a) developing a more cross-disciplinary research perspective on diagnostic reasoning, (b) integrating and refining the existing understanding of the relevant skills, and (c) presenting approaches to facilitate their learning. In explaining the theoretical foundations of this thesis, several task-related, epistemically grounded, and cognitively related reasons have been introduced, which explain why and in which regard different fields – medical education and teacher education, in particular – may be comparable or limited in their comparability with respect to diagnostic reasoning. One initial constraint in researching diagnostic reasoning across different fields is the differences in diagnostic problems, which can be evaluated in terms of several characteristics, such as content area, exemplarity, complexity, and required activities (see section 1.2). Diagnostic reasoning research should consider variations in the characteristics of diagnostic problems across different fields and, if possible, try to match the characteristics of the diagnostic problems when comparing, for example, the epistemic processes across different fields. Comparing the epistemic processes can be done by referring to epistemic aims (Chinn et al., 2011; see section 1.3.1), diagnostic activities (Fischer et al., 2014; see section 1.3.2), and diagnostic practices (Bauer et al., 2020; see section 1.3.3), which may

provide particular insights into the diagnostic reasoning of different fields. Therefore, the first study compared diagnostic activities and diagnostic practices across medical education and teacher education.

Moreover, the differences and similarities in the cognitive aspects of diagnostic reasoning have been elaborated. On the one hand, diagnostic knowledge (Förtsch et al., 2018) was considered as a field-specific basis for diagnostic reasoning (see section 1.4.1). On the other hand, cognitive processing was considered non-specific to any field (see section 1.4.2), which implies that respective research findings may be considered transferable across different fields (while considering the characteristics of the researched individuals as well as diagnostic problems and situations). A rationale was provided as to why and how diagnostic reasoning skills may be differentiated into diagnostic judgment (which aims to achieve diagnostic accuracy; e.g., Loibl et al., 2020) and the newly defined conceptualization of diagnostic argumentation (see section 1.4.3), additionally suggesting that this differentiation may be relevant to both fields of medical education and teacher education. To further elaborate on the idea of diagnostic argumentation, a conceptualization of diagnostic argumentation was introduced, including the three facets of justification, disconfirmation, and transparency (see section 1.4.4). The three facets were suggested to resemble some standards and practices involved in scientific argumentation and thus should be relevant to diagnostic argumentation in different fields, such as teacher education and medical education. Using teacher education as the context for initial investigation, the idea of distinguishing diagnostic judgment and diagnostic argumentation as two different diagnostic reasoning skills was empirically investigated in a second study, which also explored the three facets in diagnostic argumentation.

How diagnostic judgment and diagnostic argumentation may be facilitated using approaches of simulation-based learning (e.g., Chernikova, Heitzmann, Fink et al., 2020), adaptive feedback (e.g., Bimba et al., 2017) and collaborative learning (e.g., Csanadi et al., 2021) was discussed as well. In that regard, it is particularly interesting whether the proposed learning interventions exert differing effects on diagnostic argumentation and diagnostic judgments, which would support the assumption that differentiating diagnostic reasoning skills has a practical impact on the teaching, learning, and assessment of diagnostic reasoning. Therefore, a third study investigated the effects of the proposed approaches on the diagnostic accuracy of diagnostic judgments and the quality of justification in diagnostic argumentations.

The following sections offer a brief overview of the three studies regarding the investigated research questions as well as the considerations concerning the particular methodological approaches.

1.6.1 Outline of Study 1

To develop a more cross-disciplinary research perspective on diagnostic reasoning, research on epistemic processing in diagnostic reasoning across different fields was proposed (see section 1.3). In particular, researching diagnostic activities (see section 1.3.2) and diagnostic practices (see section 1.3.3) may result in not only new insights about similarities but also differences in terms of field-related specifics in epistemic ideals, standards, and processes (see Chinn et al., 2014). Therefore, the first study aimed to compare diagnostic activities and diagnostic practices in medical education and teacher education. The research questions were as follows:

RQ1: To what extent do learners' *diagnostic activities* differ between medical education and teacher education?

RQ2: To what extent do learners' *diagnostic practices* differ between medical education and teacher education?

To explore diagnostic practices in terms of the collective patterns of diagnostic activities, the novel method of Epistemic Network Analysis (ENA) was used (ENA; Shaffer, 2017). ENA is specifically designed for exploring and comparing individual and collective patterns of epistemic processing. Its algorithm analyzes the co-occurrences of specific instances (such as diagnostic activities) in a pre-defined temporal context (such as a moving window of two sentences within a diagnostic argumentation; see Siebert-Evenstone et al., 2017) and creates networks based on the relative frequencies of the observed co-occurrences within the data. In doing so, the method provides an opportunity to explore the collective patterns of diagnostic activities as diagnostic practices and compare them across medical education and teacher education.

However, research across different fields can be affected by differences in the characteristics of diagnostic problems (see section 1.2), which can systematically influence diagnostic reasoning (e.g., systematic differences in the diagnostic activities required to solve the diagnostic problems). Diagnostic problems can be compared in terms of several characteristics – content area, exemplarity, complexity, and required activities – that can affect diagnostic reasoning and should be considered when researching diagnostic reasoning across different diagnostic problems. Therefore, the research in this thesis used digital simulations of diagnostic problems for researching diagnostic reasoning across medical

education and teacher education (see Appendix A for more information about the simulation-based learning environment). Using simulations for researching diagnostic reasoning across different fields allows the presentation of specifically selected diagnostic problems and the functional adaptation of their characteristics (Fink, Radkowitsch et al., 2021; see section 1.2). For the current research, two sets of simulated diagnostic problems were designed: In medical education, the students were confronted with diagnostic problems concerning patients having fever or back pain (e.g., because of a hepatitis A virus infection or an ankylosing spondylitis). In teacher education, the students were confronted with diagnostic problems concerning problems of pupils having deficits in reading and writing or behavioral problems, which had to be distinguished as clinically relevant (e.g., because of dyslexia or ADHD) or not clinically relevant (e.g., drop in school performance because of a visual impairment or inattentiveness because of emotional stress induced by family conditions; see Appendix A for more information about the two sets of diagnostic problems). These sets of diagnostic problems were selected and designed to achieve a comparatively high degree of matching in terms of complexity and required activities. For example, both the learning environments presented diagnostic problems with a matched structure: An initial problem statement concerning a virtual patient or pupil was presented so that the diagnostic activity of *identifying problems* was not required. Next, the students had to *generate evidence* by accessing several informational sources, such as the results of different examinations and tests in medical education, and in teacher education the reports of observations from inside and outside of the classroom, and the samples of the pupils' written exercises and school certificates. The students had to *generate hypotheses* and *evaluate evidence* to *draw conclusions* and solve the diagnostic problem. Moreover, both sets of diagnostic problems were considered to involve high degrees of accountability and responsibility, thus inducing a high situational epistemic value of achieving the epistemic aim of diagnostic accuracy (see section 1.3.1). The matched design of the simulated diagnostic problems might be considered as increasing the degree of functional comparability of the observed diagnostic activities and diagnostic practices across the two fields (see section 1.3.2).

1.6.2 Outline of Study 2

To integrate and refine the existing understanding of diagnostic reasoning skills, the thesis proposed to distinguish diagnostic judgment and diagnostic argumentation as two different diagnostic reasoning skills (see section 1.4.3). However, introducing the conceptualization of diagnostic argumentation, including the three facets of justification, disconfirmation, and transparency posed several questions (see section 1.4.4). In particular,

the idea of distinguishing diagnostic reasoning skills with respect to diagnostic judgment and diagnostic argumentation required empirical investigation. Moreover, the three facets required exploration concerning whether the three facets represent distinct subskills or one joint underlying skill of diagnostic argumentation. In this regard, the degree to which justification, disconfirmation, and transparency are based on the same or different knowledge is particularly relevant. Because conceptual and strategic diagnostic knowledge are thought to be a major basis for the reasoning presented in diagnostic argumentation, they might also be important for explaining the performance differences regarding justification, disconfirmation, and transparency in diagnostic argumentation. However, apart from explicating prior reasoning, diagnostic argumentation additionally aimed to offer a comprehensible and persuasive presentation of the identified reasons for which further knowledge and skills beyond conceptual and strategic diagnostic knowledge may be relevant (see Hetmanek et al., 2018). Therefore, another matter of investigation was the extent to which conceptual diagnostic knowledge and strategic diagnostic knowledge each contribute to explaining variance in justification, disconfirmation, and transparency in diagnostic argumentation.

These open questions were empirically investigated in the second study. For this purpose, teacher education was selected as the context for the initial investigation, yet suggesting that the introduced conceptualization and research questions are relevant to both fields of medical education and teacher education. The text data collected for the first study in teacher education was reanalyzed (see Appendix B for information about the coding of the text data); further, preservice teachers' prior conceptual and strategic diagnostic knowledge as well as the diagnostic accuracy of their diagnostic judgments were also analyzed. The second study investigated the following research questions:

RQ1: Do justification, disconfirmation, and transparency represent distinct subskills or are they indicators of one joint underlying diagnostic skill?

RQ2: To which extent are justification, disconfirmation, and transparency based on conceptual diagnostic knowledge and strategic diagnostic knowledge?

RQ3: Do diagnostic judgment and diagnostic argumentation represent different diagnostic reasoning skills?

1.6.3 Outline of Study 3

To present and investigate approaches to facilitate the learning of diagnostic reasoning skills, the third study addressed the effects of adaptive feedback (e.g., Bimba et al., 2017) and collaborative learning (e.g., Csanadi et al., 2021) in the simulation-based learning of diagnostic reasoning skills (e.g., Chernikova, Heitzmann, Stadler et al., 2020). It compared

the effects of automatic adaptive feedback vs. static feedback (i.e., expert solutions) in simulation-based learning on the accuracy of diagnostic judgment and the quality of justification in diagnostic argumentation. It also addressed the potential interaction effects of the type of feedback with the social form of learning, in terms of individual vs. collaborative learning. Automatic adaptive feedback and collaborative learning were assumed to facilitate accurate diagnostic judgment and high-quality justification in diagnostic argumentation. However, there may also be variations in the result patterns of the two diagnostic reasoning skills; for example, higher benefits for diagnostic argumentation compared to diagnostic judgment (see section 1.5.2). The study addressed the following research questions:

RQ1: Is *automatic adaptive feedback* more effective than *static feedback* in fostering
(RQ1a) learners' *diagnostic accuracy*?
(RQ1b) learners' *quality of justification*?
Automatic adaptive feedback was hypothesized to be more effective than static feedback in fostering learners' diagnostic reasoning skills (see Zhu et al., 2017; Zhu et al., 2020; Hypothesis 1a for diagnostic accuracy; 1b for the quality of justification).

RQ2: Is there an interaction of the *social form of learning* and the *type of feedback* on
(RQ2a) learners' *diagnostic accuracy*?
(RQ2b) learners' *quality of justification*?
The social form of learning and the type of feedback were hypothesized to interact concerning the effects on diagnostic reasoning skills (Hypothesis 2a for diagnostic accuracy; 2b for the quality of justification).

As part of this study, an NLP-based algorithm was developed and used to implement automatic adaptive feedback in a simulation-based learning environment. Implementing such NLP-based systems initially requires training data. Manually coded text data of 118 preservice teachers who had participated in the first study (Bauer et al., 2020; see Appendix B for information about the coding schemes), was used to train the NLP algorithm. The algorithm learned from the training data to automatically analyze the preservice teachers' written task solutions and identify the included diagnostic entities (evidences and diagnoses) and diagnostic activities (hypothesis generation, evidence generation, evidence evaluation, and drawing conclusions; see Heitzmann et al., 2019). Further details on training the NLP algorithm to detect diagnostic activities were described by Schulz et al. (2019). Moreover, a technical description of the full feedback system was provided by Pfeiffer et al. (2019). After the algorithm was implemented in the simulation-based learning environment, the students could submit their written task solution and receive in-time automatic adaptive feedback

comprising predefined feedback elements, which the system selected based on the identified diagnostic entities and diagnostic activities (see study 3 in section 4). Further information on the feedback is available in Appendix D.

2

Study 1: Diagnostic Activities and Diagnostic Practices in Medical Education and Teacher Education: An Interdisciplinary Comparison

Reference: Bauer, E., Fischer, F., Kiesewetter, J., Shaffer, D. W., Fischer, M. R., Zottmann, J. M., & Sailer, M. (2020). Diagnostic activities and diagnostic practices in medical education and teacher education: An interdisciplinary comparison. *Frontiers in Psychology, 11*, Article 562665. https://doi.org/10.3389/fpsyg.2020.562665

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 20 October 2020
doi: 10.3389/fpsyg.2020.562665

Diagnostic Activities and Diagnostic Practices in Medical Education and Teacher Education: An Interdisciplinary Comparison

Elisabeth Bauer[1], Frank Fischer[1], Jan Kiesewetter[2], David Williamson Shaffer[3], Martin R. Fischer[2], Jan M. Zottmann[2] and Michael Sailer[1]*

[1] Education and Educational Psychology, Department Psychology, LMU University of Munich, Munich, Germany, [2] Institute for Medical Education, University Hospital, LMU University of Munich, Munich, Germany, [3] Epistemic Analytics Lab, Department of Educational Psychology, University of Wisconsin Madison, Madison, WI, United States

In this article, we investigate diagnostic activities and diagnostic practices in medical education and teacher education. Previous studies have tended to focus on comparing knowledge between disciplines, but such an approach is complicated due to the content specificity of knowledge. We compared 142 learners from medical education and 122 learners from teacher education who were asked to (a) diagnose eight simulated cases from their respective discipline in a simulation-based learning environment and (b) write a justificatory report for each simulated case. We coded all justificatory reports regarding four diagnostic activities: *generating hypotheses*, *generating evidence*, *evaluating evidence*, and *drawing conclusions.* Moreover, using the method of Epistemic Network Analysis, we operationalized diagnostic practices as the relative frequencies of co-occurring diagnostic activities. We found significant differences between learners from medical education and teacher education with respect to both their diagnostic activities and diagnostic practices. Learners from medical education put relatively more emphasis on generating hypotheses and drawing conclusions, therefore applying a more hypothesis-driven approach. By contrast, learners in teacher education had a stronger focus on generating and evaluating evidence, indicating a more data-driven approach. The results may be explained by different epistemic ideals and standards taught in higher education. Further research on the issue of epistemic ideals and standards in diagnosing is needed. Moreover, we recommend that educators think beyond individuals' knowledge and implement measures to systematically teach and increase the awareness of disciplinary standards.

Keywords: diagnostic activities, diagnostic practices, medical education, teacher education, interdisciplinary research

OPEN ACCESS

Edited by:
Bernhard Ertl,
Munich University of the Federal Armed Forces, Germany

Reviewed by:
Tom Rosman,
Leibniz Institute for Psychology Information and Documentation (ZPID), Germany
Mohamed Taha Mohamed,
British University in Egypt, Egypt

****Correspondence:***
Elisabeth Bauer
elisabeth.bauer@psy.lmu.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *15 May 2020*
Accepted: *23 September 2020*
Published: *20 October 2020*

Citation:
Bauer E, Fischer F, Kiesewetter J, Shaffer DW, Fischer MR, Zottmann JM and Sailer M (2020) Diagnostic Activities and Diagnostic Practices in Medical Education and Teacher Education: An Interdisciplinary Comparison. Front. Psychol. 11:562665. doi: 10.3389/fpsyg.2020.562665

INTRODUCTION

Interdisciplinary research involves various challenges, for example, the comparability of specific variables. In this article, we refer to a framework of diagnostic activities (Fischer et al., 2014; Heitzmann et al., 2019) that was applied to compare learners' diagnostic assessments within two disciplines (i.e., medical education and teacher education). We aim to investigate diagnostic

activities in these disciplines and explore their conceptual integration into diagnostic practices. Hereby, we also seek to facilitate future interdisciplinary research on diagnostic practices and the learning of diagnostic activities.

Facilitating diagnostic skills in higher education is an important objective in many disciplines (e.g., Chernikova et al., 2020). This is certainly the case in medical education, which focuses on training future physicians in the assessment of patient symptomology. Similarly, future teachers' professional challenges include diagnosing students' performance, progress, learning difficulties such as behavioral and learning disorders, or other learning prerequisites (Reinke et al., 2011). Independent of the discipline, we broadly define diagnosing as "a process of goal-oriented collection and integration of case-specific information to reduce uncertainty in order to make medical or educational decisions" (Heitzmann et al., 2019, p.4).

Professional knowledge is a crucial prerequisite for diagnosing (Blömeke et al., 2015). There are numerous models conceptualizing professional knowledge (e.g., Shulman, 1987; Kopp et al., 2009; Charlin et al., 2012), e.g., in terms of content like biological knowledge in medicine (Charlin et al., 2012) and pedagogical knowledge in teaching (Shulman, 1987). Research has even suggested that professional knowledge in diagnostic reasoning may not only be discipline-specific but case-specific, since abstract types of e.g., strategic knowledge (Kopp et al., 2009) do not seem to transfer well across cases (e.g., Wimmers et al., 2007; Schwartz and Elstein, 2009). A recently proposed interdisciplinary perspective on professional diagnostic knowledge integrated conceptualizations in medical education and teacher education into an interdisciplinary model with the two dimensions of content-related facets and abstract types of knowledge (Förtsch et al., 2018). The model acknowledges that the issue of content-specificity also affects abstractions like types of professional knowledge, and thus emphasizes limited comparability of professional diagnostic knowledge across disciplines.

We argue nonetheless that interdisciplinary research in diagnosing may still benefit from a more abstracted level of observation, namely: diagnostic practices. We build on the idea of epistemic practices, which are defined as "the specific ways members of a community propose, justify, evaluate, and legitimize knowledge claims within a disciplinary framework" (Kelly, 2008, p. 99). Epistemic practices involve community-specific or discipline-specific epistemic aims (e.g., that a claim is justified), epistemic ideals (standards and criteria to assess the achievement of aims, e.g., that the evidence supports the claim and disconfirms competing claims), and processes that are considered reliable (e.g., disconfirming competing claims; Duncan and Chinn, 2016). Transferring the idea of epistemic practices into the context of diagnosing, we define diagnostic practices as systematic approaches that are applied to collect and integrate case-specific information to reduce uncertainty, and to make and communicate informed and justifiable decisions in a professional situation (Kelly, 2008; Heitzmann et al., 2019). We assume that diagnostic practices within disciplines may involve specificities concerning their epistemic aims, ideals and processes (Duncan and Chinn, 2016), e.g., the standards for justifying a diagnosis. Therefore, comparing diagnostic practices across disciplines may improve our understanding and facilitate future research.

To conceptualize diagnostic practices across different disciplines, we refer to underlying diagnostic activities such as *generating hypotheses*, *generating evidence*, *evaluating evidence*, and *drawing conclusions* (Fischer et al., 2014; Heitzmann et al., 2019; see **Supplementary Material** section "Supplementary Illustration of the Framework of Diagnostic Activities" for further details). The activities framework has been investigated in different disciplines, e.g., social work education (Ghanem et al., 2018), teacher education (Csanadi et al., 2018), and medical education (Lenzer et al., 2017). We assume, that although concrete hypotheses, evidence, and conclusions are specific, the epistemic purpose of these diagnostic activities is conceptually transferable across disciplines (Hetmanek et al., 2018): Although different hypotheses are appropriate for different diagnostic cases, the activity of generating hypotheses holds the purpose of identifying potential explanations, which may require further investigation. Thus, in investigating diagnostic activities, the case-specific content may be less important compared to characteristics concerning the structure of cases (e.g., the form and amount of potentially available evidence).

As a starting point in investigating diagnostic practices, we can interpret and integrate disciplinary conceptualizations used in previous research in terms of diagnostic activities: In medical education, research has focused in particular on process characteristics of diagnostic reasoning (e.g., Coderre et al., 2003; Norman, 2005; Mamede and Schmidt, 2017). Several studies found that medical students conform to a diagnostic practice, which was characterized as hypothesis-driven approach: Students generated different hypotheses and evaluated evidence accordingly to draw conclusions about their initial hypotheses (e.g., Coderre et al., 2010; Kiesewetter et al., 2013). The hypothesis-driven approach reflects an epistemic ideal of differential diagnosing, which is considered a reliable process in medicine and is thereby taught in medical education (see Duncan and Chinn, 2016). However, research has also found that some medical students exhibit a data-driven approach instead, which focuses on generating and evaluating evidence without considering specific hypotheses or integrating evidence into conclusions (e.g., Gräsel and Mandl, 1993; Norman et al., 2007; Kiesewetter et al., 2013).

In teacher education, research has mostly conceptualized diagnostic practices in terms of professional vision (Goodwin, 1994). Two subcomponents of professional vision have been distinguished: noticing, which includes identifying problems and generating hypotheses, and reasoning, which comprises describing, explaining, and predicting (e.g., Seidel and Stürmer, 2014). *Describing* refers to reporting generated evidence. *Explaining* means evaluating evidence in reference to professional knowledge. Therefore, describing and explaining both focus on evidence and seldom involve generating hypotheses or drawing conclusions, both of which point to *predicting* consequences of observations. Research indicates that expert teachers integrate *describing*, *explaining*, and *predicting* into their diagnostic practice (Seidel and Prenzel, 2007).

However, *describing* seems to be a prevailing aspect, while the use of *predicting* is more variant (Stürmer et al., 2016).

Given that work surrounding diagnostic assessment has primarily emerged from the disciplines of medical education and teacher education, we aimed to compare and integrate these two theoretical approaches with respect to diagnostic activities and diagnostic practices. Specifically, we operationalized diagnostic practices as the co-occurrence of diagnostic activities, which we investigated via the use of Epistemic Network Analysis (ENA) (Shaffer, 2017). The research questions are as following:

RQ1: To what extent do learners' *diagnostic activities* differ between medical education and teacher education?

RQ2: To what extent do learners' *diagnostic practices* differ between medical education and teacher education?

METHOD

Participants

A total of 142 medical students and 122 pre-service teachers participated in two matched data collections. Medical students were in their 5th to 11th semester ($M = 8.15$; SD = 1.82). Their mean age was $M = 24.41$ (SD = 2.89). A total of 102 were women and 40 were men. Pre-service teachers were in their 1st to 13th semester ($M = 4.55$; SD = 3.40), were on average $M = 22.96$ years old (SD = 4.10), and were mostly women (106 women; 15 men; 1 non-binary). Since half of the sample in teacher education was in their 1st to 4th semester, we defined a subsample of students in teacher education in the 5th or a higher semester for additional subsample analyses (see **Supplementary Material** section "Supplementary Subsample Analyses").

Materials

We developed simulation-based learning environments for medical education and teacher education, using the authoring tool CASUS (Hege et al., 2017). Both learning environments included eight cases with a parallel structure: The cases began with an initial problem concerning a virtual patient or student. Next, learners could freely choose to access several informational sources in any sequence. Learners solved two tasks in each of the eight cases: First, they provided a diagnosis of the virtual patient or virtual student's problem; second, they had to write a justificatory report, after being prompted, to justify their diagnosis by indicating how they approached and processed the case information.

The medical education cases presented virtual patients with symptoms of fever and back pain. Medical students were asked to take over the role of a general practitioner. After reading the initial problem statement, where the patient revealed his or her reason for seeing a physician, learners accessed the patient's history and had the option to access the results of different examinations and tests, e.g., physical examination, laboratory, X-ray, ECG.

In the teacher education cases, we asked pre-service teachers to take over the role of a teacher who was encountering a student with some initial performance-related or behavioral problems that might even be clinically relevant, e.g., ADHD or dyslexia. We chose these topics because they are relevant for teachers and at the same time entail structural similarities to medical cases. After reading the initial problem, the learners could access informational sources such as reports of observations from inside and outside of the classroom as well as transcripts of conversations with the student, the parents, and other teachers. Moreover, participants could explore samples of the student's written exercises and school certificates.

For further details on the learning environment and the cases used, see **Supplementary Material** sections "Supplementary Case Materials for Medical Education" and "Supplementary Case Materials for Teacher Education."

Procedure

The data collection was computer-based and took place in a laboratory setting. We introduced participants to the aims, scope, and procedure of the study and familiarized them with the materials. Next, participants entered the simulation-based learning environment that was designed for their field of study. After giving informed consent to participate in the study, they had to answer a knowledge pretest that took up to 35 min. Afterward, they entered the learning phase, consisting of the eight simulated cases of their respective discipline. Time on task for all cases was $M = 45.1$ min (SD = 12.2) in medical education and $M = 51.8$ min (SD = 16.5) in teacher education. After four cases, participants took a break of 10 min before continuing with the second part of the learning phase and solving cases five to eight. Subsequently, they had to answer a knowledge posttest, which again took up to 35 min. Finally, participants received monetary compensation.

Data Sources and Instruments

For this paper, we analyzed only the text data from the justificatory reports that all learners wrote for the eight simulated cases. Participants wrote the justificatory reports in an empty text field, right after indicating their diagnosis for each case. There was no template or additional support apart from the standardized prompt to justify the diagnosis by indicating how they approached the case and how they processed the case information. The overall data set used in this paper consisted of 1,136 justificatory reports written by the 142 medical students (average number of words per report $M = 57.4$; SD = 32.6) and 976 justificatory reports written by the 122 pre-service teachers (average number of words per report $M = 89.6$; SD = 53.2).

Diagnostic Activities

We coded the two sets of justificatory reports on four diagnostic activities: *generating hypotheses*, *generating evidence*, *evaluating evidence*, and *drawing conclusions*. **Table 1** presents definitions and examples of the four codes. We developed a coding scheme applicable for medical education and teacher education. Coding and segmentation were done simultaneously to account for overlap in the activities as well. In both disciplines, the raters were first to second year doctoral students and student assistants (minimum 6th semester) from the respective fields. All raters were blind to this study's research questions. Raters did four

TABLE 1 | Definitions, examples, and inter-rater reliabilities (IRRs indicated as Krippendorff's α_U) for the four codes: *generating hypotheses*, *generating evidence*, *evaluating evidence*, and *drawing conclusions*.

		Medical education		Teacher education	
Code	**Definition**	**Example**	**IRR**	**Example**	**IRR**
Generating hypotheses	Explicit collection of different potential diagnoses or pointing to one diagnosis involving expressed insecurity, e.g., using conjunctive mood.	I believe this is a case of nerve entrapment.	0.60	The initial information makes me think of impaired vision, a reading disorder, or emotional problems as potential explanations for Annika's issues.	0.43
Generating evidence	Explicit description of accessing informational sources, e.g., tests, interviews, or observations.	Subsequently, I looked at the MRI and X-ray.	0.65	I observed Anna's school-related behavior and achievement.	0.56
Evaluating evidence	Explicit listing and/or interpretation of separate case information.	Among other results, the patient has an increased CRP and leukocytosis.	0.75	Markus behaves aggressively and gets offended very easily.	0.75
Drawing conclusions	Explicit conclusion or rejection of at least one diagnosis.	The patient clearly has tonsillitis involving a fever.	0.65	Consequently, I rejected the diagnosis of ADHD.	0.49

rounds of joined coding training, starting with 20 reports and increasing the number in every round of training. To evaluate inter-rater reliability (IRR), five raters in medical education and four in teacher education coded 150 reports for the respective project (13% of the data set in medical education; 15% in teacher education). The overall IRR for the simultaneous segmentation and coding was Krippendorff's $\alpha_U = 0.67$ in medical education and $\alpha_U = 0.65$ in teacher education (see **Table 1**), which we consider as satisfactory. For the analyses, we calculated the share of diagnostic activities within medical education and teacher education, respectively, as the percentages of the different diagnostic activities relative to the overall amount.

Diagnostic Practices

We operationalized diagnostic practices as the co-occurrences of diagnostic activities in the justificatory reports, using the method of ENA (Shaffer, 2017). The ENA algorithm analyzes co-occurring diagnostic activities within a moving window of two sentences (Siebert-Evenstone et al., 2017). Therefore, subsequent to the coding, we determined presence or absence of the four diagnostic activities per sentence. We accumulated the co-occurrences and created one network graph per discipline. In the network graphs, the colored edges refer to co-occurrences between diagnostic activities, with thickness indicating their relative frequencies. Relative frequencies of co-occurring activities allowed us to draw inferences about the general diagnostic practices of each discipline. Additionally, a comparison graph (i.e., showing only the difference between both graphs), allowed us to isolate the differences between the two disciplines' diagnostic practices.

We also centered the networks and created one centroid per learner as well as per discipline. The centroids' position is relative to the co-occurrences between diagnostic activities in the respective network. On the level of single learners, the representation of centroids can be used to depict the learners' distribution within the network space, which can be interpreted as an indicator of interindividual heterogeneity in diagnostic practices. On the level of disciplines, we can consider centroids as group means. ENA enables statistical testing of the group differences in overall diagnostic practices between learners in medical education and teacher education. To facilitate the testing of the group differences, we used the option of means rotation, which aligns the two disciplines' group means on the X-axis, thus depicting systematic variance on only one dimension.

Statistical Analyses

To address RQ1, the extent to which diagnostic activities differ between learners from medical education and teacher education, we calculated t tests for independent samples, one test per diagnostic activity, using Bonferroni-adjusted alpha levels of $\alpha = 0.0125$ per test ($\alpha = 0.05/4$). To statistically test RQ2, differences in diagnostic practices between learners from medical and teacher education, we used an independent-samples t test as well, comparing the two group means from the two disciplines' ENA networks at an alpha level of $\alpha = 0.05$. If Levene's test indicated unequal variances, we adjusted the degrees of freedom accordingly.

RESULTS

Comparing the two disciplines, there was a significant difference regarding the number of semesters studied (medical education $M = 8.15$; SD = 1.82; teacher education $M = 4.55$; SD = 3.40), $t(173) = 10.35$, $p < 0.001$, Cohen's $d = 2.75$. Therefore, we analyzed the relation with the percentages of diagnostic activities within the disciplines. There was no significant correlation found between number of semesters studied and the percentages of the different diagnostic activities (for details see **Supplementary Material** section "Supplementary Results of a Correlation Between Semesters Studied and Number of Diagnostic Activities"). However, to ensure that the number of semesters studied did not bias the results, we performed the following analyses not only with the full sample as reported in the following sections, but a second time, comparing learners from medical education to the specified subsample of learners from teacher education in their 5th or a higher semester (see **Supplementary Material** section "Supplementary Subsample Analyses").

Diagnostic Activities in Medical Education and Teacher Education (RQ1)

In both disciplines, *evaluating evidence* was clearly the most prominent activity found in the justificatory reports with a share of more than half of the diagnostic activities found in the reports (medical education $M = 60.96\%$; SD = 10.24%; teacher education $M = 66.08\%$; SD = 17.02%). The difference in the relative frequencies for *evaluating evidence* was significant with a small effect size [$t(192) = 2.91$, $p = 0.004$, Cohen's $d = 0.37$]. We found that in medical education, the share for *generating hypotheses* was about twice as high ($M = 16.26\%$; SD = 7.96%) as in teacher education ($M = 8.37\%$; SD = 6.41%). This difference was significant with a large effect size [$t(261) = 8.92$, $p < 0.001$, Cohen's $d = 1.08$]. By contrast, the share for *generating evidence* was about twice as high in teacher education ($M = 13.74\%$; SD = 14.81%) as in medical education ($M = 6.79\%$; SD = 8.26%), and this was also significantly different with a medium-sized effect [$t(183) = 4.60$, $p < 0.001$, Cohen's $d = 0.59$]. In medical education, we also found a significantly higher share for *drawing conclusions* ($M = 15.99\%$; SD = 6.39%) than in teacher education ($M = 11.82\%$; SD = 6.83%), with a medium effect size [$t(262) = 5.13$, $p < 0.001$, Cohen's $d = 0.63$].

Comparing medical education with the specified subsample from teacher education (see section "Participants"), the results show the same results pattern (for detailed results see **Supplementary Material** section "Supplementary Subsample Analyses"). However, there was no significant difference in the relative frequencies for *evaluating evidence* [medical education $M = 60.96\%$; SD = 10.24%; teacher education $M = 65.40\%$; SD = 18.00%; $t(77) = 1.81$, $p = 0.075$, Cohen's $d = 0.34$].

Diagnostic Practices in Medical Education and Teacher Education (RQ2)

In **Figure 1**, we present the diagnostic practices of learners from medical education (**Figure 1A**) and teacher education (**Figure 1C**) as network graphs. The colored edges and their thickness reflect the relative frequencies of co-occurrences of diagnostic activities. The overall network across all learners from medical education (**Figure 1A**) showed some similarities to the overall network across all learners from teacher education (**Figure 1C**): First, in both disciplines, we found that the relative frequencies of co-occurrences were in accordance with the relative frequencies of the individual diagnostic activities (see the results for RQ1). In both network graphs, the three relatively most frequent co-occurrences were the ones including *evaluating evidence*. This is why we found *evaluating evidence* near the center of the disciplines' overall networks. However, by looking at its temporal context indicated by co-occurrences with other diagnostic activities, we can draw inferences about the purpose of *evaluating evidence* within the respective context. When it co-occurs with *drawing conclusions* or *generating hypotheses*, *evaluating evidence* serves the purpose of *explaining*; whereas when co-occurring with *generating evidence*, *evaluating evidence* may rather *describe* the evidence (see **Table 2** for examples). To compare learners from medical education and teacher education, the comparison graph (**Figure 1B**) shows the difference between the two disciplines' overall networks, therefore indicating only the differences in co-occurrences. In medical education, there was a relatively higher frequency of *evaluating evidence* co-occurring with *generating hypotheses*, pointing to a rather hypothesis-driven approach that puts more emphasis on *explaining* evidence; whereas learners in teacher education exhibited a relatively higher frequency of co-occurrences between *evaluating evidence* and *generating evidence*, indicating a tendency toward *describing* evidence or a data-driven approach.

In addition to the disciplines' overall networks, **Figure 2** presents the distribution of single learners across the two disciplines' overall networks. The colored points represent the networks' centroids on the level of single learners from medical education (**Figure 2A**) and teacher education (**Figure 2C**). In teacher education, single learners' centroids (red colored points) are more scattered across the network space, compared to the positioning of the single learners' centroids in medical education (blue colored points). This indicates that the diagnostic practices

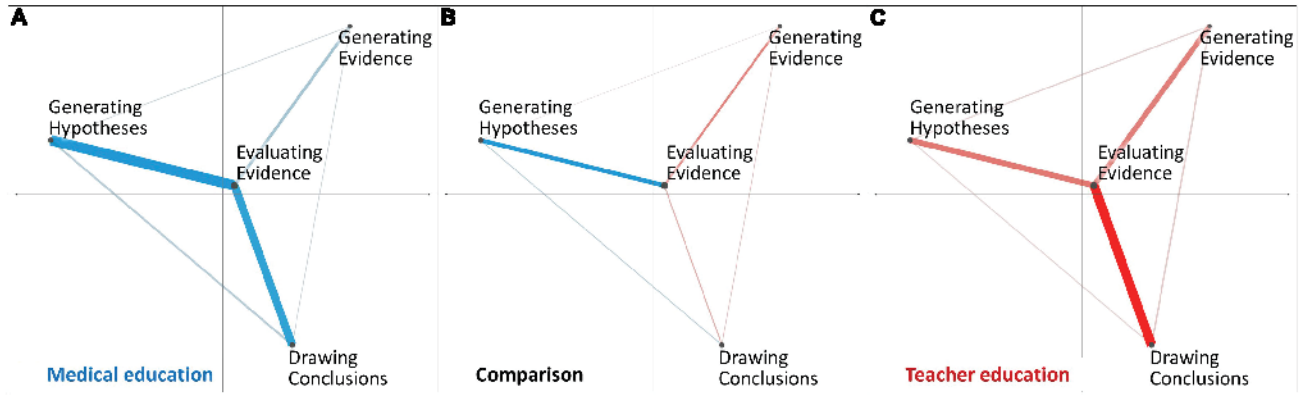


FIGURE 1 | ENA networks from medical education **(A)**, and teacher education **(C)**. The comparison network **(B)** depicts only the differences between the other two networks.

TABLE 2 | Examples of *evaluating evidence*, co-occurring with *generating evidence*, *generating hypotheses*, or *drawing conclusions* in a temporal context of one to two sentences in the disciplines of medical education and teacher education.

Case	Text	Generating hypotheses	Generating evidence	Evaluative evidence	Drawing conclusions
Section a: Examples of *evaluating evidence* co-occurring with *drawing conclusions* or *generating hypotheses* in the discipline of medical education					
2	Due to his age and the sudden symptomatology in only his lumbar spine, I would diagnose a rheumatic disease.	0	0	1	1
7	Upon physical examination, she mostly indicated pain in the upper abdomen, which highlights the region of the liver, gall bladder, and eventually the biliary tract and pancreatic duct.	0	0	1	0
	Laboratory results indicated increased liver values, which is why I believe the patient has hepatitis.	1	0	1	0
Section b: Examples of *evaluating evidence* co-occurring with *drawing conclusions* or *generating hypotheses* in the discipline of teacher education					
8	The characteristic writing, confusion of characters, deficits in stringing together syllables, as well as deficits in syllabification and slow reading speed, combined with an otherwise good school performance, clearly indicate dyslexia.	0	0	1	1
6	Thomas might have eventually developed ADHD and therefore low concentration.	1	0	0	0
	This assumption is backed by the fact that his performance in all subjects decreased and that he does not fully answer all questions on exams.	0	0	1	0
Section c: Examples of *evaluating evidence* co-occurring with *generating evidence* in the discipline of medical education					
7	First, I examined all the available information, before focusing on the most relevant points.	0	1	0	0
	They mostly seemed to be related to the liver.	0	0	1	0
8	Even after being treated by the general practitioner, the patient still had a fever and symptoms of a systemic infection.	0	0	1	0
	This is why, considering the anamnesis regarding previous travels, I decided to administer an HIV test.	0	1	1	0
Section d: Examples of *evaluating evidence* co-occurring with *generating evidence* in the discipline of teacher education					
6	I examined the teacher's report and the available documents.	0	1	0	0
	It seems that Thomas' symptoms have only been observable recently and that he has repeatedly complained about small font sizes.	0	0	1	0
5	Initially, I collected information from observations, conversations, the annual report, and recent school exams.	0	1	0	0
2	My attention was caught by the mother's description of her reading behavior at home, especially in terms of reading aloud.	0	0	1	0

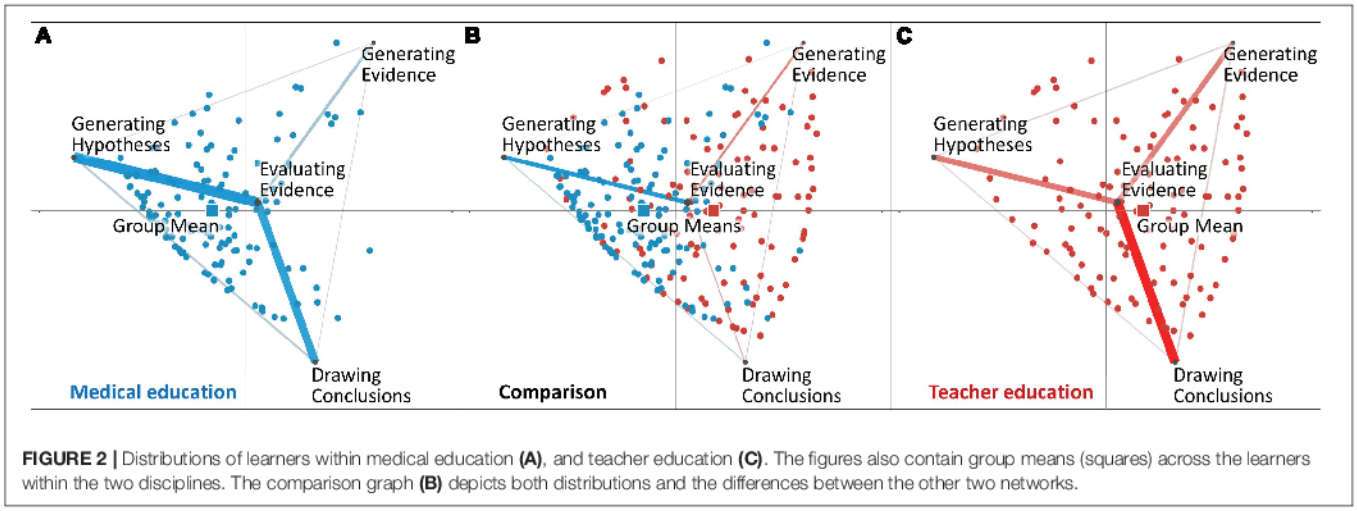


FIGURE 2 | Distributions of learners within medical education **(A)**, and teacher education **(C)**. The figures also contain group means (squares) across the learners within the two disciplines. The comparison graph **(B)** depicts both distributions and the differences between the other two networks.

of learners from medical education are more homogeneous compared with the diagnostic practices of learners from teacher education.

Figure 2 presents centroids on the group level, representing the means of all learners within the two disciplines of medical education and teacher education as indicated by

the colored squares. The positioning of the group mean of learners from medical education ($M = -0.36$, SD = 0.63, $N = 142$) was statistically significantly different from the positioning of the group mean of learners from teacher education [$M = 0.42$, SD = 0.74, $N = 122$; $t(240.48) = -9.16$, $p < 0.01$, Cohen's $d = 1.14$]. This result indicates a significant difference in diagnostic practices between teacher education and medical education. Repeating these analyses, comparing students from medical education with the specified subsample from teacher education, revealed basically the same result (for details see **Supplementary Material** section "Supplementary Subsample Analyses").

DISCUSSION

In analyzing learners' reports of their diagnostic activities in medical education and teacher education, we found that future physicians and future teachers put the most focus toward *evaluating evidence*. Moreover, learners from teacher education focused more on *generating evidence*, whereas learners from medical education put more focus toward *generating hypotheses* and *drawing conclusions*. These results support the notion that the relative emphasis on each diagnostic activity differs between these disciplines.

The disciplinary differences in the use of diagnostic activities is also reflected by overall diagnostic practices. Because the overall network across all learners from medical education was similar to the network across all learners from teacher education, this similarity suggests that the overall diagnostic practices are similar. Still, there were significant disciplinary differences in the relative frequencies of the co-occurrences of diagnostic activities. In general, we found that learners from medical education showed a more explanation-driven or hypothesis-driven approach (see Coderre et al., 2010; Kiesewetter et al., 2013; Seidel and Stürmer, 2014), whereas learners from teacher education showed a more description-driven or data-driven approach (see Gräsel and Mandl, 1993; Norman et al., 2007; Kiesewetter et al., 2013; Seidel and Stürmer, 2014). Furthermore, learners from teacher education showed greater variability in their diagnostic practices than learners from medical education.

We interpret the results relating to epistemic ideals as the "criteria or standards used to evaluate epistemic products" (Duncan and Chinn, 2016, p. 158). In the context of medical education, differential diagnosing is considered as ideal for ensuring a reliable process. Differential diagnosing essentially refers to a hypothesis-driven approach of generating and testing hypotheses (see Fischer et al., 2014), which is what we observed in learners from medical education. This diagnostic standard is put into practice on different levels (e.g., in guidelines and university curricula), and is systematically taught to future physicians in their medical programs. In teacher education, we are not aware of a widespread use of such specific standards for diagnosing in general and particularly regarding the topic of students' behavioral and performance-related disorders. Research in teacher education was referred to as a rather "young" field (Grossman and McDonald, 2008) and thus, the evolvement of standards for diagnosing might be less advanced than in medical education. In comparison with medical students, pre-service teachers also seem to show greater variability in their diagnostic practices, which may support the notion of lower standardization in diagnostic practices or at least in educating pre-service teachers to apply diagnostic practices. However, there might be some implicit ideals that enhance pre-service teachers' tendency to embrace a data-driven approach in their diagnostic practices. First, as a reaction to findings of teachers' biases in diagnostic tasks (e.g., Südkamp et al., 2012), some teacher education programs have subsequently taught the concept of professional vision (Goodwin, 1994) to pre-service teachers, emphasizing the need to focus on *describing* observations before *explaining* them (e.g., Seidel and Stürmer, 2014). This development may complement other implicit values (see Duncan and Chinn, 2016) in teaching, such as to avoid being judgmental toward students (Aalberts et al., 2012). Therefore, the findings may reflect disciplinary differences in epistemic ideals implemented in higher education and diagnostic practices, respectively.

Limitations

One limitation of the study involves the inter-rater reliabilities for *generating hypotheses* and *drawing conclusions*, which were relatively low in the teacher education data. This could limit the conclusions that can be drawn about the variability in diagnostic practices of teacher education learners in particular.

Another limitation may be the learners' study progress: In the full sample, learners from medical education had completed significantly more semesters than learners from teacher education. However, the number of semesters did not correlate with the proportion of the different diagnostic activities. The subsample analyses, which compared students from medical education with students from teacher education in their 5th or a higher semester revealed the same patterns of results as the analyses of the full sample. Hence, it seems unlikely that the *a priori* difference in the number of semesters would lead to substantial bias in our results.

Furthermore, we acknowledge that although we argue for the interdisciplinary comparability of the diagnostic activities' epistemic purpose, this conceptualization may still not fully eliminate the issues associated with comparing disciplinary diagnostic practices. Yet, we think that diagnostic activities and diagnostic practices are more advantageous in terms of interdisciplinary comparability than other investigated approaches, e.g., professional diagnostic knowledge.

The choice of clinical topics in both disciplines served the purpose of having similarly structured problems. Nevertheless, in teacher education there are other than clinical areas where diagnosing is relevant (e.g., assessing a student's level of skill). Thus, our choice might limit the generalizability of the findings to other areas of assessment in teacher education. However, if we consider diagnostic practices as discipline-specific approaches, it is reasonable to assume that the findings may replicate in other areas of teachers' diagnostic assessments, which could be investigated in further research.

Finally, similar to verbal protocols, assessing reported activities raises the question of validity, concerning the degree to which the reports effectively represent actually performed activities. Therefore, further research might additionally complement reported diagnostic activities with behavioral data like user-logs.

CONCLUSION

In this article, we have argued that interdisciplinary research on diagnostic assessments benefits from comparisons drawn at the level of diagnostic activities (Fischer et al., 2014) and diagnostic practices (Kelly, 2008; Heitzmann et al., 2019) as comparing professional diagnostic knowledge has been found to be difficult due to its content specificity. In an interdisciplinary comparison of justifications by learners from teacher education and medical education, we found significant differences in their diagnostic activities and diagnostic practices. We found a more hypothesis-driven approach in justifications of learners from medical education, who put relatively more emphasis on generating hypotheses and drawing conclusions. Learners from teacher education instead seemed to apply a more data-driven approach, with a stronger focus on generating and evaluating evidence. The results may allude to different epistemic ideals and diagnostic standards (see Duncan and Chinn, 2016) taught in higher education and thereby put into diagnostic practices.

Diagnostic activities can provide a useful and interdisciplinary framework to analyze diagnostic practices across disciplines. For future interdisciplinary research, we recommend considering matched study designs, as implemented in our project, to maximize interdisciplinary comparability. Additionally, from a practically oriented viewpoint, we recommend that educators from both the medical education and teacher education fields reflect further on their standards in diagnosing and their underlying epistemic ideals to further increase the awareness of practitioners and systematization in teaching. Finally, we encourage researchers to further investigate the potential relation between epistemic ideals and diagnostic practices in terms of interdisciplinary differences, commonalities, and their continuing evolvement.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors. Requests to access the data should be directed to elisabeth.bauer@psy.lmu.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Medical Faculty of LMU Munich (no. 17-249). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

EB, FF, MF, JK, MS, and JZ developed the study concept and contributed to the study design. EB and MS performed the data analysis. EB, FF, and MS interpreted the data. EB drafted the manuscript. FF, MF, JK, MS, DS, and JZ provided critical revisions. All authors approved the final version of the manuscript for submission.

FUNDING

This research was funded by a grant of the German Federal Ministry of Research and Education (16DHL1040 und 16DHL1039) and the Elite Network of Bavaria (K-GS-2012-209). Parts of this work were funded by the National Science Foundation (DRL-1661036, DRL-1713110), the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.562665/full#supplementary-material

REFERENCES

Aalberts, J., Koster, E., and Boschhuizen, R. (2012). From prejudice to reasonable judgement: integrating (moral) value discussions in university courses. *J. Moral Educ.* 41, 437–455. doi: 10.1080/03057240.2012.677600

Artelt, C., and Rausch, T. (2014). "Accuracy of teacher judgments: when and for what reasons?," in *Teachers' Professional Development*, eds S. Krolak-Schwerdt, S. Glock, and M. Böhmer (Leiden: Brill | Sense), 27–43. doi: 10.1007/978-94-6209-536-6_3

Blömeke, S., Gustafsson, J.-E., and Shavelson, R. J. (2015). Beyond dichotomies: competence viewed as a continuum. *Z. Psychol.* 223, 3–13. doi: 10.1027/2151-2604/a000194

Charlin, B., Lubarsky, S., Millette, B., Crevier, F., Audétat, M. C., Charbonneau, A., et al. (2012). Clinical reasoning processes: unravelling complexity through graphical representation. *Med. Educ.* 46, 454–463. doi: 10.1111/j.1365-2923.2012.04242.x

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., and Fischer, F. (2020). Facilitating diagnostic competences in higher education—a meta-analysis in medical and teacher education. *Educ. Psychol. Rev.* 32, 157–196. doi: 10.1007/s10648-019-09492-2

Coderre, S., Mandin, H., Harasym, P. H., and Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Med. Educ.* 37, 695–703. doi: 10.1046/j.1365-2923.2003.01577.x

Coderre, S., Wright, B., and McLaughlin, K. (2010). To think is good: querying an initial hypothesis reduces diagnostic error in medical students. *Acad. Med.* 85, 1125–1129. doi: 10.1097/acm.0b013e3181e1b229

Csanadi, A., Eagan, B., Kollar, I., Shaffer, D. W., and Fischer, F. (2018). When coding-and-counting is not enough: using epistemic network analysis (ENA)

to analyze verbal data in CSCL research. *Int. J. Comput. Support. Collab. Learn.* 13, 419–438. doi: 10.1007/s11412-018-9292-z

Duncan, R. G., and Chinn, C. A. (2016). New directions for research on argumentation: insights from the AIR framework for epistemic cognition. *Z. Padagog. Psychol.* 30, 155–161. doi: 10.1024/1010-0652/a000178

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Front. Learn. Res.* 2, 28–45. doi: 10.14786/flr.v2i3.96

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M., Girwidz, R., Obersteiner, A., et al. (2018). Systematizing professional knowledge of medical doctors and teachers: development of an interdisciplinary framework in the context of diagnostic competences. *Educ. Sci.* 8:207. doi: 10.3390/educsci8040207

Ghanem, C., Kollar, I., Fischer, F., Lawson, T. R., and Pankofer, S. (2018). How do social work novices and experts solve professional problems? A micro-analysis of epistemic activities and the use of evidence. *Eur. J. Soc. Work.* 21, 3–19. doi: 10.1080/13691457.2016.1255931

Goodwin, C. (1994). Professional vision. *Am. Anthropol.* 96, 606–633.

Gräsel, C., and Mandl, H. (1993). Förderung des Erwerbs diagnostischer Strategien in fallbasierten Lernumgebungen. *Unterrichtswissenschaft* 21, 355–369.

Grossman, P., and McDonald, M. (2008). Back to the future: directions for research in teaching and teacher education. *Am. Educ. Res. J.* 45, 184–205. doi: 10.3102/0002831207312906

Hege, I., Kononowicz, A. A., and Adler, M. (2017). A clinical reasoning tool for virtual patients: design-based research study. *JMIR Med. Educ.* 3:e21. doi: 10.2196/mededu.8100

Heitzmann, N., Seidel, T., Hetmanek, A., Wecker, C., Fischer, M. R., Ufer, S., et al. (2019). Facilitating diagnostic competences in simulations in higher education. *Front. Learn. Res.* 7, 1–24. doi: 10.14786/flr.v7i4.384

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teach. Teach. Educ.* 76, 181–193. doi: 10.1016/j.tate.2017.12.001

Hetmanek, A., Engelmann, K., Opitz, A., and Fischer, F. (2018). "Beyond intelligence and domain knowledge," in *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*, eds F. Fischer, C. A. Chinn, K. Engelmann, and J. Osborne (New York: Routledge). doi: 10.1016/j.engappai.2013.11.002

Kelly, G. (2008). "Inquiry, activity and epistemic practice," in *Teaching Scientific Inquiry*, eds R. A. Duschl and R. E. Grandy (Leiden: Brill | Sense), 99–117. doi: 10.1163/9789460911453_009

Kiesewetter, J., Ebersbach, R., Görlitz, A., Holzer, M., Fischer, M. R., and Schmidmaier, R. (2013). Cognitive problem solving patterns of medical students correlate with success in diagnostic case solutions. *PLoS One* 8:e71486. doi: 10.1371/journal.pone.0071486

Kopp, V., Stark, R., Kühne-Eversmann, L., and Fischer, M. R. (2009). Do worked examples foster medical students' diagnostic knowledge of hyperthyroidism? *Med. Educ.* 43, 1210–1217. doi: 10.1111/j.1365-2923.2009.03531.x

Lenzer, B., Ghanem, C., Weidenbusch, M., Fischer, M. R., and Zottmann, J. (2017). Scientific reasoning in medical education: a novel approach for the analysis of epistemic activities in clinical case discussions. *Paper Presented at the Conference of the Association for Medical Education in Europe (AMEE)*, Helsinki, Finland.

Mamede, S., and Schmidt, H. G. (2017). Reflection in medical diagnosis: a literature review. *Health Prof. Educ.* 3, 15–25. doi: 10.1016/j.hpe.2017.01.003

Norman, G. (2005). Research in clinical reasoning: past history and current trends. *Med. Educ.* 39, 418–427. doi: 10.1111/j.1365-2929.2005.02127.x

Norman, G., Young, M., and Brooks, L. (2007). Non-analytical models of clinical reasoning: the role of experience. *Med. Educ.* 41, 1140–1145.

Reinke, W. M., Stormont, M., Herman, K. C., Puri, R., and Goel, N. (2011). Supporting children's mental health in schools: teacher perceptions of needs, roles, and barriers. *Sch. Psychol. Q.* 26, 1–13. doi: 10.1037/a0022714

Schwartz, A., and Elstein, A. S. (2009). "Clinical problem solving and diagnostic decision making: a selective review of the cognitive research literature," in *The Evidence Base of Clinical Diagnosis*, eds J. A. Knottnerus and F. Buntinx (Hoboken, NJ: Wiley), 237–255. doi: 10.1002/9781444300574.ch12

Seidel, T., and Prenzel, M. (2007). How teachers perceive lessons-assessing educational competencies by means of videos. *Zeitschrift für Erziehungswissenschaft* 10, 201–216.

Seidel, T., and Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *Am. Educ. Res. J.* 51, 739–771. doi: 10.3102/0002831214531321

Shaffer, D. W. (2017). *Quantitative Ethnography*. Madison: Cathcart Press.

Shulman, L. (1987). Knowledge and teaching: foundations of the new reform. *Harv. Educ. Rev.* 57, 1–23. doi: 10.17763/haer.57.1.j463w79r56455411

Siebert-Evenstone, A. L., Irgens, G. A., Collier, W., Swiecki, Z., Ruis, A. R., and Shaffer, D. W. (2017). In search of conversational grain size: modelling semantic structure using moving stanza windows. *J. Learn. Anal.* 4, 123–139.

Stürmer, K., Seidel, T., and Holzberger, D. (2016). Intra-individual differences in developing professional vision: preservice teachers' changes in the course of an innovative teacher education program. *Instr. Sci.* 44, 293–309. doi: 10.1007/s11251-016-9373-1

Südkamp, A., Kaiser, J., and Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *J. Educ. Psychol.* 104:743. doi: 10.1037/a0027627

Wimmers, P. F., Splinter, T. A., Hancock, G. R., and Schmidt, H. G. (2007). Clinical competence: general ability or case-specific? *Adv. Health Sci. Educ.* 12, 299–314. doi: 10.1007/s10459-006-9002-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers

Supplementary Material

1 Supplementary illustration of the framework of diagnostic activities

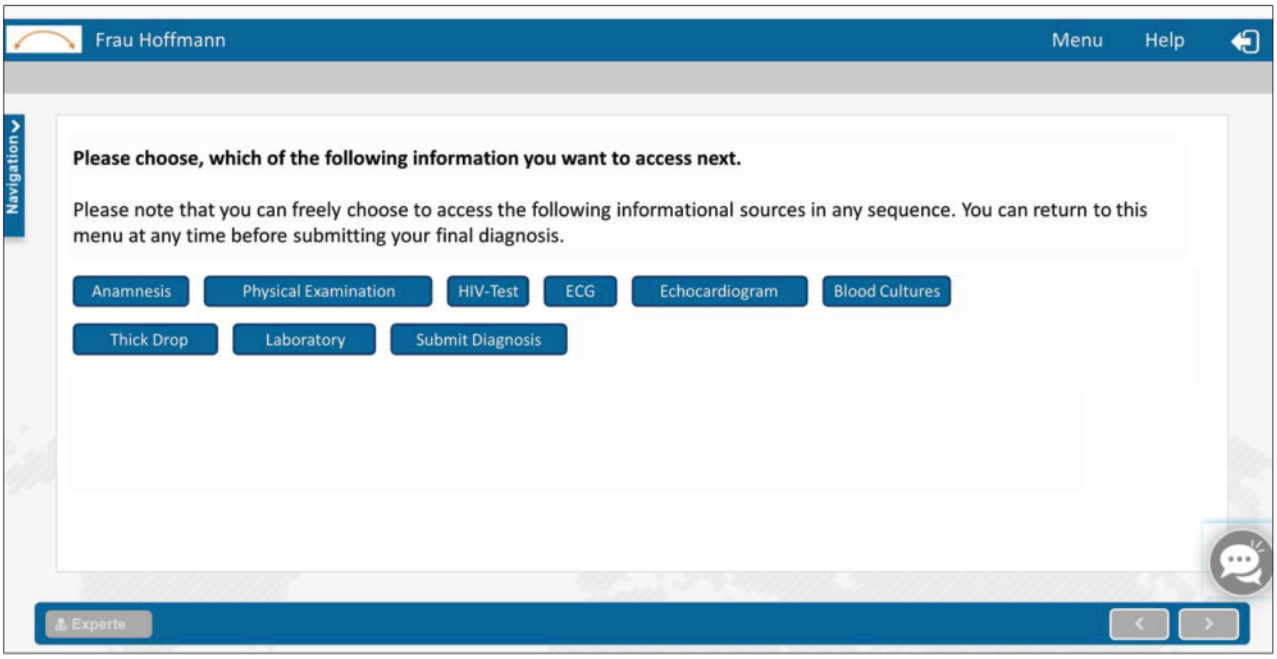
In our research, we refer to a framework of diagnostic activities such as generating hypotheses, generating evidence, evaluating evidence, and drawing conclusions (Fischer et al., 2014; Heitzmann et al., 2019). The full framework includes eight diagnostic activities, which are illustrated in the supplementary Table 1.

Supplementary Table 1. The framework of diagnostic activities, adapted from Heitzmann et al., 2019.

Diagnostic Activity	**Examples from medical education and teacher education**
Problem identification	A physician encounters a patient who reports non-specific symptoms such as shortness of breath; A teacher faces a student who wrongly answers a question in class.
Questioning	A physician asks what the reason for the symptoms could be; A teacher asks what the reason for a student's error could be.
Hypothesis generation	A physician suspects a specific disease, such as a pulmonary embolism; A teacher suspects a specific misconception.
Construction and redesign of artefacts	A medical report which indicates the need for further examination, e.g. a computer tomography; The development of a task which provides insight into the presence of a misconception.
Evidence generation	Conducting further examination, for example through computed tomography; Observation of the student's solution of the task.
Evidence evaluation	Evaluation of the computer tomography with signs of a pulmonary embolism; Evaluation of the solution of the task with some but not all of the signs for the hypothesized misconception.
Drawing conclusions	Deciding that the most likely cause of the patient's symptoms is a pulmonary embolism; Deciding that the most likely reason for the student's error is the assumed misconception, which impedes further learning.
Communication and scrutinization	A medical report with the diagnosis of a pulmonary embolism for another physician; Informing another teacher about the discovered misconception held by a certain student so that teacher can adapt the teaching.

2 Supplementary case materials for medical education

The medical education cases presented virtual patients with symptoms of fever and back pain and medical students were asked to take over the role of a general practitioner. One exemplary case was about a 36 year-old female, Mrs. Hoffmann, who had a febrile and flu-like infection for almost a week before seeing the doctor. In addition, she experienced fatigue, loss of appetite, sickness and diarrhea. One month earlier, she returned from a trip to Costa Rica, for which she received the recommended vaccinations prior to departure. The anamnesis provided the information that no other persons in her surrounding had the same symptoms; that she does not know about any pre-existing illnesses and does not consume any prescribed drugs, apart from occasionally using homeopathic globules; and that, she is allergic to penicillin and nickel; moreover, she is a non-smoker, occasionally consumes alcohol, and excluded the option of pregnancy. To gather more information, learners could access the patient's history and had the option to access different tests and test results, e.g. physical examination, laboratory, x-ray, ECG, HIV test, and others (see supplementary Figure 1). Overall, the symptoms and test results point to an acute hepatitis A infection. Typical symptoms of the patient's current stage of the infection are fatigue, limb pain, fever, sickness, diarrhea and joint pain.

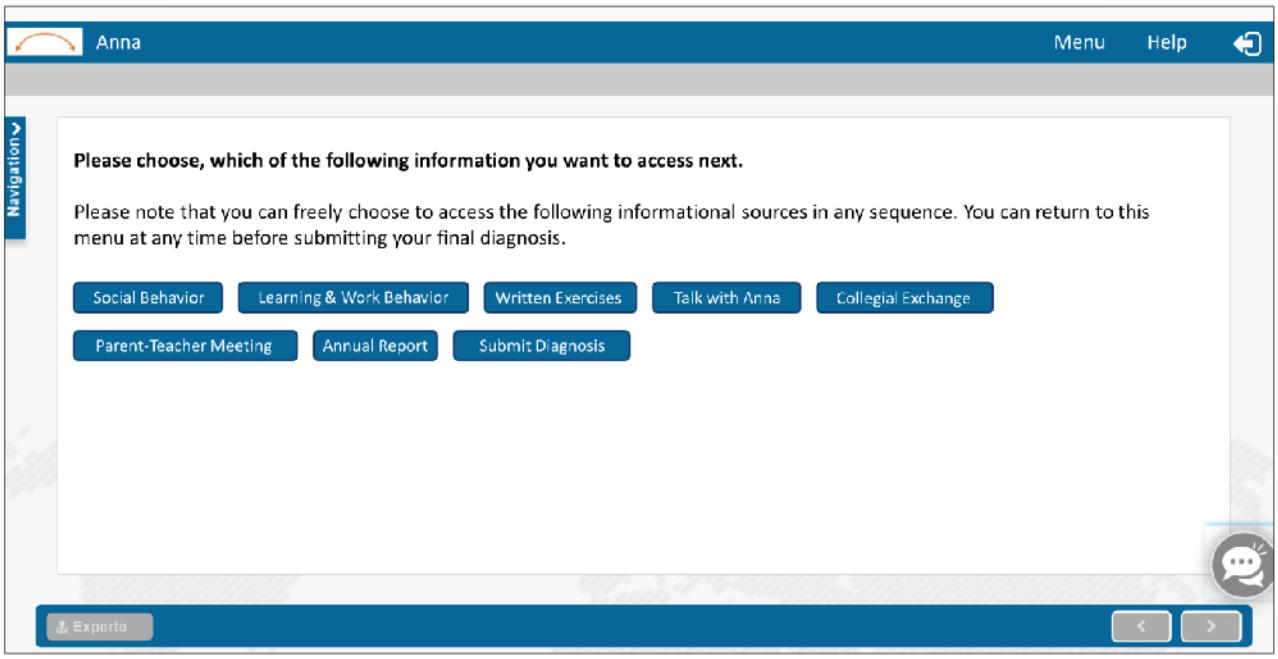


Supplementary Figure 1. Screenshot of user interface for the medical education case in the CASUS learning environment.

3 Supplementary case materials for teacher education

In the teacher education cases, we asked pre-service teachers to imagine themselves in the position of a teacher who was encountering a student with some initial performance-related or behavioral problems that might even be clinically relevant, e.g. ADHD or dyslexia. One example is the case of a secondary student named Anna who is displaying symptoms of an attention-deficit disorder. The leaners are asked to put themselves into the role of Anna's class teacher, who teaches German classes and music lessons. The initial problem statement for the case describes Anna as a 5th grade student,

eleven years old, who constantly needs to be pushed to finish her tasks and who has bad grades in many subjects, especially the main subjects. The learners could examine written observations of Anna's in-class and out-of-class behavior, read recordings of conversations with Anna, or with her parents and several teachers, or look at Anna's last annual report and an example of a written exercise (see supplementary Figure 2). Her behavior is described as very calm and distracted. She is slow in reading and it is difficult for her to answer questions about a text that she just read. She often fails to fulfill the exact instruction of a task or fails to fully complete a task. Moreover, she often does not bring all required school supplies or comes late in the mornings. In a parent-teacher meeting, Anna's mother backs up the impression of a disorganized and slow learning behavior when talking about the homework situation. Anna's last annual report and the conversations with the other teachers show that her grades are mostly affected by her inattentiveness as well, with the exception of artistic subjects and gym classes. Anna mostly interacts with her one friend and is rather distanced from the other students. Anna herself points out that it is hard for her to concentrate since she feels easily distracted. However, at home, where there are fewer ambient noises, she can focus on and enjoy reading, drawing and painting. Overall, the case information was designed in a way that, the diagnosis of an attention-deficit disorder is the most likely clinical diagnosis, despite several differential diagnoses being relevant.



Supplementary Figure 2. Screenshot of user interface for the teacher education case in the CASUS learning environment.

4 Supplementary results of a correlation between semesters studied and number of diagnostic activities.

We analyzed the relation between the relative percentages of diagnostic activities within the disciplines and number of semesters completed. There was no significant correlation found between number of semesters studied and the percentages of the different diagnostic activities. The correlation coefficients and p-values are presented in the supplementary Table 2.

Supplementary Table 2. Results of the two Pearson correlation analyses with the variables *semester* and *percentages of diagnostic activities* within the disciplines of medical education (section a) and teacher education (section b).

		Generating hypotheses	Generating evidence	Evaluating evidence	Drawing conclusions
Section a: Medical Education (N = 142)					
Semester	Pearson's r	.009	-.129	.105	-.012
	p-value	.917	.127	.215	.885
Section b: Teacher Education (N = 119)					
Semester	Pearson's r	.014	.104	-.034	-.140
	p-value	.876	.262	.716	.128

5 Supplementary subsample analyses

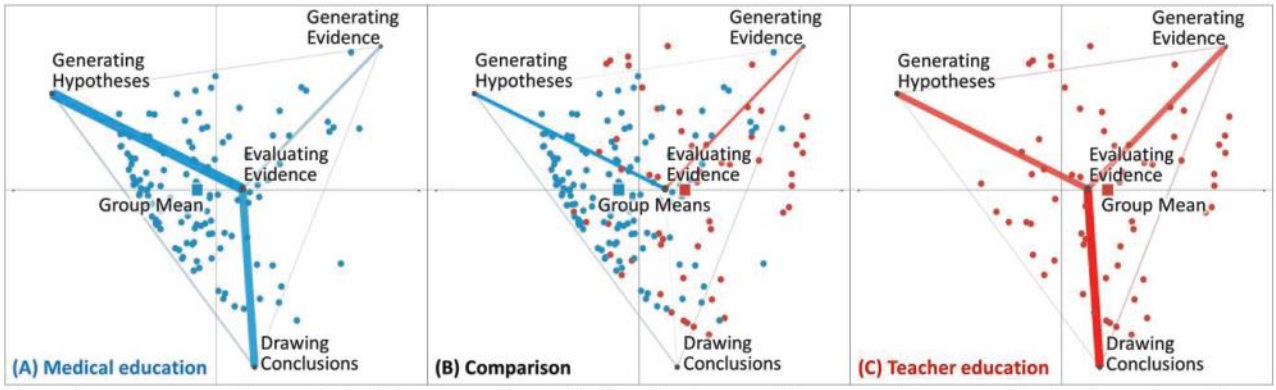
Since half of the sample of pre-service teachers were in their 1st to 4th semester, which was excluded in medical education, we defined a comparative subsample of 61 pre-service teachers in their 5ht to 13th semester (M = 7.38; SD = 2.30), who were on average M = 24.20 years old (SD = 3.55), and were mostly women (52 women; 8 men; 1 nonbinary). The comparative subsample of 61 pre-service teachers in the 5th or a higher semester accounted for 488 justificatory reports (average number of words per report M = 94.3; SD = 62.3).

5.1 Diagnostic activities in medical education and teacher education (RQ1)

Comparing medical education with the subsample of 5th or higher semester students from teacher education, there was no significant difference in the relative frequencies for *evaluating evidence* (medical education M = 60.96%; SD = 10.24%; teacher education M = 65.40%; SD = 18.00%; $t(77)$ = 1.81, p = .075, Cohen's d = 0.34). Concerning the other three diagnostic activities, the differences between the disciplines was significant: In medical education, the share for *generating hypotheses* was still about twice as high (M = 16.26%; SD = 7.96%) as in teacher education (M = 8.50%; SD = 5.50%), with a significant, large-sized effect ($t(161)$ = 8.00, p < .001, Cohen's d = 1.06). The share for *generating evidence* was still about twice as high in teacher education (M = 15.00%; SD = 15.80%) as in medical education (M = 6.79%; SD = 8.26%), with a significant medium-sized effect ($t(74)$ = 3.84, p < .001, Cohen's d = -0.74). In medical education, we still found a significantly higher share for *drawing conclusions* (M = 15.99%; SD = 6.39%) than in teacher education (M = 11.10%; SD = 7.15%), with a medium effect size ($t(201)$ = 4.82, p < .001, Cohen's d = 0.74).

5.2 Diagnostic practices in medical education and teacher education (RQ2)

In the supplementary Figure 3A and 3B, we compared students from medical education with the subsample of 5th or higher semester students from teacher education. The positioning of the group mean of learners from medical education (M = -.21, SD = .60, N = 142) was statistically significantly different from the positioning of the group mean of learners from teacher education in their 5th or a higher semester (M = .49, SD = .68, N = 61; $t(100.89)$ = 6.97, p < .01, Cohen's d = 1.13).



Supplementary Figure 3. ENA networks and distributions of learners from medical education **(A),** and teacher education **(C)**. The figures also contain group means (squares) across the learners within the two disciplines. The comparison graph **(B)** depicts both distributions and the differences between the other two networks.

The results of this subsample analysis support the previously described findings. Furthermore, in the supplementary Figure 3C, we additionally distinguished between students from teacher education who are in their 1st to 4th semester and students from teacher education who are in their 5th or a higher semester. As indicated by the positioning of the two group means displayed in the supplementary Figure 3C, the two subsamples of students from teacher education are overall rather similar.

This re-analysis, which compared students from medical education with a subsample of students from teacher education in their 5th or a higher semester of study, supported the findings from the initial analyses of the full sample, apart from the significant difference in evaluating evidence.

3

Study 2: Diagnostic Argumentation in Teacher Education: Making the Case for Justification, Disconfirmation, and Transparency

Reference: 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

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/feduc.2022.977631

OPEN ACCESS

EDITED BY
Robin Stark,
Saarland University,
Germany

REVIEWED BY
Timothy Fukawa-Connelly,
Temple University,
United States
Tom Rosman,
Leibniz-Institute
for Psychology (ZPID), Germany
Hendrik Lohse-Bossenz,
University of Education Heidelberg,
Germany

*CORRESPONDENCE
Elisabeth Bauer
elisabeth.bauer@psy.lmu.de

SPECIALTY SECTION
This article was submitted to
Teacher Education,
a section of the journal
Frontiers in Education

RECEIVED 24 June 2022
ACCEPTED 05 October 2022
PUBLISHED 03 November 2022

CITATION
Bauer E, Sailer M, Kiesewetter J,
Fischer MR and Fischer F (2022) Diagnostic
argumentation in teacher education:
Making the case for justification,
disconfirmation, and transparency.
Front. Educ. 7:977631.
doi: 10.3389/feduc.2022.977631

Diagnostic argumentation in teacher education: Making the case for justification, disconfirmation, and transparency

Elisabeth Bauer[1]*, Michael Sailer[1], Jan Kiesewetter[2], Martin R. Fischer[2] and Frank Fischer[1]

[1]Education and Educational Psychology, Department of Psychology, Ludwig-Maximilians-Universität in Munich, Munich, Germany, [2]Institute for Medical Education, University Hospital, Ludwig-Maximilians-Universität in Munich, Munich, Germany

Research on diagnosing in teacher education has primarily emphasized the accuracy of diagnostic judgments and has explained it in terms of factors such as diagnostic knowledge. However, approaches to scientific argumentation and information processing suggest differentiating between *diagnostic judgment* and *diagnostic argumentation*: When making accurate diagnostic judgments, the underlying reasoning can remain intuitive, whereas diagnostic argumentation requires controlled and explicable reasoning about a diagnostic problem to explain the reasoning in a comprehensible and persuasive manner. We suggest three facets of argumentation for conceptualizing diagnostic argumentation, which are yet to be addressed in teacher education research: *justification* of a diagnosis with evidence, *disconfirmation* of differential diagnoses, and *transparency* regarding the processes of evidence generation. Therefore, we explored whether preservice teachers' diagnostic argumentation and diagnostic judgment might represent different diagnostic skills. We also explored whether justification, disconfirmation, and transparency should be considered distinct subskills of preservice teachers' diagnostic argumentation. We reanalyzed data of 118 preservice teachers who learned about students' learning difficulties with simulated cases. For each student case, the preservice teachers had to indicate a diagnostic judgment and provide a diagnostic argumentation. We found that preservice teachers' diagnostic argumentation seldom involved all three facets, suggesting a need for more specific training. Moreover, the correlational results suggested that making accurate diagnostic judgments and formulating diagnostic argumentation may represent different diagnostic skills and that justification, disconfirmation, and transparency may be considered distinct subskills of diagnostic argumentation. The introduced concepts of justification, disconfirmation, and transparency may provide a starting point for developing standards in diagnostic argumentation in teacher education.

KEYWORDS

teacher education, diagnostic argumentation, scientific argumentation, diagnostic accuracy, diagnostic judgment, diagnostic knowledge

Introduction

Diagnostic skills are relevant in many fields, one of which is teacher education (Heitzmann et al., 2019). Teachers' diagnosing is a prototypical practice scenario for evidence-oriented practice, and as such, it is crucial for teachers' professionalism (Fischer, 2021). Previous research on teachers' diagnosing has primarily investigated diagnostic accuracy—i.e., the correctness of diagnostic judgments—because inaccurate judgments can easily disadvantage students by, for example, leading to unsuitable or insufficient educational interventions (e.g., Loibl et al., 2020; Urhahne and Wijnia, 2020; Kramer et al., 2021a). Besides making accurate diagnostic judgments, communicating diagnostic considerations is another vital aspect of diagnostic skills, for example, for purposes such as reporting diagnostic findings (Bauer et al., 2020) or collaborative diagnosing (Kiesewetter et al., 2017). However, thus far, there is no clear conceptualization of diagnostic argumentation, which we define as explaining a diagnostic judgment and the underlying reasoning comprehensibly and persuasively (see Walton, 1990; Berland and Reiser, 2009). It is also unclear whether professionals (e.g., teachers) who can make accurate diagnostic judgments are capable of offering sufficient diagnostic argumentation. This raises the question of whether accurate diagnostic judgment and diagnostic argumentation are fully based on the same knowledge—reflecting one overarching diagnostic skill—or whether they need to be considered different subskills of diagnosing. This differentiation might have implications for teaching diagnostic skills, such as the definition of learning objectives and the design and implementation of learning environments (see Grossman et al., 2009).

To our knowledge, no systematic research has differentiated between the concepts of diagnostic argumentation and diagnostic judgment. Therefore, we propose a conceptualization of diagnostic argumentation that consists of three facets: *justification* of a diagnosis with evidence, *disconfirmation* of differential diagnoses, and *transparency* regarding the processes of evidence generation. We explore diagnostic argumentation in terms of these three facets and investigate whether they indicate one joint underlying skill or different aspects of diagnostic skills by analyzing their interrelations with one another and with a potentially joint knowledge base. We also explore how justification, disconfirmation, and transparency in diagnostic argumentation are related to the accuracy of diagnostic judgments in the context of teacher education.

Diagnosing in teacher education

Teacher education is one of the fields in which learning diagnostic skills is an important matter of professionalization (Grossman, 2021). In particular, teachers have to diagnose students' performance, progress, and learning prerequisites (e.g., Praetorius et al., 2013; Südkamp et al., 2018). However, these aspects also include the initial identification of clinical problems, such as learning difficulties (e.g., dyslexia) and behavioral disorders (e.g., attention deficit hyperactivity disorder, i.e., ADHD; e.g., Poznanski et al., 2021). In all these contexts, we broadly define *diagnosing* as a "goal-oriented collection and interpretation of case-specific or problem-specific information to reduce uncertainty in order to make […] educational decisions" (Heitzmann et al., 2019, p. 4). Other associated terms are used for diagnosing in teacher education as well, such as *assessment* (e.g., Herppich et al., 2018). As part of teachers' professional activities, diagnosing is crucially related to the discussion around teachers' evidence-oriented practice (Stark, 2017) and is possibly a prototypical practice scenario (Fischer, 2021). Teachers are expected to use knowledge on theories, methods, procedures, and findings from educational research (e.g., Kiemer and Kollar, 2021) to reflect their experiences, possibly overcome dysfunctional intuitive approaches and—at least partially—guide their diagnostic activities and interventions. Teacher education programs are increasingly acknowledging the relevance of facilitating diagnostic skills, and research in teacher education has also addressed the issue of how diagnostic skills are learned (e.g., Chernikova et al., 2020; Loibl et al., 2020; Sailer et al., 2022).

Teachers' diagnostic judgments

Previous research on teachers' diagnosing has focused on how teachers make diagnostic judgments (e.g., Loibl et al., 2020; Urhahne and Wijnia, 2020; Kramer et al., 2021a). Loibl et al. (2020) suggested distinguishing between the processes and products of teachers' diagnostic judgments In terms of product indicators, research on teachers' diagnostic judgments has focused on diagnostic accuracy—i.e., the correctness of diagnostic judgments—because inaccurate judgments can lead to unsuitable or insufficient educational interventions that easily disadvantage students (e.g., Urhahne and Wijnia, 2020). There is also an increasing amount of research investigating teachers' judgment processes, for example, in terms of diagnostic activities such as generating hypotheses, generating and evaluating evidence, and drawing conclusions (e.g., Wildgans-Lang et al., 2020; Codreanu et al., 2021; Kramer et al., 2021a). In addition, research has begun to focus more on the role of information processing in teachers' judgment processes (e.g., Loibl et al., 2020). Teachers' diagnostic judgment processes can involve intuitive information processing—i.e., fast recognition of patterns of information—which facilitates flexible and adaptive acting in the classroom; teachers can also engage in controlled information processing when spending time and effort on consciously evaluating evidence and its causal relations (Kahneman, 2003; Evans, 2008). Teachers' information processing in making diagnostic judgments depends on situational characteristics (Loibl et al., 2020), such as the available time for making a judgment (Rieu et al., 2022), the consistency and conclusiveness of the available evidence, and teachers' perceptions of their situational accountability (Pit-ten

Cate et al., 2020). In classrooms with multiple students, teachers often need to make intuitive judgments, prioritize tasks, and decide where to invest their time and cognitive resources (Feldon, 2007; Vanlommel et al., 2017). With respect to achieving diagnostic accuracy, research suggests regarding judgment processes (e.g., in terms of information processing) as processes that interact with teachers' characteristics, especially their diagnostic knowledge (e.g., Loibl et al., 2020; Kramer et al., 2021a).

The role of diagnostic knowledge

Diagnostic knowledge is generally considered an important basis of diagnostic skills (Heitzmann et al., 2019). Having a sufficient base of specific diagnostic knowledge seems to be a necessary condition for achieving accurate diagnostic judgments (Kolovou et al., 2021). In addition, advanced diagnosticians' well-organized knowledge structures enable them to recognize patterns of critical case information correctly, without necessarily conducting a controlled analysis of the underlying causal relations (see Kahneman, 2003; Evans, 2008; Boshuizen et al., 2020). Research has suggested that performing complex cognitive tasks requires not only knowledge about relevant concepts but also knowledge about *how* to systematically approach the task (e.g., Van Gog et al., 2004). In the context of teacher education, Shulman (1986) suggested that, besides domain-specific content, distinguishing between different types of knowledge—such as conceptual and strategic knowledge—is relevant to capturing different functionalities of knowledge, such as acting adaptively in response to various problems and situations. In the course of developing strategic knowledge, basic aspects of conceptual knowledge are abstracted and integrated with episodic knowledge into cognitive scripts about approaching certain problems or situations (e.g., Shulman, 1986; Schmidmaier et al., 2013; Boshuizen et al., 2020). This means that conceptual and strategic knowledge about the same specific content are likely related but address different aspects of solving a task. Conceptual and strategic knowledge have been adapted and empirically investigated in the context of diagnosing in medical education (e.g., Stark et al., 2011; Schmidmaier et al., 2013): *conceptual diagnostic knowledge* (CDK) consists of concepts, such as diagnoses and their relations with each other and with evidence, whereas *strategic diagnostic knowledge* (SDK) refers to how to proceed in diagnosing a specific problem (i.e., how to reject or confirm differential diagnoses and which informational sources provide critical evidence for doing so). Researchers addressing diagnosing in teacher education have also suggested distinguishing between CDK and SDK (e.g., Förtsch et al., 2018). Therefore, CDK and SDK seem crucial for correctly processing relevant case information and making accurate diagnostic judgments.

Diagnostic argumentation

Beyond making accurate diagnostic judgments, there are instances in which teachers or other diagnosticians need to explain their reasoning and the resulting diagnostic judgment in a comprehensible and persuasive manner, which we suggest to designate as *diagnostic argumentation* (see Walton, 1990; Berland and Reiser, 2009). Diagnostic argumentation is required in situations in which explanations are directed toward a recipient, such as a collaborating teacher or school psychologist (e.g., Kiesewetter et al., 2017; Csanadi et al., 2020; Radkowitsch et al., 2021). The context of identifying students' clinical problems is one example in which diagnostic argumentation is particularly relevant for teachers, as in many educational systems, final judgments about clinical diagnoses are made by clinical professionals (e.g., school psychologists), with whom teachers might need to collaborate (Albritton et al., 2021). However, also in other contexts, diagnostic argumentation facilitates a collaborative process of considering and reconciling competing explanations and thus, if necessary, can help improve the diagnosing (see Berland and Reiser, 2009; Csanadi et al., 2020). There are also nonimmediate dialogical situations (see Walton, 1990), such as writing a report about diagnostic findings (Bauer et al., 2020), in which information may need to be comprehensible and persuasive to potential recipients at a later point in time.

Especially when engaging in a face-to-face critical exchange of arguments in collaborative or otherwise dialogical diagnosing, teachers might involve in argumentation processes and a controlled analysis of the available evidence and potential explanations before making a diagnostic judgment. Collaborative generation and evaluation of evidence and a critical evaluation of others' arguments can improve the quality of argumentative outcomes (Mercier and Sperber, 2017; Csanadi et al., 2020). In other contexts, teachers might make intuitive judgments without a controlled analysis of all the available evidence and causal relations. If the information processing for a diagnostic judgment mainly involves intuitive pattern recognition, parts of the reasoning can remain implicit (Evans, 2008). However, comprehensively explaining a judgment and its underlying reasoning initially requires that the reasoning be explicable or at least constructible in retrospect. In terms of nondialogical situations, such as writing reports, initial evidence suggests that compared to medical education, there seems to be a lower standardization in teacher education (Bauer et al., 2020), which could facilitate constructing persuasive explanations in retrospect. For these reasons, it might not necessarily be a given that teachers who make accurate judgments in nondialogical diagnostic situations are capable of subsequently providing comprehensible and persuasive explanations of their reasoning. This open question has yet to be explored by research.

Justification, disconfirmation, and transparency in diagnostic argumentation

To explore how diagnostic judgment and diagnostic argumentation are related, it is first necessary to define what kind

of information is expected to be provided in the context of diagnostic argumentation. We argue that besides providing comprehensible explanations, diagnostic argumentation also aims to persuade potential recipients of the presented reasoning (Berland and Reiser, 2009) and, thus, requires providing information that enables a recipient's understanding and evaluation of the efforts made during diagnosing (see Chinn and Duncan, 2018). Therefore, to further define the concept of diagnostic argumentation, we suggest three facets that might facilitate recipients' understanding of the presented reasoning: justification, disconfirmation, and transparency. We propose that these three facets of diagnostic argumentation resemble approaches in scientific argumentation (see Sampson and Clark, 2008; Mercier and Heintz, 2014), namely justifying one's reasoning with evidence (e.g., Toulmin, 1958), considering and disconfirming alternative explanations (e.g., Lawson, 2003), and emphasizing the credibility of informational sources with methodological transparency (e.g., Chinn et al., 2014). In what follows, we explain the three facets in further detail.

Justification denotes the provision of evidence in support of a claim (e.g., Toulmin, 1958; Hitchcock, 2005), which allows recipients to raise potential issues about the reasoning that was presented. In the context of diagnostic argumentation, diagnostic judgments are claims that need to be justified by providing evidence derived from the case information. Therefore, justifications evaluate relevant case information as evidence from which to draw conclusions concerning a judgment (see Fischer et al., 2014).

Disconfirmation emphasizes discussing differential diagnoses that may have been hypothesized when diagnosing a given case. As a process of uncertainty reduction (Heitzmann et al., 2019), diagnosing involves generating and evaluating different hypotheses (Klahr and Dunbar, 1988; Fischer et al., 2014) that resemble competing claims in argumentation. Similar to the scientific approach of disconfirmation (e.g., Gorman et al., 1984), a rebuttal of competing claims supports the persuasiveness of the final claim (e.g., Toulmin, 1958; Lawson, 2003). In diagnostic argumentation, differential diagnoses are competing claims that should be explicated and discussed to facilitate the persuasiveness of the final judgment by demonstrating that alternative explanations have been considered. Recipients can build on this information to evaluate and criticize whether relevant differential diagnoses have been missed or mistakenly rejected.

Transparency regarding the processes of evidence generation provides information about the reliability of the methodology for generating evidence from informational sources (Chinn et al., 2014; Fischer et al., 2014). In diagnostic argumentation, transparency is achieved by describing the processes underlying evidence generation, thus allowing recipients to evaluate the presented evidence and diagnostic conclusions. Explicating how evidence was generated facilitates a recipient's understanding and ability to criticize the quality of the evidence and, ultimately, the validity of the conclusions (Vazire, 2017).

Analogously to approaches involved in scientific argumentation (see Sampson and Clark, 2008; Mercier and Heintz, 2014), we suggest that justification, disconfirmation, and transparency in diagnostic argumentation facilitate a recipient's understanding and evaluation of the efforts made during diagnosing. We are unaware of any research in teacher education that has conceptualized or investigated a skill similar to what we have defined as diagnostic argumentation, including the facets of justification, disconfirmation, and transparency. Therefore, in this study, we aimed to explore the interrelations between justification, disconfirmation, and transparency in diagnostic argumentation, as well as their relations with making accurate diagnostic judgments, and the explanatory roles of CDK and SDK (see Figure 1).

Research questions

We propose that justification, disconfirmation, and transparency are three relevant facets of diagnostic argumentation and that diagnostic argumentation and diagnostic judgment might represent two distinct diagnostic skills that may, however, both be partially explained by CDK and SDK. Understanding the interrelations between these skills and knowledge might provide relevant information for teacher educators and the field of teacher education.

In investigating the proposed concept of diagnostic argumentation, it is also important to explore whether justification, disconfirmation, and transparency might represent distinct subskills or indicators of one joint underlying diagnostic skill (RQ1). To approach this question, we investigated how the individual facets (1a) and different combinations of the facets (1b) occur within preservice teachers' diagnostic argumentation and analyzed the facets' relations (1c) in preservice teachers' diagnostic argumentation. We assumed that finding close relationships would indicate a joint basis of knowledge and skills; by contrast, small relationships or a lack thereof would indicate that the three facets represent different subskills of diagnostic argumentation.

In terms of distinguishing between the three facets as different subskills, a related question is to what extent justification, disconfirmation, and transparency are based on conceptual diagnostic knowledge and strategic diagnostic knowledge (RQ2). Because CDK and SDK are thought to be a major basis for the reasoning presented in diagnostic argumentation (Heitzmann et al., 2019), we assumed that they also partially explain justification, disconfirmation, and transparency; that is, CDK and SDK may be needed to generate evidence from informational sources (explicated in transparency) and to make a warranted connection between the evidence and a diagnosis (explicated in justification) or several differential diagnoses (explicated in disconfirmation). Exploring the degree to which CDK and SDK explain justification, disconfirmation, and transparency in diagnostic argumentation can provide an initial basis for future research on teachers' prerequisites for diagnostic argumentation.

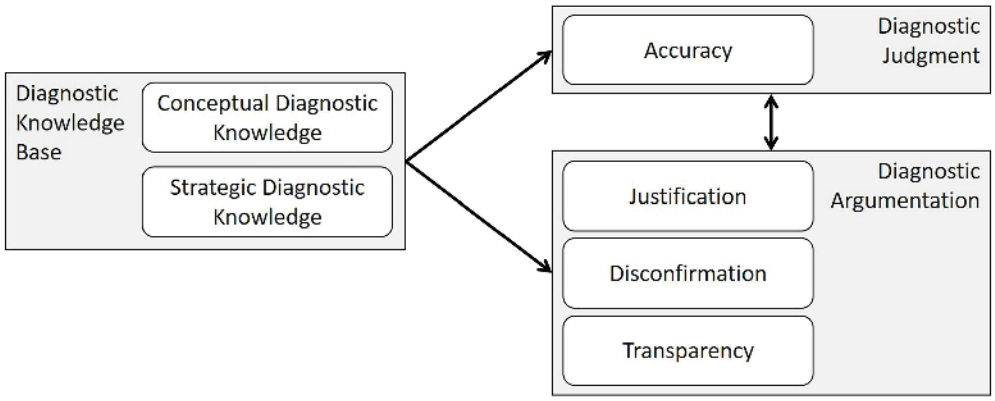


FIGURE 1
Theoretical model of the potential relationships between diagnostic knowledge and the skills of diagnostic judgment and diagnostic argumentation.

Given that diagnostic argumentation additionally aims to be persuasive instead of solely verbalizing the reasoning made while processing information, further knowledge and skills beyond CDK and SDK may contribute to justification, disconfirmation, and transparency in diagnostic argumentation.

For the same reason, we assumed that, despite a presumably joint basis of CDK and SDK, diagnostic accuracy might not necessarily be related to justification, disconfirmation, and transparency. Therefore, we explored whether diagnostic judgment (indicated by diagnostic accuracy) and diagnostic argumentation (indicated by justification, disconfirmation, and transparency) might represent different diagnostic skills (RQ3). In doing so, we assumed that identifying close relationships would indicate a joint underlying diagnostic skill; by contrast, small relationships or a lack thereof would indicate that diagnostic argumentation and diagnostic judgment might represent different diagnostic skills.

Materials and methods

Participants

In this study, we reanalyzed data that were originally collected to train an AI-based adaptive feedback component for a simulation-based learning environment (see Pfeiffer et al., 2019). A total of 118 preservice teachers participated in the data collection and processed simulated cases pertaining to students' clinical problems. Participants were $M = 22.96$ years old ($SD = 4.10$), the majority were women (102 women, 15 men, and 1 nonbinary), and they were in their first to 13th semester ($M = 4.62$, $SD = 3.40$) of a teacher education program. We recruited preservice teachers in all semesters because relevant courses about students' clinical problems were not compulsory or bound to a specific semester but could be taken in any semester. Participants subjectively rated their prior knowledge of students' clinical problems prior to receiving any instruction about the content of the study. On average, they indicated a medium rating of their own prior knowledge (on a rating scale ranging from 1 to 5 points: prior knowledge about ADHD, $M = 2.78$, $SD = 0.81$; prior knowledge about dyslexia, $M = 2.47$, $SD = 0.76$). We assumed that this sample mirrors the diverse population of preservice teachers.

Research design

We chose a quantitative and correlational research design to determine the relationships between the following variables: justification, disconfirmation, and transparency in diagnostic argumentation; CDK and SDK; and the accuracy of diagnostic judgment.

Simulation and tasks

We asked participants to take on the role of a teacher and process eight cases of primary and secondary students with performance-related or behavioral problems that might or might not indicate a clinical diagnosis in the range of ADHD or dyslexia. Two independent domain experts, one school psychologist and one psychotherapist for children and adolescents, validated the case materials before they were implemented in CASUS, a case-based online learning environment.[1] Participants solved the cases consecutively. The cases included several informational sources, such as samples of the students' written exercises and school certificates, reports of observations from inside and outside the classroom, and conversations with the respective students, their parents, and other teachers (the German-language case materials can be accessed at https://osf.io/hn7wm/). Participants could freely

1 http://www.casus.net/

choose how many and which informational sources to consult and in which order they wanted to do so (see Figure 2).

One example is the case of a secondary school student named Anna, who is showing symptoms of attention-deficit disorder (ADD). An initial problem statement describes Anna as a fifth-grade student, 11 years old, who constantly needs to be pushed to finish her tasks and who has poor grades in many subjects, especially the core subjects, such as math and the language subjects. The learners could examine written observations of Anna's in-class and out-of-class behavior, read recordings of conversations with Anna or with her parents and several teachers, or look at Anna's last annual report and an example of a written exercise. Her behavior is described as very calm and distracted. She reads very slowly, and it is difficult for her to answer questions about a text that she has just read. She often fails to follow the exact instructions for tasks or fails to complete them fully. Moreover, she often does not bring all the required school supplies or arrives late in the morning. At a parent–teacher conference, Anna's mother backs up the impression of disorganized and slow learning behavior when talking about Anna's homework. Anna's last annual report and the conversations with the other teachers show that her grades are also affected by her inattentiveness, except artistic subjects and gym class. She mostly interacts with one friend and tends to remain distant from the other students. Anna herself points out that it is hard for her to concentrate because she feels easily distracted. However, at home, where there are fewer ambient noises, she can focus on and enjoy reading, drawing, and painting. Overall, the case information is designed in such a way that the diagnosis of ADD is the most likely diagnosis, despite the fact that several differential diagnoses may be relevant. The other cases included the same kinds of informational sources as Anna's case.

To complete a case and move on to the next case, participants had to complete two tasks. First, they had to make a diagnostic judgment, answering the question of whether the simulated student has issues that warrant further diagnosing of a clinical problem and, if so, which diagnosis may apply. Second, we asked participants to write an argumentation text about their conclusions and their reasoning about the case. For the purpose of this study, participants received no further guidance or support regarding how to write their diagnostic argumentation.

Procedure

The data were collected on computers in a laboratory setting, with three to 20 participants simultaneously joining the study. They worked individually at separate desks and were not permitted to speak to each other. We introduced the participants to the aims and procedure of the study and familiarized them with the learning environment. After giving informed consent to participate in the study, participants received randomly assigned codes to log on to the CASUS learning environment to anonymize the data. When entering the online learning environment, participants first received a 25 min theoretical input concerning the topic of diagnosing in general and the diagnosing of ADHD and dyslexia in particular to activate existing knowledge and ensure the minimum amount of knowledge required for solving the cases. Participants were not

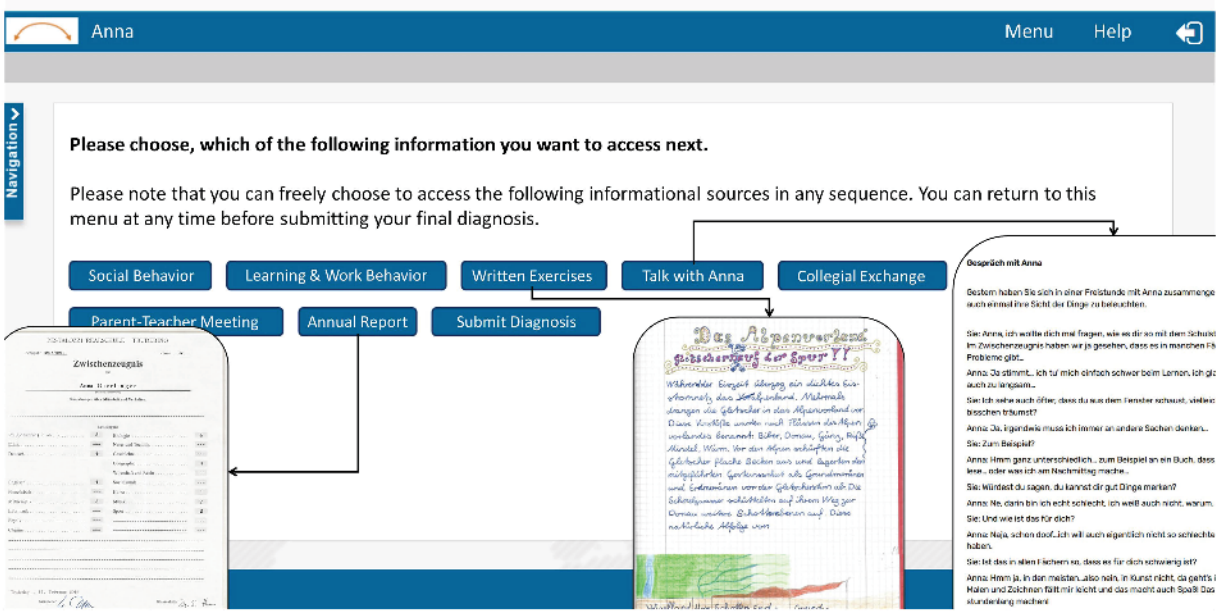


FIGURE 2
Example of the case materials in the simulation-based learning environment.

allowed to take any notes or go back to the input part at a later point to avoid biases in subsequent testing and learning. Following the theoretical input, participants spent around 25 min on a pretest that assessed their CDK and SDK. Subsequently, participants entered the learning phase consisting of the eight simulated cases, with a break of 10 min after four cases. They had to finish one case at a time to gain access to the next case. All participants received the cases in the same sequence. The time on task for all cases was around 1 h. Subsequently, participants spent around 25 min on a posttest. Generally, participants were allowed to work at their own pace. Overall, participants spent around 3 h from login to logout. During the study, researchers were available to help with technical issues or questions about navigation but did not answer any content-related questions. Participants received monetary compensation of 35 euros.

Data sources and measurements

The data sources used for the presented analyses are the CDK and SDK scores from the pretest as well as the written diagnostic judgments and diagnostic argumentation texts from six of the eight cases. We decided to exclude two cases from the analysis because their case information turned out to be more ambiguous and inconclusive compared to the other cases.

Diagnostic knowledge

Conceptual diagnostic knowledge

CDK was assessed in the pretest after participants received the theoretical input. We used 14 single-choice items about diagnosing ADHD and dyslexia with four answer options each (one correct answer and three distractors). The CDK questionnaire was developed prior to the study to assess participants' CDK, which was considered relevant for processing the simulated cases. Two independent domain experts, one school psychologist and one psychotherapist for children and adolescents, validated the CDK questionnaire. One example item is "Which of the following is *not* one of the cardinal symptoms of ADHD?" with the answer options (a) Inattentiveness, (b) Hyperactivity, (c) Impulsivity, and (d) Impatience. Participants received one point per correct answer. The points were aggregated into a total score, ranging from 0 to 14 points, for CDK.

As suggested by Stadler et al. (2021; see also Diamantopoulos and Siguaw, 2006; Taber, 2018), we calculated variance inflation factors (VIFs) for all items to avoid having redundant items representing the formative knowledge construct. The maximum VIF was $VIF_{max} = 1.30$, which is well below the recommended cut-off of 3.3.

Strategic diagnostic knowledge

Subsequent to assessing CDK, we measured SDK using four key-feature cases (two key-feature cases about ADHD and two about dyslexia) with two multiple-choice questions each (see Page et al., 1995). Key-feature cases present a brief description consisting of a few sentences before asking about the strategic approaches used to diagnose the case. The key-feature cases were developed prior to the study to assess participants' SDK, which was considered relevant for processing the simulated cases. Two independent domain experts, one school psychologist and one psychotherapist for children and adolescents, validated the key-feature cases. One example key-feature case introduced the fourth grader Luis, who has always been a rather poor reader but has begun to fall farther behind his classmates over the last few months and just recently again received the lowest grade in the class on a reading test. He cannot summarize the contents of a short text even immediately after reading it and can only read aloud very slowly. Apart from his performance issues, he has a chronic disease due to which he cannot regularly attend school for stretches of several weeks. After reading this brief case description, two multiple-choice questions were asked.

The first of the two multiple-choice questions per key-feature case asked participants to choose all relevant differential diagnoses out of a list of clinical as well as non-clinical differential diagnoses (one to three correct options out of seven to nine answer options). Participants received points for correctly choosing relevant options and not choosing irrelevant options. We calculated one mean score across all options per key-feature case, resulting in a diagnosis score of 0 to 1 for the first question for each key-feature case.

The second of the two multiple-choice questions per key-feature case asked participants to choose from a list of further approaches and resources relevant to confirm or disconfirm a given set of differential diagnoses (three to six correct options out of seven to 10 answer options). Participants received points for correctly choosing relevant options and not choosing irrelevant options. We calculated one mean score across all options per key-feature case, resulting in a resource score of 0 to 1 for the second question for each key-feature case.

The four diagnosis scores and four resources scores were accumulated into a total score of 0 to 8 points for SDK on the pretest. There were no redundant items ($VIF_{max} = 1.09$).

Accuracy of diagnostic judgment

To measure diagnostic accuracy, we coded all the written diagnoses as accurate (1 point), partially accurate (0.5 points), or inaccurate (0 points). We coded written diagnoses as accurate if indicating a diagnosis that was considered the correct solution when designing the cases (e.g., ADD for the case Anna). The written diagnoses were coded as partially accurate if correctly indicating the higher-level class of diagnoses for the accurate diagnosis (e.g., if the correct diagnosis was ADD and the participants indicated ADHD). A total of 12.5% of the diagnoses were double-coded, resulting in an interrater reliability (IRR) of Cohen's $\kappa = 0.80$ (Cohen, 1960). The internal consistency across the six cases was

McDonald's ω = 0.37 (McDonald, 1999). For further analyses, we calculated a total score from the points achieved for diagnostic accuracy with a possible range of 0 to 6 points.

Justification, disconfirmation, and transparency in diagnostic argumentation

We operationalized justification, disconfirmation, and transparency based on a coding of the six cases' diagnostic argumentation texts.

Justification

We operationalized the presence or absence of justification in diagnostic argumentation as *evaluating evidence* co-occurring with *drawing conclusions* within the temporal context of two sentences, resulting in 1 or 0 points per diagnostic argumentation. In this study, we reanalyzed data that were originally used to train an AI-based adaptive feedback algorithm for a simulation-based learning environment (see Pfeiffer et al., 2019). Four expert raters coded the diagnostic argumentation texts segmented by sentences regarding the categories *evaluating evidence* and *drawing conclusions*. They initially read the complete diagnostic argumentation before coding *evaluating evidence* and *drawing conclusions* for the individual sentences. *Evaluating evidence* was defined as explicitly presenting or interpreting case information (e.g., "Markus behaves aggressively and gets offended very easily"). *Drawing conclusions* was defined as explicitly accepting or rejecting at least one diagnosis (e.g., "I think most likely the diagnosis is ADHD"). The raters simultaneously coded 15% of the data before dividing the rest of the data because of substantial agreement (IRRs: Fleiss' κ = 0.71 for *drawing conclusions*; Fleiss' κ = 0.75 for *evaluating evidence*; Fleiss, 1971; Landis and Koch, 1977). The internal consistency across six cases was sufficient (McDonald's ω = 0.60; McDonald, 1999). We calculated a total justification score for each participant, with a possible range of 0 to 6 points.

Disconfirmation

We operationalized disconfirmation as present if two or more *differential diagnoses* were addressed, resulting in 1 or 0 points per diagnostic argumentation. This round of coding was done separately from the coding of justification and transparency for the purpose of our reanalysis. Two expert raters coded the diagnostic argumentation texts of six cases regarding a set of *differential diagnoses*. The coding scheme consisted of 27 differential diagnoses, which included non-clinical (e.g., insufficient schooling, emotional stress, and problematic home environment) and clinical differential diagnoses (e.g., ADHD, ADD, dyslexia, and autism). The raters considered the facet of disconfirmation as being included in the diagnostic argumentation if two or more of these differential diagnoses were discussed in one diagnostic argumentation, independent of which diagnosis the participant indicated as the final diagnosis. The raters simultaneously coded 15% of the data before dividing the rest of the data (overall IRR: Cohen's κ = 0.92; Cohen, 1960). The internal consistency was sufficient (McDonald's ω = 0.60; McDonald, 1999). We calculated a total disconfirmation score for each participant, with a possible range of 0 to 6 points.

Transparency

We operationalized transparency in diagnostic argumentation as at least one explication of *generating evidence*, resulting in 1 or 0 points per diagnostic argumentation. The coding for transparency was done in in the same round as the coding for justification. Four expert raters coded the diagnostic argumentation texts regarding *generating evidence*, which was defined as an explicit description of accessing informational sources (i.e., tests or observations; e.g., "I observed Anna's school-related behavior and achievement"). The raters simultaneously coded 15% of the data before dividing the rest of the data because of substantial agreement (IRR: Fleiss' κ = 0.70; Landis and Koch, 1977). The internal consistency was sufficient (McDonald's ω = 0.71; McDonald, 1999). We calculated a total transparency score for each participant, with a possible range of 0 to 6 points.

Statistical analyses

For RQ1, we explored the descriptive statistics of justification, disconfirmation, and transparency in preservice teachers' diagnostic argumentation texts in terms of both individual facets (1a) and facet combinations (1b). We considered facet combinations as types of argumentation texts and depicted them in relation to the individual facets using Epistemic Network Analysis (ENA; Shaffer, 2017). The ENA algorithm analyzes and accumulates co-occurrences of elements in coded data, such as the three facets of argumentation within individual argumentation texts, to create a multidimensional network model, which is depicted as a dynamic network graph. To determine the types of argumentation texts, we grouped the argumentation texts according to the presence or absence of each argumentation facet in each argumentation text. The ENA algorithm then accumulated co-occurrences of the three facets across the argumentation texts to create a network model. We depicted this model as a two-dimensional network graph that showed the relative location of the argumentation types within the resulting two-dimensional space. We used the ENA online tool to create the network graphs.[2] In addition to the descriptive analyses, we calculated Pearson correlations with participants' overall justification, disconfirmation, and transparency scores (1c). To investigate RQ2, we calculated a multivariate multiple linear regression with the predictors CDK and SDK and the dependent variables justification, disconfirmation, and transparency. For RQ3, we first created two separate ENA networks by grouping the diagnostic argumentation texts that addressed either accurate or inaccurate

2 https://www.epistemicnetwork.org/

TABLE 1 Prevalence of the individual facets justification, disconfirmation, and transparency in the 709 diagnostic argumentation texts.

	Number of argumentation texts including the facet	Number of argumentation texts missing the facet
Justification	468 (66%)	241 (34%)
Disconfirmation	183 (26%)	526 (74%)
Transparency	327 (46%)	382 (54%)

TABLE 2 Prevalence of the argumentation types, indicated by combinations of the facets Justification (J), Disconfirmation (D), and Transparency (T), in the 709 argumentation texts.

Argumentation types, indicated by combinations of the three facets	Number of facets included	Number of argumentation texts	Percent of argumentation texts
JDT	3	83	11.7%
JDO	2	90	12.7%
JOT	2	129	18.2%
ODT	2	7	1.0%
JOO	1	166	23.4%
ODO	1	3	0.4%
OOT	1	108	15.2%
OOO	0	123	17.3%

diagnostic judgments; we tested the difference between the group means' locations in the network space using a *t*-test. To facilitate the statistical testing of the groups' network differences, we used the option of means rotation, which aligns the two group means on the X-axis of the network, thus, depicting systematic variance in only one dimension in the two-dimensional space (Shaffer, 2017). Moreover, we again calculated Pearson correlations, including the participants' overall scores for diagnostic accuracy, justification, disconfirmation, and transparency. We also explored partial correlations, controlling for CDK and SDK. For RQ1c and RQ3, including multiple comparisons (three Pearson correlations each), the significance level was Bonferroni-adjusted to $\alpha = 0.0167$ ($\alpha = 0.05/3$). For the other analyses, the significance level was set to $\alpha = 0.05$.

Results

RQ1: Justification, disconfirmation, and transparency

To investigate whether justification, disconfirmation, and transparency represent distinct subskills or one joint underlying diagnostic skill (RQ1), we analyzed the prevalence of the individual facets (1a) and the combinations of the facets (1b) in preservice teachers' individual argumentation texts. Moreover, we analyzed the relationships between justification, disconfirmation, and transparency in preservice teachers' diagnostic argumentation (1c). We considered findings of close relations to indicate a joint basis of knowledge and skills, and small or no relations to indicate that the three facets represent different aspects of diagnostic skills.

RQ1a: Prevalence of the facets in preservice teachers' argumentation texts

Analyzing the descriptive statistics of the prevalence of justification, disconfirmation, and transparency in preservice teachers' individual argumentation texts, we found that *justification* was the most common of the three facets in all diagnostic argumentation texts (see Table 1): Participants explicitly stated conclusions and justified them by evaluating evidence alongside the conclusion in 66% ($M = 0.66$; $SD = 0.47$) of all argumentation texts. *Disconfirmation* was found in 26% ($M = 0.26$; $SD = 0.44$) of all diagnostic argumentation texts, indicating that the majority of diagnostic argumentation texts did not involve differential diagnoses but tended to focus on one final diagnosis. Moreover, we found *transparency* concerning the processes of evidence generation in 46% ($M = 0.46$; $SD = 0.50$) of all argumentation texts, indicating that approximately half of the diagnostic argumentation texts explained the processes of evidence generation.

RQ1b: Combinations of the facets in preservice teachers' argumentation texts

Descriptive statistics of the combinations of justification, disconfirmation, and transparency are outlined in Table 2. The combinations of the three facets can be considered different types of diagnostic argumentation texts, which we distinguished using the following abbreviations: *J* indicates the presence of justification, *D* indicates the presence of disconfirmation, *T* indicates the presence of transparency, and *O* indicates the absence of a facet (e.g., *JOT* indicates justification and transparency without disconfirmation; see Table 2 for all argumentation types and their prevalence). A

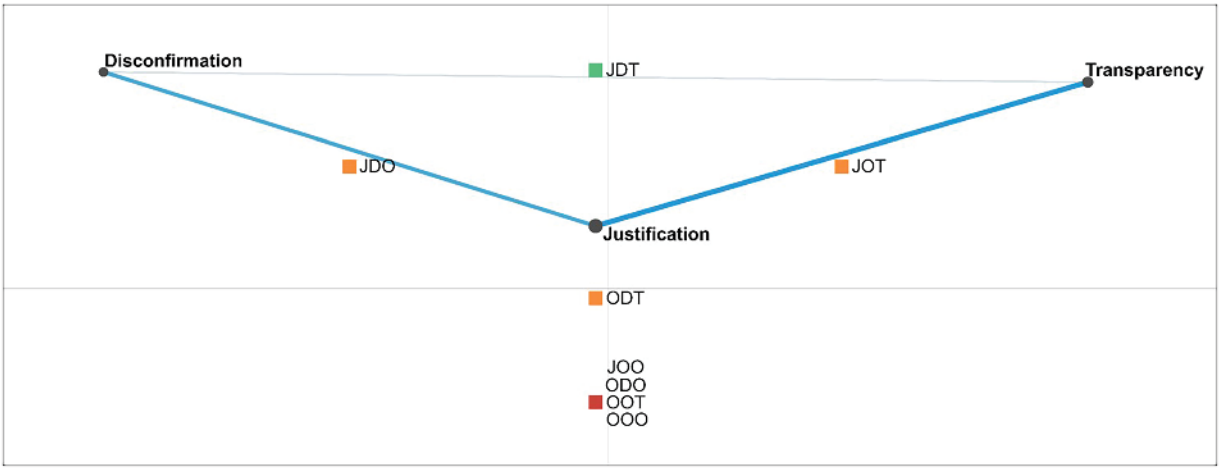


FIGURE 3
Argumentation types (indicated by colored squares) plotted in a two-dimensional space to indicate the relationship between the argumentation types and the individual facets (indicated by gray nodes) and the co-occurrences of the facets (indicated by blue lines). argumentation types are characterized by: J, Justification; D, Disconfirmation; T, Transparency; O, Absence of a Facet. For example: JOT, Justification and transparency without disconfirmation.

notable pattern was that argumentation texts addressing more than one diagnosis usually discussed the different diagnoses by evaluating evidence to make and justify conclusions (*JDT* and *JDO*), whereas hardly any argumentation texts addressed differential diagnoses without making and justifying related conclusions (*ODT* and *ODO*). However, diagnostic argumentation texts frequently presented a confirmatory justification of a single diagnosis without discussing alternative explanations (*JOT* and *JOO*). Consequently, including disconfirmation in diagnostic argumentation was dependent on including justification, but justification in diagnostic argumentation was not dependent on including the facet of disconfirmation, suggesting a relationship of unidirectional dependency.

To illustrate the types of argumentation texts and their relationships with the individual facets, we used ENA to plot both the argumentation types (indicated by colored squares) and the individual facets (indicated by gray nodes) in a two-dimensional space (see Figure 3). The two-dimensional space was built based on the co-occurrences of two argumentation facets each, which are indicated by the blue lines. The thickness of the blue lines represents the relative frequency of the co-occurrences (e.g., the thick line between justification and transparency relates to the 212 co-occurrences of justification and transparency in *JDT* and *JOT*). The positioning of argumentation types (indicated by the colored squares) along the X-axis is relative to the facets' co-occurrences, which is why *JOT* is located toward the right-sided node of transparency and *JDO* is located toward the left-sided node of disconfirmation. The central positioning of justification is due to its high overall prevalence (see Table 1). The positioning of argumentation types along the Y-axis indicates the argumentation texts' comprehensiveness regarding the three facets, with the extremes of *JDT* (all facets are present) and *OOO* (all facets are missing).

Overall, the findings indicate that preservice teachers tend to primarily provide justification in their diagnostic argumentation as an antecedent to including disconfirmation, transparency, or both. Moreover, the results suggest that there may be a relationship of unidirectional dependency of disconfirmation on justification.

RQ1c: Relations of justification, disconfirmation, and transparency

Beyond exploring the three facets in the individual argumentation texts, we also analyzed the descriptive statistics and correlations of preservice teachers' justification, disconfirmation, and transparency across the cases. The descriptive results of the facets' total scores (see Table 3) were consistent with the pattern found in the individual argumentation texts (see Table 1). Participants mostly focused on *justification* (M = 3.83, SD = 1.58), rarely used *disconfirmation* (M = 1.53, SD = 1.41), and put a medium emphasis on *transparency* (M = 2.67, SD = 1.81). The correlational analysis (see Table 3) indicated that justification and disconfirmation were significantly correlated, with a large effect ($r = 0.568$, $p < 0.001$). By contrast, transparency was not significantly correlated with justification ($r = 0.055$, $p = 0.554$) or disconfirmation ($r = 0.025$, $p = 0.787$). Considering the unidirectional dependency of disconfirmation on justification (see the results of RQ1b), we interpreted the overall result pattern as suggesting that justification, disconfirmation, and transparency are distinct facets of diagnostic argumentation rather than indicators of a uniform skill.

TABLE 3 Descriptive results and Pearson correlations of preservice teachers' scores for the three argumentation facets justification, disconfirmation, and transparency, as well as conceptual and strategic diagnostic knowledge and diagnostic accuracy.

		M	SD	1.	2.	3.	4.	5.
1.	Justification	3.83	1.58					
2.	Disconfirmation	1.53	1.41	$r = 0.568, p = 0.000$				
3.	Transparency	2.67	1.81	$r = 0.055, p = 0.554$	$r = 0.025, p = 0.787$			
4.	Conceptual diagnostic knowledge	8.86	1.66	$r = 0.265, p = 0.004$	$r = 0.234, p = 0.011$	$r = 0.041, p = 0.659$		
5.	Strategic diagnostic knowledge	6.70	0.39	$r = 0.252, p = 0.006$	$r = 0.042, p = 0.652$	$r = 0.194, p = 0.035$	$r = 0.130, p = 0.161$	
6.	Diagnostic accuracy	4.42	0.94	$r = 0.284, p = 0.002$	$r = 0.105, p = 0.259$	$r = 0.059, p = 0.526$	$r = 0.185, p = 0.045$	$r = 0.222, p = 0.016$

RQ2: Relations of conceptual and strategic diagnostic knowledge with justification, disconfirmation, and transparency

To explore the extent to which CDK and SDK predicted the dependent variables of justification, disconfirmation, and transparency, we calculated a multivariate multiple linear regression. Participants achieved $M = 8.86$ points ($SD = 1.66$) out of a maximum of 14 points on the CDK test and $M = 6.70$ points ($SD = 0.39$) out of a maximum of eight points on the SDK test (see Table 3). The Pearson correlations of the three argumentation facets with the variables CDK and SDK are reported in Table 3. The overall regression model with the predictors CDK and SDK significantly predicted justification—$F(2, 115) = 7.725$, $p = 0.001$—and explained 11.8% of the variance. Both CDK ($\beta = 0.236$, $p = 0.009$) and SDK ($\beta = 0.222$, $p = 0.013$) contributed significantly to the model. Similarly, disconfirmation was significantly predicted by the overall regression model, with the predictors CDK and SDK—$F(2, 115) = 3.331$, $p = 0.039$—explaining 5.5% of the variance. Whereas CDK ($\beta = 0.232$, $p = 0.012$) contributed significantly to the model, SDK did not ($\beta = 0.012$, $p = 0.898$). By contrast, transparency was not significantly predicted by the overall regression model, including both predictors, CDK and SDK—$F(2, 115) = 2.264$, $p = 0.109$—which explained 3.8% of the variance. CDK ($\beta = 0.016$, $p = 0.861$) was not a significant predictor of transparency; however, SDK ($\beta = 0.192$, $p = 0.040$) was a significant predictor of transparency in diagnostic argumentation.

Overall, justification, disconfirmation, and transparency were each partially explained by CDK, SDK, or both, with small effect sizes. Across the three facets, there were considerable differences in the amounts of variance explained by CDK and SDK. Moreover, the pattern in which CDK and SDK predicted justification, disconfirmation, and transparency differed considerably.

RQ3: Relationship between diagnostic judgment and diagnostic argumentation

To explore whether diagnostic judgment and diagnostic argumentation represent different diagnostic skills, we started again by plotting argumentation texts in ENA. First, we grouped argumentation texts according to diagnostic accuracy to compare argumentation concerning inaccurate versus accurate judgments. Second, we explored preservice teachers' total scores to investigate whether diagnostic accuracy correlated with justification, disconfirmation, and transparency in diagnostic argumentation.

To explore whether the argumentation texts differed if concerning an accurate vs. an inaccurate judgment, we grouped the individual argumentation texts by diagnostic accuracy and created one overall ENA network per group. We descriptively compared the networks of the groups of argumentation texts concerning accurate judgments (see Figure 4A) and inaccurate judgments (see Figure 4C), which we found to be highly similar (see also the comparison plot in Figure 4B, which shows the other two networks' differences). To determine whether the two groups of argumentation texts differed significantly, we centered the networks, resulting in the two group means (indicated by colored squares, with confidence intervals indicated by colored dashed boxes) depicted in Figure 4B.

All networks in Figure 4 were rotated to align both group means to the X-axis, which enabled statistical testing of group differences in a single dimension (Shaffer, 2017). The positioning of the group mean of argumentation texts concerning inaccurate judgments ($M = -0.01$, $SD = 0.38$, $n = 100$) was not statistically significantly different from the positioning of the group mean of argumentation texts concerning accurate judgments ($M = 0.01$, $SD = 0.41$, $n = 457$; $t(153.53) = 0.56$, $p = 0.58$, Cohen's $d = 0.06$). The analysis suggests that, overall, argumentation texts did not differ if addressing an accurate versus an inaccurate judgment.

We proceeded with a correlational analysis of preservice teachers' total scores to investigate whether their overall diagnostic accuracy was correlated with justification, disconfirmation, and transparency (see Table 3). On average, participants achieved a diagnostic accuracy of $M = 4.42$ points ($SD = 0.94$) out of a maximum of six achievable points. We found that participants' diagnostic accuracy and justification were significantly correlated, with a small effect ($r = 0.284$, $p = 0.002$). By contrast, diagnostic accuracy was not significantly correlated with either disconfirmation ($r = 0.105$, $p = 0.259$) or transparency ($r = 0.059$, $p = 0.526$).

To determine the role of CDK and SDK in explaining the relationship between diagnostic accuracy and justification, we calculated a partial correlation between diagnostic accuracy and

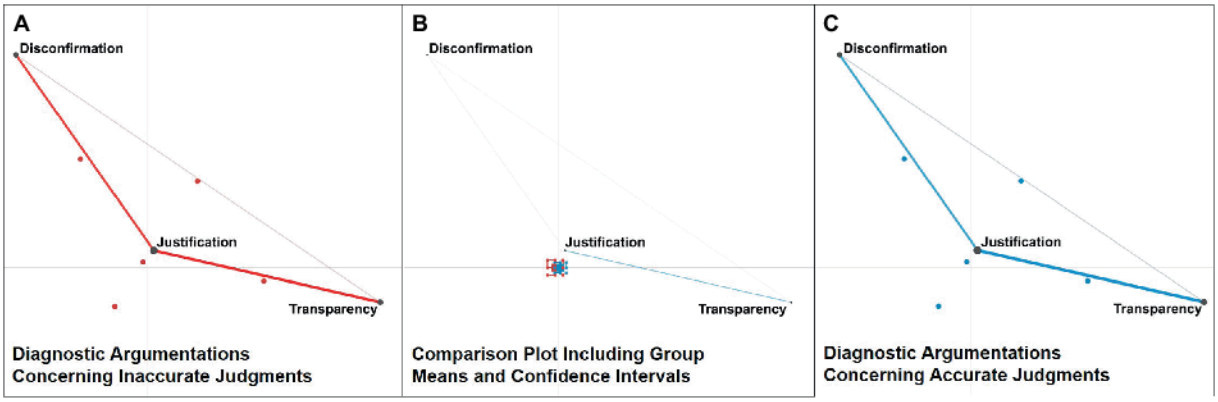


FIGURE 4
Networks across diagnostic argumentation texts, grouped by diagnostic accuracy and rotated by group means: **(A)** shows the network of diagnostic argumentation texts concerning inaccurate judgments. **(C)** shows the network of diagnostic argumentation texts concerning accurate judgments. **(B)** shows the comparison plot, which depicts the differences between the other two networks, as well as the group means (indicated by colored squares) and confidence intervals (indicated by colored dashed boxes) of the other two networks.

justification, statistically controlling for CDK and SDK (see Table 3 for the Pearson correlations of CDK and SDK with the argumentation facets and diagnostic accuracy). We found that the resulting partial correlation between diagnostic accuracy and justification in diagnostic argumentation remained significant, with a small effect ($r = 0.211$, $p = 0.023$). Thus, controlling for CDK and SDK hardly decreased the effect size of the correlation between diagnostic accuracy and justification. Consequently, our results suggest that CDK and SDK are not the variables that primarily explain the relationship between diagnostic accuracy and justification.

Overall, the results only indicate a weak relationship between the accuracy of preservice teachers' diagnostic judgments on the one hand, and justification, disconfirmation, and transparency in their diagnostic argumentation on the other. CDK and SDK did not explain the small correlation between diagnostic accuracy and justification. Moreover, groups of argumentation texts concerning inaccurate versus accurate judgments did not show a statistically significant difference. These findings suggest that diagnostic judgment and diagnostic argumentation can be considered different diagnostic skills.

Discussion

In exploring whether justification, disconfirmation, and transparency represent distinct subskills or one joint underlying diagnostic skill (RQ1), we found that preservice teachers primarily provide justification in their diagnostic argumentation as an antecedent to including disconfirmation or transparency in their diagnostic argumentation. Furthermore, we found a unidirectional dependency of disconfirmation on justification; diagnostic argumentation texts presenting more than one diagnosis usually discussed the differential diagnoses by evaluating evidence to make conclusions; however, preservice teachers often only argued for their final diagnosis without discussing competing explanations. Concerning the interrelations between justification, disconfirmation, and transparency, we found that they were distinguishable facets of diagnostic argumentation. Determining the extent to which justification, disconfirmation, and transparency were explained by CDK and SDK (RQ2), we found that justification was predicted by CDK about diagnoses and evidence as well as SDK about diagnostic approaches and activities. Disconfirmation of different diagnoses was only predicted by CDK of diagnoses. By contrast, transparency about the diagnostic approaches for generating evidence was only predicted by SDK of diagnostic proceedings for generating evidence. However, the variance explained by CDK and SDK was low. Furthermore, the accuracy of diagnostic judgments and justification, disconfirmation, and transparency in diagnostic argumentation did not necessarily seem to be related (RQ3). Overall, groups of argumentation texts addressing either accurate or inaccurate diagnostic judgments did not show a statistically significant difference. However, in contrast to disconfirmation and transparency, we found that justification in diagnostic argumentation was significantly correlated with the accuracy of diagnostic judgments. Despite statistically controlling for CDK and SDK, the relationship between the accuracy of diagnostic judgments and justification in diagnostic argumentation remained significant, suggesting that other variables may be important in explaining the relationship.

Overall, we interpreted the results as suggesting that diagnostic judgment and diagnostic argumentation might be different diagnostic skills. Finding a relationship between the accuracy of diagnostic judgments and justification in diagnostic argumentation supports the relevance and validity of the construct of diagnostic argumentation. Yet, the argumentation facets seemed to be sufficiently distinguishable from one another and from diagnostic accuracy. Finding differences regarding the predictive patterns of CDK and SDK (see Förtsch et al., 2018) supports the notion that justification, disconfirmation, and transparency are distinct subskills of diagnostic argumentation. Justification involves explicitly evaluating evidence as the basis for

concluding a diagnosis (see Fischer et al., 2014; Heitzmann et al., 2019). Therefore, justification requires CDK about relevant concepts (e.g., diagnoses, evidence, and their interrelations; see Förtsch et al., 2018). Moreover, justification requires making warranted connections between evidence and diagnoses (e.g., Toulmin, 1958) to conclude or reject diagnoses, which seems to be facilitated by SDK (see Förtsch et al., 2018). Disconfirmation involves addressing differential diagnoses to demonstrate that alternative explanations have been considered (e.g., Toulmin, 1958; Lawson, 2003), which seems to primarily require CDK about differential diagnoses. By contrast, transparency, which involves describing the processes behind evidence generation (see Chinn et al., 2014; Vazire, 2017), seems to rely on SDK when it comes to the process of diagnosing a specific problem (e.g., which informational sources can deliver critical evidence).

Large amounts of variance in justification, disconfirmation, and transparency remained unexplained by CDK and SDK. Our findings raise the question of which additional kinds of knowledge and skills may be used when formulating justified, disconfirming, and transparent diagnostic argumentation. Beyond CDK and SDK, we propose two additional variables that might play a role in explaining justification, disconfirmation, and transparency within diagnostic argumentation: (1) knowledge about standards in diagnosing and diagnostic argumentation (see Chinn et al., 2014; Bauer et al., 2020) and (2) argumentation skills that are transferrable across domains (Hetmanek et al., 2018). In teacher education, there seems to be limited agreement about standards in diagnostic practices compared with other fields, such as medical education (Bauer et al., 2020). Teacher education programs do not yet systematically teach agreed-upon standards for communicating in situations that require what we defined as diagnostic argumentation. Consequently, preservice teachers likely do not have much knowledge about standards in diagnostic argumentation. There might also be differences between teacher and medical education in what are considered suitable standards for diagnostic argumentation (Bauer et al., 2020). Moreover, teachers and teacher educators might vary in their views regarding the role of scientific standards in diagnostic argumentation. Therefore, it is important to continue to discuss such standards in teacher education. We suggest using justification, disconfirmation, and transparency as a starting point from which to further discuss, systematize, and teach standards for diagnostic argumentation in teacher education.

The performance differences and higher prevalence of justification observed in the current study may be explained by argumentation skills that are transferrable across domains. It has been suggested that cross-domain transferable skills can, to some extent, compensate for a lack of more specifically relevant knowledge (e.g., knowledge about standards in diagnostic argumentation; Hetmanek et al., 2018). Accordingly, knowledge about standards in diagnostic argumentation, as well as cross-domain transferable argumentation skills, may be relevant for explaining justification, disconfirmation, and transparency in preservice teachers' diagnostic argumentation beyond their CDK and SDK. Other possible sources of variance are additional kinds of knowledge used in diagnosing that were not considered in this study, such as scientific knowledge that is not pertinent to the context (e.g., Hetmanek et al., 2015) or subjective theories, beliefs, and epistemic goals (Stark, 2017).

CDK and SDK also did not explain the relationship found between the accuracy of diagnostic judgments and justification in diagnostic argumentation. Beyond a joint knowledge base, another variable that could potentially explain the relationship between accuracy and justification may be the different types of information processing that occur during the judgment process (see Loibl et al., 2020). The literature on dual-process theories (see Kahneman, 2003; Evans, 2008) suggests that controlled information processing results in more conscious and explicable reasoning compared to intuitive information processing (e.g., pattern recognition; see Evans, 2008). Thus, a controlled analysis of evidence during the judgment process could affect the accuracy of diagnostic judgments (see Coderre et al., 2010; Norman et al., 2017) and at the same time facilitate justification in diagnostic argumentation.

Limitations and future research

One methodological limitation that needs to be discussed is the low internal consistency of diagnostic accuracy across diagnostic judgments, which may hide further correlations that were not observed in the results. Low internal consistency values are a common issue in measurement instruments with small numbers of items (e.g., Monteiro et al., 2020). However, we did not assume that low internal consistency was a major issue for our interpretations because we still found the theoretically expected relations of diagnostic accuracy with the variables CDK and SDK.

The operationalization of the judgment process in the simulation-based learning environment might be considered to limit generalizability to real-life practice situations, in which teachers' judgment processes might take place over several days or weeks and involve higher degrees of complexity and ambiguity compared to our simulated cases. However, in our simulation, preservice teachers could decide by themselves how much evidence they wanted to collect, and in which order they would access which informational sources (e.g., conversation protocols). Therefore, we argue that, for the purpose of our research goals, the simulation provided a sufficient representation of a real-world diagnostic situation.

Descriptive results of the participants' performance in all three argumentation facets across the measurement points of the different cases suggest that participants' performance generally decreased throughout the data collection. The long duration of the study might have exhausted the participants or decreased their motivation. In addition, some participants might have concluded from the order of the tasks in the simulated cases that they would not need to include their initially indicated diagnostic judgments as a conclusion in their subsequently written diagnostic argumentation texts. Given that the operationalization of justification required participants not only to evaluate evidence but also to explicate conclusions in their argumentation texts, their argumentation skills in terms of justification might have

been underestimated in our study. Therefore, generalizing to teachers in authentic classroom situations based on our participants' performance should be done with caution.

There are areas other than students' clinical problems in which teachers' diagnosing is relevant (e.g., assessing a student's level of skill). Our choice of topic might limit the generalizability of the findings to other areas of diagnosing in teacher education. However, we consider the conceptualization of diagnostic argumentation (i.e., justification, disconfirmation, and transparency) presented in this article nonspecific to the content area of clinical problems. Thus, we expect the result pattern to be replicable in other areas of teachers' diagnosing, which could be investigated in further research.

To explore the research questions addressed in this paper, we reanalyzed the data collected in a prior cross-sectional study. The sample was too small to employ structural equation modeling, which would have been preferable to analyzing the data with correlation and regression analyses. Although our results provide initial evidence of the potential relationships between the investigated constructs, they must be replicated in future research using larger samples and advanced methods.

Future research is necessary to further validate the findings that diagnostic argumentation is a diagnostic skill that is distinct from diagnostic judgment. For this purpose, we recommend the approach to investigate preservice teachers' performance based on both qualitative and quantitative data as illustrated in our study. In particular, possible joint predictors of accurate diagnostic judgments and justified diagnostic argumentation, such as controlled information processing during the judgment process, require further clarification because CDK and SDK did not seem to explain the relation between accuracy and justification. Additionally, further research in teacher education should investigate the knowledge and skills that underlie justification, disconfirmation, and transparency beyond CDK and SDK, such as knowledge about standards in diagnosing, cross-domain transferrable argumentation skills, as well as subjective theories, beliefs, and experiential knowledge regarding evidence-oriented practice.

In our study, we did not specify a particular recipient to whom preservice teachers should direct their diagnostic argumentation. However, diagnostic argumentation might vary considerably depending on the recipient (e.g., a teacher colleague, a school psychologist, or a parent) and the argumentative situation (during a collaborative judgment process or subsequent to making a judgment). For example, prior research in collaborative diagnosing has emphasized the potential role of meta-knowledge about the collaborating professional's role and responsibilities (Radkowitsch et al., 2021). Therefore, future studies might systematically investigate the role of different recipients in teachers' diagnostic argumentation.

Research may also validate whether professionals in teacher education perceive justification, disconfirmation, and transparency as facilitating comprehensibility and persuasiveness in diagnostic argumentation or whether our suggested conception of argumentation facets needs to be further specified for the area of teacher education. One interesting and potentially relevant direction in which to further develop our conception might be found in the literature on professional vision, which distinguishes between describing and interpreting evidence as two different forms of how evidence is reported and evaluated in the context of teachers' diagnosing (Seidel and Stürmer, 2014; Kramer et al., 2021b). Moreover, researchers could explore the potential of different learning opportunities and support measures for fostering preservice teachers' diagnostic argumentation. Similarly, researchers could investigate whether diagnostic judgment and diagnostic argumentation have similar or different developmental trajectories and might benefit from similar or different forms of instruction.

Conclusion

In this article, we presented evidence suggesting that diagnostic judgment and diagnostic argumentation might represent different diagnostic skills. Preservice teachers do not necessarily seem to be equally capable of making accurate diagnostic judgments on the one hand, and formulating justified, disconfirming, and transparent diagnostic argumentation on the other. We suggest that justification, disconfirmation, and transparency can be considered relevant facets and distinct subskills of diagnostic argumentation, as our results appear to indicate differences in the underlying knowledge bases. Despite the fact that CDK and SDK explain some variance in justification, disconfirmation, and transparency, the portion of variance they explain might be rather small. Thus, additional variables may be relevant predictors of justification, disconfirmation, and transparency in diagnostic argumentation, such as knowledge of diagnostic standards or cross-domain transferable argumentation skills. Including these additional constructs in further investigations would be a promising direction for future research on diagnostic argumentation. In addition, it seems particularly important that researchers and educators in the field of teacher education, as well as in-service teachers as practitioners in the field, further reflect on standards in diagnosing and diagnostic argumentation. Justification, disconfirmation, and transparency may serve as a productive set of constructs for establishing standards for teachers' diagnostic argumentation in the future.

Data availability statement

The data presented in this article will be made available by the authors upon request. Requests to access the data should be directed to elisabeth.bauer@psy.lmu.de.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Medical Faculty of LMU Munich (no. 17-249). The participants provided their written informed consent to participate in this study.

Author contributions

EB, MS, JK, MF, and FF developed the study concept and contributed to the study design. EB performed the data analysis. EB, MS, and FF interpreted the data. EB drafted the manuscript. MS, JK, MF, and FF provided critical revisions. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the German Federal Ministry of Research and Education (FAMULUS-Project 16DHL1040) and the Elite Network of Bavaria (K-GS-2012-209).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Albritton, K., Chen, C.-I., Bauer, S. G., Johnson, A., and Mathews, R. E. (2021). Collaborating with school psychologists: moving beyond traditional assessment practices. *Young Except. Children* 24, 28–38. doi: 10.1177/1096250619871951

Bauer, E., Fischer, F., Kiesewetter, J., Shaffer, D. W., Fischer, M. R., Zottmann, J. M., et al. (2020). Diagnostic activities and diagnostic practices in medical education and teacher education: an interdisciplinary comparison. *Front. Psychol.* 11, 1–9. doi: 10.3389/fpsyg.2020.562665

Berland, L. K., and Reiser, B. J. (2009). Making sense of argumentation and explanation. *Sci. Educ.* 93, 26–55. doi: 10.1002/sce.20286

Boshuizen, H. P., Gruber, H., and Strasser, J. (2020). Knowledge restructuring through case processing: the key to generalise expertise development theory across domains? *Educ. Res. Rev.* 29:100310. doi: 10.1016/j.edurev.2020.100310

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., and Fischer, F. (2020). Facilitating diagnostic competences in higher education - a meta-analysis in medical and teacher education. *Educ. Psychol. Rev.* 32, 157–196. doi: 10.1007/s10648-019-09492-2

Chinn, C. A., and Duncan, R. G. (2018). "What is the value of general knowledge of scientific reasoning? Scientific reasoning and argumentation: the roles of domain-specific and domain-general knowledge," in *Scientific reasoning and argumentation*. eds. F. Fischer, A. C. Clark, K. Engelmann and J. Osborne (New York: Routledge), 77–101.

Chinn, C. A., Rinehart, R. W., and Buckland, L. A. (2014). "Epistemic cognition and evaluating information: applying the AIR model of epistemic cognition," in *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. eds. D. Rapp and J. Braasch (Cambridge, MA: MIT Press), 425–453.

Coderre, S., Wright, B., and McLaughlin, K. (2010). To think is good: querying an initial hypothesis reduces diagnostic error in medical students. *Acad. Med.* 85, 1125–1129. doi: 10.1097/ACM.0b013e3181e1b229

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., and Seidel, T. (2021). Exploring the process of preservice teachers' diagnostic activities in a video-based simulation. *Front. Educ.* 6:626666. doi: 10.3389/feduc.2021.626666

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Csanadi, A., Kollar, I., and Fischer, F. (2020). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: better together? *Eur. J. Psychol. Educ.* 36, 147–168. doi: 10.1007/s10212-020-00467-4

Diamantopoulos, A., and Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: a comparison and empirical illustration. *Br. J. Manag.* 17, 263–282. doi: 10.1111/j.1467-8551.2006.00500.x

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

Feldon, D. F. (2007). Cognitive load and classroom teaching: the double-edged sword of automaticity. *Educ. Psychol.* 42, 123–137. doi: 10.1080/00461520701416173

Fischer, F. (2021). Some reasons why evidence from educational research is not particularly popular among (pre-service) teachers: a discussion. *Zeitschrift für Pädagogische Psychologie* 35, 209–214. doi: 10.1024/1010-0652/a000311

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* 2, 28–45. doi: 10.14786/flr.v2i2.96

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76:378, –382. doi: 10.1037/h0031619

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M. R., Girwidz, R., Obersteiner, A., et al. (2018). Systematizing professional knowledge of medical doctors and teachers: development of an interdisciplinary framework in the context of diagnostic competences. *Educ. Sci.* 8:207. doi: 10.3390/educsci8040207

Gorman, M. E., Gorman, M. E., Latta, R. M., and Cunningham, G. (1984). How disconfirmatory, confirmatory and combined strategies affect group problem solving. *Br. J. Psychol.* 75, 65–79. doi: 10.1111/j.2044-8295.1984.tb02790.x

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

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., and Williamson, P. (2009). Teaching practice: a cross-professional perspective. *Teach. Coll. Rec.* 111, 2055–2100. doi: 10.1177/016146810911100905

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

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teach. Teach. Educ.* 76, 181–193. doi: 10.1016/j.tate.2017.12.001

Hetmanek, A., Engelmann, K., Opitz, A., and Fischer, F. (2018). "Beyond intelligence and domain knowledge: scientific reasoning and argumentation as a set of cross-domain skills. Scientific reasoning and argumentation: the roles of domain-specific and domain-general knowledge," in *Scientific reasoning and argumentation*. eds. F. Fischer, A. C. Clark, K. Engelmann and J. Osborne (New York: Routledge), 203–226. doi: 10.4324/9780203731826

Hetmanek, A., Wecker, C., Kiesewetter, J., Trempler, K., Fischer, M. R., Gräsel, C., et al. (2015). Wozu nutzen Lehrkräfte welche Ressourcen? Eine Interviewstudie zur Schnittstelle zwischen bildungswissenschaftlicher Forschung und professionellem Handeln im Bildungsbereich [for which purposes do teachers use which resources? An interview study on the relation between educational research and professional educational practice]. *Unterrichtswissenschaft* 43, 194–210.

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

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *Am. Psychol.* 58, 697–720. doi: 10.1037/0003-066X.58.9.697

Kiemer, K., and Kollar, I. (2021). Source selection and source use as a basis for evidence-informed teaching. *Zeitschrift für Pädagogische Psychologie* 35, 127–141. doi: 10.1024/1010-0652/a000302

Kiesewetter, J., Fischer, F., and Fischer, M. R. (2017). Collaborative clinical reasoning—a systematic review of empirical studies. *J. Contin. Educ. Health Prof.* 37, 123–128. doi: 10.1097/CEH.0000000000000158

Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognit. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Kolovou, D., Naumann, A., Hochweber, J., and Praetorius, A.-K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teach. Teach. Educ.* 100:103298. doi: 10.1016/j. tate.2021.103298

Kramer, M., Förtsch, C., Boone, W. J., Seidel, T., and Neuhaus, B. J. (2021a). Investigating pre-service biology teachers' diagnostic competences: relationships between professional knowledge, diagnostic activities, and diagnostic accuracy. *Educ. Sci.* 11:89. doi: 10.3390/educsci11030089

Kramer, M., Förtsch, C., Seidel, T., and Neuhaus, B. J. (2021b). Comparing two constructs for describing and analyzing teachers' diagnostic processes. *Stud. Educ. Eval.* 68:100973. doi: 10.1016/j.stueduc.2020.100973

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lawson, A. (2003). The nature and development of hypothetico-predictive argumentation with implications for science teaching. *Int. J. Sci. Educ.* 25, 1387–1408. doi: 10.1080/0950069032000052117

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

McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mercier, H., and Heintz, C. (2014). Scientists' argumentative reasoning. *Topoi* 33, 513–524. doi: 10.1007/s11245-013-9217-4

Mercier, H., and Sperber, D. (2017). *The enigma of reason*. Cambridge, Massachusetts: Harvard University Press.

Monteiro, S. D., Sherbino, J., Schmidt, H., Mamede, S., Ilgen, J., and Norman, G. (2020). It's the destination: diagnostic accuracy and reasoning. *Adv. Health Sci. Educ.* 25, 19–29. doi: 10.1007/s10459-019-09903-7

Norman, G. R., Monteiro, S. D., Sherbino, J., Ilgen, J. S., Schmidt, H. G., and Mamede, S. (2017). The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. *Acad. Med.* 92, 23–30. doi: 10.1097/ACM.0000000000001421

Page, G., Bordage, G., and Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Acad. Med.* 70, 194–201. doi: 10.1097/00001888-199503000-00009

Pfeiffer, J., Meyer, C. M., Schulz, C., Kiesewetter, J., Zottmann, J., Sailer, M., et al (2019). "FAMULUS: Interactive Annotation and Feedback Generation for Teaching Diagnostic Reasoning," *in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations,*. eds. S. Padó and R. Huang Stroudsburg, PA, USA: Association for Computational Linguistics., 73–78

Pit-ten Cate, I. M., Hörstermann, T., Krolak-Schwerdt, S., Gräsel, C., Böhmer, I., and Glock, S. (2020). Teachers' information processing and judgement accuracy: effects of information consistency and accountability. *Eur. J. Psychol. Educ.* 35, 675–702. doi: 10.1007/s10212-019-00436-6

Poznanski, B., Hart, K. C., and Graziano, P. A. (2021). What do preschool teachers know about attention-deficit/hyperactivity disorder (ADHD) and does it impact ratings of child impairment? *Sch. Ment. Heal.* 13, 114–128. doi: 10.1007/s12310-020-09395-6

Praetorius, A. K., Berner, V. D., Zeinz, H., Scheunpflug, A., and Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *J. Educ. Res.* 106, 64–76. doi: 10.1080/00220671.2012.667010

Radkowitsch, A., Sailer, M., Schmidmaier, R., Fischer, M. R., and Fischer, F. (2021). Learning to diagnose collaboratively–effects of adaptive collaboration scripts in agent-based medical simulations. *Learn. Instr.* 75:101487. doi: 10.1016/j. learninstruc.2021.101487

Rieu, A., Leuders, T., and Loibl, K. (2022). Teachers' diagnostic judgments on tasks as information processing–the role of pedagogical content knowledge for task diagnosis. *Teach. Teach. Educ.* 111:103621. doi: 10.1016/j.tate.2021.103621

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., et al. (2022). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learn. Instr.* 101620, 1–10. doi: 10.1016/j.learninstruc.2022.101620

Sampson, V., and Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: current perspectives and recommendations for future directions. *Sci. Educ.* 92, 447–472. doi: 10.1002/sce.20276

Schmidmaier, R., Eiber, S., Ebersbach, R., Schiller, M., Hege, I., Holzer, M., et al. (2013). Learning the facts in medical school is not enough: which factors predict successful application of procedural knowledge in a laboratory setting? *BMC Med. Educ.* 13, 1–9. doi: 10.1186/1472-6920-13-28

Seidel, T., and Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *Am. Educ. Res. J.* 51, 739–771. doi: 10.3102/0002831214531321

Shaffer, D. W. (2017). *Quantitative ethnography*, Madison, Wisconsin: Cathcart Press.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15, 4–14. doi: 10.3102/0013189X015002004

Stadler, M., Sailer, M., and Fischer, F. (2021). Knowledge as a formative construct: a good alpha is not always better. *New Ideas Psychol.* 60:100832. doi: 10.1016/j. newideapsych.2020.100832

Stark, R. (2017). Probleme evidenzbasierter bzw. -orientierter pädagogischer praxis [problems of evidence-based or rather evidence-oriented educational practice]. *Zeitschrift für Pädagogische Psychologie* 31, 99–110. doi: 10.1024/1010-0652/a000201

Stark, R., Kopp, V., and Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: two experimental studies in undergraduate medical education. *Learn. Instr.* 21, 22–33. doi: 10.1016/j.learninstruc.2009.10.001

Südkamp, A., Praetorius, A.-K., and Spinath, B. (2018). Teachers' judgment accuracy concerning consistent and inconsistent student profiles. *Teach. Teach. Educ.* 76, 204–213. doi: 10.1016/j.tate.2017.09.016

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

Toulmin, S. E. (1958). *The uses of argument*, Cambridge: Cambridge University Press.

Urhahne, D., and Wijnia, L. (2020). A review on the accuracy of teacher judgments. *Educ. Res. Rev.* 32:100374. doi: 10.1016/j.edurev.2020.100374

Van Gog, T., Paas, F., and Van Merriënboer, J. J. (2004). Process-oriented worked examples: improving transfer performance through enhanced understanding. *Instr. Sci.* 32, 83–98. doi: 10.1023/B:TRUC.0000021810.70784.b0

Vanlommel, K., Van Gasse, R., Vanhoof, J., and Van Petegem, P. (2017). Teachers' decision-making: data based or intuition driven? *Int. J. Educ. Res.* 83, 75–83. doi: 10.1016/j.ijer.2017.02.013

Vazire, S. (2017). Quality uncertainty erodes trust in science. Collabra. *Psychology* 3. doi: 10.1525/collabra.74

Walton, D. N. (1990). What is reasoning? What is an argument? *J. Philos.* 87, 399–419. doi: 10.2307/2026735

Wildgans-Lang, A., Scheuerer, S., Obersteiner, A., Fischer, F., and Reiss, K. (2020). Analyzing prospective mathematics teachers' diagnostic processes in a simulated environment. *ZDM* 52, 241–254. doi: 10.1007/s11858-020-01139-9

4

Study 3: Adaptive Feedback from Artificial Neural Networks Facilitates Pre-service Teachers' Diagnostic Reasoning in Simulation-Based Learning

Reference: Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction, 83*, Article 101620. https://doi.org/10.1016/j.learninstruc.2022.101620

Learning and Instruction 83 (2023) 101620

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning

Michael Sailer [a,*], Elisabeth Bauer [a], Riikka Hofmann [b], Jan Kiesewetter [c], Julia Glas [a], Iryna Gurevych [d], Frank Fischer [a]

[a] Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany
[b] Faculty of Education, University of Cambridge, Cambridge, UK
[c] Institute for Medical Education, University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany
[d] Ubiquitous Knowledge Processing Lab, Department of Computer Science, Technical University of Darmstadt, Darmstadt, Germany

ARTICLE INFO

Keywords:
Simulation-based learning
Teacher education
Artificial intelligence
Adaptive feedback
Natural language processing

ABSTRACT

In simulations, pre-service teachers need sophisticated feedback to develop complex skills such as diagnostic reasoning. In an experimental study with $N = 178$ pre-service teachers about simulated pupils with learning difficulties, we investigated the effects of automatic adaptive feedback, which is based on artificial neural networks, on pre-service teachers' diagnostic reasoning. Diagnostic reasoning was operationalised as diagnostic accuracy and the quality of justifications. We compared automatic adaptive feedback with static feedback, which we provided in form of an expert solution. Further, we experimentally manipulated whether the learners worked individually or in dyads on the computer lab-based simulations. Results show that adaptive feedback facilitates pre-service teachers' quality of justifications in written assignments, but not their diagnostic accuracy. Further, static feedback even had detrimental effects on the learning process in dyads. Automatic adaptive feedback in simulations offers scalable, elaborate, process-oriented feedback in real-time to high numbers of students in higher education.

1. Introduction

Teachers' diagnostic reasoning skills are essential for dealing with increasing diversity and heterogeneity in classrooms: pupils have diverse and changing learning prerequisites that teachers must consider in order to offer individual support (Reinke, Stormont, Herman, Puri, & Goel, 2011). However, there are indications that diagnostic reasoning is often neglected in teacher education and that teachers themselves consider their diagnostic skills insufficient (Poznanski, Hart, & Graziano, 2021). In teacher education as in many other higher education (HE) programmes, it is often not possible to offer extensive real-life practice of specific instances of diagnostic reasoning (Grossman et al., 2009; Heitzmann et al., 2019).

One promising option to overcome this gap between education and practice is to provide pre-service teachers with simulation-based learning opportunities, which are less overwhelming than real-life situations by isolating skills early on in professional learning (Chernikova et al., 2020). However, simulations might not be helpful per se, but need to be accompanied by further instructional guidance like targeted feedback to become effective. Specifically, due to the complexity involved in simulation-based learning of diagnostic reasoning, learners may need specific support and feedback to make full use of their learning opportunities (Kiesewetter et al., 2020). Feedback that is adapted to learners' needs is resource intensive for HE teachers (see Henderson, Ryan, & Phillips, 2019); partially automating the feedback seems promising but also challenging.

Involving collaborative learning scenarios is another pedagogical approach in the context of simulation-based learning. Existing studies found that, compared to individuals, collaborative learners often perform better in solving reasoning problems in simulated scenarios (Csanadi, Kollar, & Fischer, 2021). Moreover, learners seem to be better in critical evaluation of other's arguments than their own arguments (Mercier & Sperber, 2017), suggesting collaborative scenarios may be beneficial for learning complex diagnostic tasks. We conducted a study in which we investigated the effects of automated adaptive feedback on pre-service teachers' simulation-based learning of diagnostic reasoning in individual and collaborative learning settings.

* Corresponding author. Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstraße 13, 80802, Munich, Germany.
E-mail address: Michael.Sailer@psy.lmu.de (M. Sailer).

https://doi.org/10.1016/j.learninstruc.2022.101620
Received 30 June 2021; Received in revised form 10 January 2022; Accepted 23 March 2022
Available online 11 April 2022

1.1. Teachers' diagnostic reasoning

Facilitating pre-service teachers' learning of diagnostic reasoning seems important to prepare them for diagnostic reasoning in real classroom settings and in school. We consider teachers' diagnostic reasoning broadly as the goal-oriented collection and integration of information, aiming to reduce uncertainty in order to make educational decisions (Heitzmann et al., 2019). Studies have shown that teachers' diagnosis of their pupils' various learning prerequisites is important in order to provide individual pupils with suitable support (Reinke et al., 2011). Part of pupils' learning prerequisites may also be learning difficulties such as dyslexia or behavioural disorders like attention deficit hyperactivity disorder (ADHD). These require early identification to intervene in a timely manner and provide affected pupils with the necessary support. To avoid disadvantaging single pupils due to insufficient or unsuited support, generating greater problems in the future, it is vital that teachers identify cues to learning difficulties early in a pupil's school career. We conceptualise the recognition of these cues as an important part of teachers' effective diagnostic reasoning (Poznanski et al., 2021).

Diagnostic reasoning skills can develop by engaging in practice, i.e., by repeated knowledge application and exposure to various diagnostic problems. Thereby, knowledge becomes increasingly encapsulated into higher-level concepts (Schmidt & Rikers, 2007). With further experience and thus practice, knowledge is integrated into episodic representations of diagnostic problems, which is referred to as script formation (Charlin et al., 2012; Lachner, Jarodzka, & Nückles, 2016). Diagnostic reasoning can be assessed regarding the achievement of target criteria, such as diagnostic accuracy, which indicates the degree of correctness of a teacher's diagnostic judgement (Kolovou, Naumann, Hochweber, & Praetorius, 2021). Beyond achieving diagnostic accuracy, justifying diagnoses and explaining the underlying diagnostic reasoning are helpful and crucial for collaborating with other teachers or school psychologists (Csanadi et al., 2021). Justifying diagnoses by providing supporting evidence facilitates collaborators' understanding of the diagnostic reasoning (see Hitchcock, 2005). Justifications can also facilitate a process of considering and reconciling explanations within collaborative diagnostic reasoning and thus help improve the diagnosing (see Berland & Reiser, 2009). Therefore, we conceptualise the pre-service teachers' learning of diagnostic reasoning with both outcomes of teachers' reasoning, diagnostic accuracy and the quality of justifications.

1.2. Simulation-based learning to foster diagnostic reasoning

To foster diagnostic reasoning, there is evidence that pre-service teachers' practicing diagnostic reasoning in authentic contexts or with authentic cases during HE can be effective (Van Merriënboer, 2013; VanLehn, 1996), which is one reason why simulation-based learning has been identified as an innovative way forward in teacher education. Simulations are partial representations of professional situations, with a set of features that can be manipulated by learners (see Codreanu, Sommerhoff, Huber, Ufer, & Seidel, 2020). This can involve authentic cases of simulated pupils that teachers will deal with as part of their work. Authentic cases of simulated pupils provide learning opportunities to practice diagnostic reasoning, which is needed as an in-service teacher. Simulations are particularly beneficial in terms of practicing critical but infrequent situations and focusing on specific subsets of practices in which learners can repeatedly engage (Grossman et al., 2009). Therefore, simulation-based learning is considered a highly promising instructional approach for learning diagnostic reasoning in teacher education (Codreanu et al., 2020).

However, diagnosing simulated pupils even in simulation-based learning is a complex task for pre-service teachers and might not be effective per se (see Kiesewetter et al., 2020). Research emphasised that especially novice learners, who lack a certain level of prior knowledge and skills, need particular support and feedback to effectively learn complex skills (Cook et al., 2013; Wisniewski, Zierer, & Hattie, 2019). Therefore, pre-service teachers may need specific support and feedback to effectively learn diagnostic reasoning in simulation-based learning.

1.2.1. Adaptive feedback in simulation-based learning

Receiving feedback is considered a necessary condition for harnessing the potentials of simulation-based learning of complex skills, such as diagnostic reasoning (Cook et al., 2013; Scheuer, McLaren, Loll, & Pinkwart, 2012). In order to be effective in supporting the learning of complex skills, feedback needs to elaborate on ways to appropriately process the task, not only provide information about correct task solutions (Narciss et al., 2014; Wisniewski et al., 2019). Elaborating on appropriate or optimal processing of the task is often done by presenting expert solutions, which exemplify the processing of the task following the learner's own efforts to solve the problem (Renkl, 2014). Presenting an expert solution as a form of static feedback (i.e., non-adaptive feedback) is resource-efficient in HE, because all learners receive the same generic feedback; besides, it can easily be provided automatically in digital learning environments. However, learners need to determine their current state of knowledge and performance and figure out options for improvement by themselves, by comparing their own processing and solution with the expert solution. This process can be demanding and difficult for learners, involving being confronted with a large amount of information, possibly exceeding learners' cognitive capacity (Sweller, van Merriënboer, & Paas, 2019). In contrast to such static feedback, adaptive feedback can accommodate learners' specific needs by making appropriate adjustments to the feedback based on learners' performance (see Plass & Pawar, 2020). Such adaptive feedback can highlight and thus facilitate learners' understanding of their current state of knowledge and options for improvement, for example by identifying gaps between a learner's current and desired knowledge state or providing additional explanations if the task processing was flawed (Bimba, Idris, Al-Hunaiyyan, Mahmud, & Shuib, 2017; Narciss et al., 2014; Plass & Pawar, 2020). Adaptive feedback might thus increase germane cognitive load, that is, the cognitive resources invested in actual learning processes (Sweller et al., 2019). Freeing up cognitive resources for learning processes to actually happen might be particularly helpful in learning complex skills like diagnostic reasoning. Therefore, pre-service teachers' simulation-based learning of diagnostic reasoning might be effectively supported by process-oriented, adaptive feedback on their diagnostic reasoning (Wisniewski et al., 2019).

1.2.2. Automation of adaptive feedback in simulation-based learning

Adaptive feedback, however, is resource-intensive for HE teachers if done manually for every learner's task solution. Automating adaptive feedback on the learners' task processing to make process-oriented, adaptive feedback accessible to numerous learners is a potential solution. Research explored various possible applications of automatically assessing closed format questions or log data in cognitive tutors and intelligent tutoring systems (Graesser, Hu, & Sottilare, 2018). However, complex reasoning tasks in simulations require justifications consolidated in written explanations. For open ended explanations, recent advancements in artificial intelligence and machine learning offer new technical capabilities with help of artificial neural networks. In particular, methods of Natural Language Processing (NLP) aim to parse, analyse, and understand human language (Manning & Schütze, 2005) and thus, enable automating a real-time measurement of certain aspects of learners' written solutions without a human corrector (see Plass & Pawar, 2020). In the context of diagnostic reasoning, artificial neural network-based NLP models for sequence tagging can be specialised for the particular context of diagnostic reasoning: they can be trained to automatically detect diagnostic entities (e.g., cues or diagnoses) and epistemic activities (e.g., hypothesis generation or evidence evaluation; see Schulz, Meyer, and Gurevych (2019) in learners' written explanations. Based on predictions provided by the models, pre-service teachers can be automatically offered predefined feedback elements that are

adapted to the corresponding detected diagnostic entities and epistemic activities in written explanations of their diagnostic reasoning (Pfeiffer et al., 2019). However, the use of NLP involves challenges: the predominant learning paradigm in NLP utilises transfer learning strategies where models or word representations are pre-trained on freely available text corpora (Howard & Ruder, 2018; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). These models are subsequently fine-tuned on the target task, which, depending on the similarity of the source and target domain, can result in a considerable decrease in performance. This is particularly challenging when the target task involves domain-specific terms, which might have been seldomly used in text corpora or even in training data that are used for fine-tuning the target task. Further, as the pre-training approaches rely on unsupervised methods, i.e., trained on unlabelled data, biases (e.g. gender) can be encoded in the pre-trained representations (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016), which have to be monitored and considered. Despite these challenges, there is initial evidence that using NLP to automate adaptive feedback on learners' written solutions concerning an online task about climate change leads learners to revise their solutions, which improved the quality of their justifications (Zhu et al., 2017; Zhu, Liu, & Lee, 2020).

In summary, we assume that compared to providing an expert solution as a form of non-adaptive, static feedback, NLP-based automatic adaptive feedback may be employed to support pre-service teachers in simulation-based learning of diagnostic reasoning in terms of the quality of their justifications. To what extent NLP-based adaptive feedback can also advance diagnostic accuracy has hardly been investigated.

1.2.3. Individual and collaborative learning in simulation-based learning

The need for adaptive feedback on pre-service teachers' learning of diagnostic reasoning in simulation-based learning may differ depending on the social form of learning. Learners seem better in the critical evaluation of other's arguments than their own arguments (Mercier & Sperber, 2017). Throughout collaboration, learners can use their partners as resources in negotiating meaning and as an additional source of feedback (Weinberger, Stegmann, & Fischer, 2010), by adaptively correcting each other and filling the collaboration partner's knowledge gaps. Therefore, collaborative learners' need for adaptive feedback in diagnostic reasoning may be lower compared to individual learners. However, there is also evidence that collaborative learners show higher learning gains with adaptive feedback (Chuang & O'Neil, 2012; Hsieh & O'Neil, 2002). As collaborative learners might be affected by transaction costs, because they need cognitive capacity for interacting, expressing thoughts, and monitoring another's understanding, they might be particularly in danger of cognitive overload in complex tasks like diagnostic reasoning (Dillenbourg, 2002; Janssen & Kirschner, 2020). Thus, the effects of adaptive feedback on germane cognitive load might be even more pronounced in collaborative learning contexts where cognitive resources are taxed through additional *collaboration load* (Kirschner, Paas, & Kirschner, 2009).

Regarding diagnostic reasoning, research comparing collaborative processing of diagnostic reasoning tasks with individual processing indicates differences in the processing approaches. These different approaches may affect diagnostic accuracy and the quality of justifications: on the one hand, collaborative learners generate more hypotheses and evaluate more evidence before suggesting a solution (Csanadi et al., 2021) and they apply a more reflective approach by processing the task information under different perspectives (Okada & Simon, 1997). Individuals, on the other hand, seem more determined in proposing solutions (Csanadi et al., 2021).

Based on these contrasting findings regarding the benefits and costs of collaboration – particularly in the field of simulation-based learning (Cook et al., 2012, 2013) – it is difficult to derive directed hypotheses. However, it is plausible to assume that the effects of different kinds of feedback *differ* depending on whether learners learn alone or together. Thus, we hypothesise an interaction of feedback type with the social form of learning (i.e., an undirected hypothesis).

1.3. The present study

In this study, we employ an automatic adaptive feedback algorithm in a simulation setting that is based on NLP methods. The algorithm was implemented to provide feedback on pre-service teachers' written explanations of their diagnostic reasoning about simulated pupils with learning difficulties. In this context, diagnostic reasoning is expressed in diagnostic accuracy (i.e., whether or not diagnoses are correct), and in the quality of diagnostic justifications (i.e., the extent to which relevant supporting pieces of evidence for the diagnoses are presented). We investigate effects of automatic adaptive NLP-based feedback compared to static feedback (i.e., expert solutions). Automatic adaptive feedback provides process-related feedback and can foster learners' understanding of their current state of knowledge and suggest where and how to improve current performance (Narciss et al., 2014; Plass & Pawar, 2020; Wisniewski et al., 2019). We hypothesise that automatic adaptive feedback is more effective than static feedback in fostering learners' diagnostic reasoning in the learning process (see Zhu et al., 2017; Zhu et al., 2020; Hypothesis 1a for diagnostic accuracy; 1b for the quality of justification). We further investigate whether potential effects of automatic adaptive feedback might interact with the social form of learning, that is, whether pre-service teachers learn individually or collaboratively. On the one hand, adaptive feedback might have higher impact for individual learners than for collaborators since collaborative learners can provide each other with adaptive feedback during the task processing (Weinberger et al., 2010). On the other hand, collaborators' needs for adaptive feedback might be higher compared to individual learners because of transaction costs (Janssen & Kirschner, 2020). We hypothesise an interaction of the social form of learning and the type of feedback on diagnostic reasoning in the learning process (Hypothesis 2a for diagnostic accuracy; 2b for the quality of justification). Further, we hypothesise a positive effect of automatic adaptive feedback on learners' diagnostic reasoning skills in a post-test (Hypothesis 3a for diagnostic accuracy; 3b for the quality of justification). We also hypothesise an interaction effect of the social form of learning and the type of feedback on diagnostic reasoning skills in a post-test (Hypothesis 4a for diagnostic accuracy; 4b for the quality of justification).

2. Method

2.1. Sample and design

A total of $N = 178$ pre-service teachers for primary school and higher track secondary school from a German university participated in the study. We recruited the pre-service teachers by advertising in lectures, in online courses, and on campus. The study utilised a 2X2 between-subjects experimental design. Learners were randomly assigned to one of four conditions: they received either static or adaptive feedback and worked either collaboratively or individually. Sixty pre-service teachers learned individually, half of them received static feedback while the other half received adaptive feedback. The remaining 118 pre-service teachers were randomly grouped into dyads. One dyad was excluded, because one of the partners turned out not to be a pre-service teacher. Half of the dyads received static feedback; the other half received adaptive feedback. The adjusted sample size is $N = 118$ units (60 individual learners and 58 dyads).

The 137 (77%) women and 39 (22%) men were on average 23.34 years old ($SD = 3.58$, min = 18, max = 35) and their study semester varied from semester 1 to 16 ($M = 5.84$, $SD = 3.73$). The distribution regarding age and gender in our study is comparable with the population of pre-service teachers in Germany.

2.2. Learning environment and the learners' task

The study was conducted using the computer-based learning platform CASUS (https://www.instruct.eu/casus/), on which pre-service teachers learned with simulated pupil cases. The simulated pupils are constructed building child profiles with various learning difficulties. In the learning phase, the learners worked on six simulated pupil cases, which are available in an open science repository https://osf.io/knfmu/. Three of the cases were concerned with children with specific learning difficulties with impairment in reading and/or writing (dyslexia or isolated reading or spelling disorder). The other three cases dealt with diseases from the spectrum of Attention-Deficit and/or Hyperactivity Disorder (ADD or ADHD). We used document-based simulations (Heitzmann et al., 2019): the learners had access to different types of materials that described the behaviour of the simulated pupil. The material included a transcript of a conversation between the teacher and the parents of the child, the pupil's school assignments and certificates, and a description of the pupil's learning and social behaviour. Learners decided how many information sources they examined and in which order. Following each case, pre-service teachers wrote an explanation of their diagnostic reasoning. After that, the learners received either automated or static feedback on their written explanation. A detailed explanation of the learning environment can be found in Bauer et al. (2022).

2.3. Manipulation of independent variables

Depending on the experimental condition, the learning environment provided either static or adaptive feedback and the pre-service teachers learned either individually or collaboratively.

2.3.1. Static and adaptive feedback

Static feedback: After learners had entered and justified their diagnosis, they received an expert solution of the case and were asked to compare it with their own solution. Two independent domain experts validated the expert solutions prior to their use in the study. An example of static feedback is shown in Fig. 1.

Automatic adaptive feedback: Learners' diagnostic explanation was analysed in real-time using NLP: we applied a sequence labelling approach to identify diagnostic classes, consisting of diagnostic entities (e.g., reading problems, hyperactivity) and diagnostic activities (hypothesis generation, evidence generation, evidence evaluation, and drawing conclusions), in pre-service teachers' written explanations. To automatically and adaptively provide feedback on learners' current diagnostic reasoning, a system consisting of three components (NeuralWeb, INCEpTION, and CASUS) was implemented (for in-depth explanation see Pfeiffer et al., 2019). First, in a "cold-start" phase, domain experts coded explanations written by learners of a prior study with $N = 118$ pre-service teachers, who worked on the same simulations in the learning environment CASUS (see Bauer et al., 2020). The experts used the annotation platform INCEpTION (https://inception-project.github.io/) and coded the data according to diagnostic entities and epistemic activities (for details see Schulz, Meyer, & Gurevych, 2019). Second, the coded data was used to initially train a predictive model in NeuralWeb, a Python-based web service. Third, the written explanations of new learners, who participated in the present study, were processed through the NeuralWeb model to output a label-set of discrete diagnostic classes (diagnostic entities and epistemic activities). In a nutshell, we utilised state-of-the-art artificial neural network-based models for sequence tagging (see Akbik et al., 2019), specialised for the setting of diagnostic reasoning (see Pfeiffer et al., 2019), which have been shown to outperform standard baselines for these types of tasks (for a comparison of alternative models see Schulz, Meyer, & Gurevych, 2019).

Depending on the automatically identified classes, specific paragraphs of predefined feedback text were adaptively activated. The feedback paragraphs were created by domain experts and validated by independent domain experts prior to the study. Experts created approximately 40 feedback paragraphs for every case. For example, these feedback paragraphs informed the learner that a specific symptom was correctly identified in the simulated pupil case. When the corresponding element of a pupil's profile was not detected in a learner's written explanation, the feedback informed the learner that they missed mentioning that symptom. The identified classes and the automatic adaptive feedback, consisting of a range of different feedback paragraphs, were sent back to the learning environment CASUS. CASUS then presented this adaptive feedback to the user.

The automatic adaptive feedback targets two levels of the learner's written explanation of their diagnostic reasoning: diagnostic activities

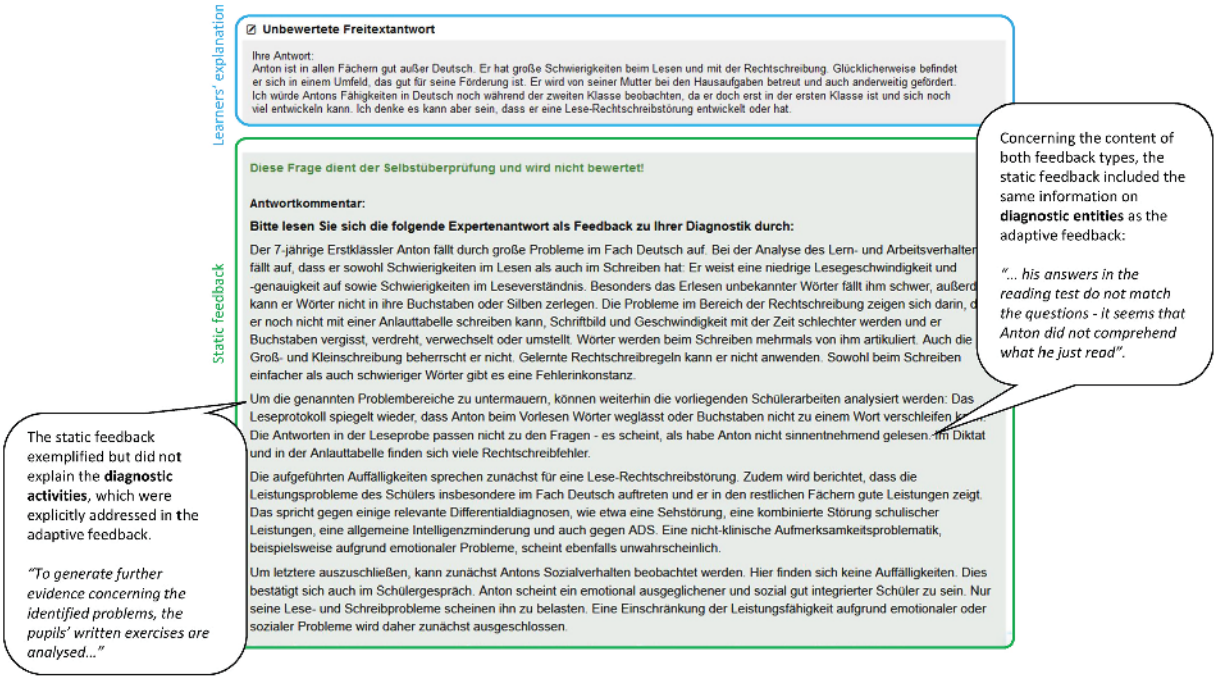


Fig. 1. Static feedback in CASUS.

(whether appropriate reasoning activities were applied or missing) and diagnostic entities (whether the chosen diagnosis and its justification are correct, incorrect, or missing in terms of the domain-specific and case-specific content). By clicking on feedback paragraphs, the relating text part, which was identified in their answer, was highlighted (see Fig. 2).

After the feedback, learners were presented with the next case. There was no opportunity for a revision of the initial written explanation after the feedback.

2.3.2. Individual and collaborative learning

In the individual condition, every learner worked on their own in a computer lab. In the collaborative condition, two learners worked together as a dyad. The two partners each worked on their own computer in separate computer labs and communicated via headsets. We lightly structured the collaboration of the dyads by a scene level script (see Vogel, Wecker, Kollar, & Fischer, 2017). We assigned two roles to the learners, namely main user and secondary user. The main user actively operated in the CASUS learning environment and shared their screen with the collaborating partner. The secondary user was able to advise and discuss with the main user (for details see Kiesewetter et al., 2022). After three of the six learning cases, the roles changed. In the beginning of the study, collaborative learners were explicitly told to exchange ideas with their learning partner and to write their explanation together.

2.4. Procedure

On average the learners needed $M = 170.17$ min to complete the study ($SD = 31.59$ min). At first, we introduced the participants to the learning platform CASUS by a short video. Second, the pre-service teachers completed a questionnaire containing demographics and a conceptual knowledge test. Third, they watched an 18-min video with theoretical input about specific learning difficulties included in the simulation to compensate for possible differences in prior knowledge. Fourth, the participants worked on the six learning cases (individually or collaboratively and receiving static or adaptive feedback). After three of the six learning cases, participants had a 10-min break. After that, participants completed the remaining three learning cases. Lastly, the participants completed a post-test with two unsupported cases; these were cases similar to the learning cases in the learning phase and also concerned with pupils with learning difficulties (dyslexia and ADD), however, without any feedback, and without a learning partner. For their participation, the participants received 50 Euros as compensation.

2.5. Measures

2.5.1. Prior conceptual knowledge

Conceptual knowledge, which refers to knowledge about concepts and their interrelations in a certain domain, is considered a necessary prerequisite for successful diagnostic reasoning (Heitzmann et al., 2019). We operationalised prior conceptual knowledge about learning disorders with 14 questions about reading and writing difficulties as well as behavioural disorders from the spectrum of ADHD or ADD. The questions were single choice questions with four answer options and one right answer option for each question. Choosing the correct answer was awarded with one point. For dyads, which we used as one unit of analysis for the collaborative learning setting, we calculated the mean score of both collaborators for every item. Further, we calculated the mean score of the 14 items to operationalise it. As the conceptual knowledge represented several potentially independent areas (e.g., ADHD, dyslexia) rather than only one, we assume that the scale reflects a formative instead of a reflective construct and thus we will report the variance inflation factor (VIF). A VIF statistic for formative constructs should be lower than 3.3, meaning that less than 70% of the indicator's variance is explained by the other indicators (Stadler, Sailer, & Fischer, 2021). The items for prior conceptual knowledge showed no variance inflation, indicating an appropriate measurement as a formative construct ($VIF_{min} = 1.20$; $VIF_{max} = 1.78$).

2.5.2. Diagnostic accuracy

We used the written explanations of learners' diagnostic reasoning as the data source to determine leaners' diagnostic accuracy in each case. Based on the expert solutions, we developed a coding scheme to operationalise diagnostic accuracy for each learning case and each post-test

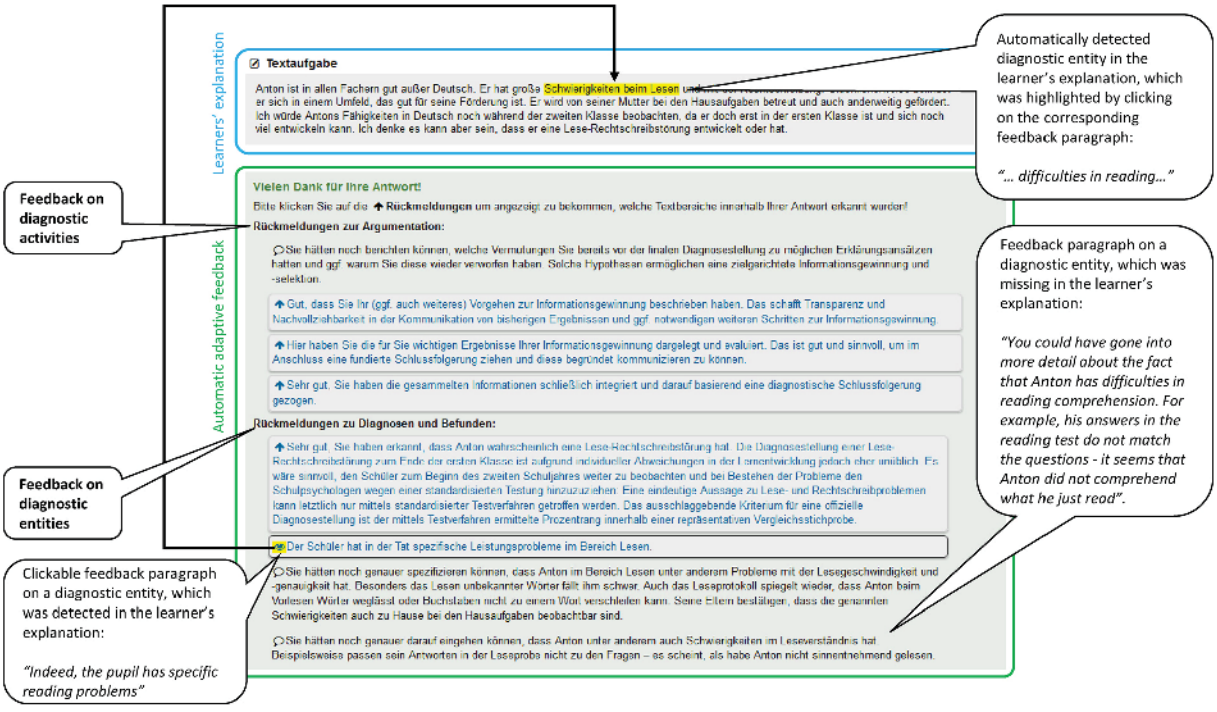


Fig. 2. Automatic adaptive feedback in CASUS.

case. We used one category per case, which represented the presence (coded as 1) or absence (coded as 0) of the correct diagnosis in the learners' explanation. Two trained coders independently coded the written explanations of 11 individual learners and nine dyads (16.9% of the overall data) regarding diagnostic accuracy. The written explanations used in the training were from learners of all four conditions. The inter-rater agreement for diagnostic accuracy, assessed with Cohen's kappa, was high ($\kappa = 0.95$). The remaining material was coded individually.

Diagnostic accuracy in the learning process consisted of five categories, each indicating the presence of the correct diagnoses in the written explanation of learning cases numbered two to six. The first learning case was not included in the learning process measurement as the learners received the first feedback after the completion of the cases. We calculated the mean score of the frequencies of these five categories to operationalise diagnostic accuracy in the learning process. The reliability of this variable was acceptable (McDonald's $\omega = 0.61$).

Post-test diagnostic accuracy consisted of two categories, each indicating the presence of the correct diagnoses in the written explanation of both unsupported post-test cases. All participants solved the two post-test cases individually. To obtain the post-test diagnostic accuracy for the dyads, we calculated the mean value for every category for every dyad. Further, we calculated the mean score of the frequency of the two categories to operationalise post-test diagnostic accuracy. These two categories showed a significant correlation of $r = 0.43$ ($p < .001$), indicating sufficient internal consistency.

2.5.3. Quality of justification

To determine learners' quality of justifications in each case, we used the written explanations of learners' diagnostic reasoning as the data source. We developed a coding scheme, which is based on expert solutions, to operationalise the quality of diagnostic justifications. We used six categories for each case, which indicated the presence (coded as 1) or absence (coded as 0) of the six primary supporting pieces of evidence for the correct diagnosis. Based on experts' solutions, employing all six pieces of evidence in a case is considered a high-quality justification. Again, two trained coders independently coded the written explanations of 11 individual learners and nine dyads (17% of the overall data) regarding the quality of justifications. Cohen's kappa for the quality of justification was high ($\kappa = 0.91$).

We used the described coding of the learners' explanations to operationalise the quality of justification in both the learning process and the post-test quality of justification.

The quality of justification in the learning process was operationalised by 30 categories. We used learners' explanations from five learning cases (learning cases two to six), in which we used six categories representing the presence or absence of the primary supporting pieces of evidence for the correct diagnosis in each learning case. As the learners received the first feedback after completing the first learning cases, we excluded the first case for the measurement of the quality of justification in the learning process. We calculated the mean score of the 30 categories. The reliability of the variable is acceptable (McDonald's $\omega = 0.75$).

Post-test quality of justification consisted of 12 categories, which were assigned in the learners' explanations from two unsupported post-test cases that all students completed individually at the end of the study. In each of these two post-test cases we used six categories, which were coded for the presence or absence of supporting evidence in the corresponding case. To obtain the quality of justification values for the dyads, we calculated the mean value of each of the 12 categories of every dyad. Then, we calculated the mean score for the 12 categories for the quality of justifications in the post-test. The reliability of the variable was rather low (McDonald's $\omega = 0.50$).

2.5.4. Time-on-task

As a control measure, we included time-on-task, which is the time that learners spent with the learning material (e.g., report of grades, description of a pupil's social behaviour), not including the time they spent with writing the explanation and reading the feedback. We measured time-on-task for all six learning cases and computed a sum score.

2.6. Statistical analyses

To investigate Hypotheses 1(a, b) and 2(a, b) we calculated a MANCOVA, using the type of feedback and the social form of learning as independent variables. The diagnostic accuracy and the quality of justifications in the learning process were dependent variables. To control for prior conceptual knowledge, we used this variable as a covariate. To account for individual differences in time-on-task, we included it as a covariate as well. To account for non-normal distribution, time-on-task values were log transformed (see Van der Linden, 2016). Hypotheses 3 (a, b) and 4(a, b) were also tested by a MANCOVA. We included the same independent variables and covariates in this MANCOVA as in the MANCOVA for Hypotheses 1(a, b) and 2(a, b), however this time we used post-test diagnostic accuracy and the post-test quality of justifications as dependent variables. We analysed the data with IBM SPSS Statistics 26 and set the alpha level to $\alpha = 0.05$.

2.7. Ethics clearance

The study was approved by the Medical Faculty's Ethics Committee of Ludwig-Maximilians-Universität München (no.17-249).

3. Results

3.1. Prior conceptual knowledge (Randomisation check)

Descriptive results of the participants' prior conceptual knowledge in the four conditions of the 2X2 design are shown in Table 1. The descriptives indicate comparable levels of prior knowledge across all conditions. An ANOVA with the independent variables type of feedback and social form of learning did not reveal indications for systematic apriori differences (type of feedback, $F(1,114) = 0.09$, $p = .76$, $\eta_p = 0.001$, social form of learning, $F(1,114) = 0.66$, $p = .42$, $\eta_p = 0.006$, and interaction type of feedback X social form of learning, $F(1,114) = 2.00$, $p = .16$, $\eta_p = 0.017$). Thus, the randomisation was successful with respect to prior conceptual knowledge. A correlation matrix of all variables is included in Appendix 1.

3.2. Effects on diagnostic reasoning in the learning process

To analyse effects of automatic adaptive feedback, the social form of learning as well as their interaction on diagnostic reasoning outcomes in the learning process, we conducted a MANCOVA. In this analysis we included log transformed time-on-task, (diagnostic accuracy, $F(1,112) = 11.34$, $p = .001$, $\eta_p = 0.092$, quality of justification, $F(1,112) = 55.34$, $p < .001$, $\eta_p = 0.331$), and prior conceptual knowledge as covariates (diagnostic accuracy, $F(1,112) = 0.22$, $p = .643$, $\eta_p = 0.002$, quality of justification, $F(1,112) = 1.51$, $p = .222$, $\eta_p = 0.013$). The model explains 32% of variance in diagnostic accuracy and 47% of variance in the quality of justifications in the learning process.

Table 1
Means (*M*) and standard deviations (*SD*) for prior conceptual knowledge split by conditions of the 2X2 design.

Social form of learning	Type of feedback	*M*	*SD*
individual	static	.70	.12
	adaptive	.68	.13
collaborative	static	.69	.09
	adaptive	.72	.08

Regarding the *diagnostic accuracy* in the learning process, we found a significant medium-sized interaction effect of the type of feedback and the social form of learning, $F(1,112) = 13.28$, $p < .001$, $\eta_p = 0.106$, supporting Hypothesis 2a. Based on the estimated marginal means in Table 2 and Bonferroni-corrected post hoc comparisons, we conclude that collaborative learners, who have received static feedback scored significantly lower on diagnostic accuracy compared to all other groups. No other post hoc comparisons were significant (see Appendix 2). As the interaction is disordinal the main effects for type of feedback, $F(1,112) = 16.06$, $p < .001$, $\eta_p = 0.125$, and social form of learning, $F(1,112) = 12.41$, $p = .001$, $\eta_p = 0.100$, cannot be meaningfully interpreted. Thus, Hypothesis 1a, in which a main effect of automatic adaptive feedback was hypothesised, was not supported.

Regarding the *quality of justification* in the learning process, we found a large main effect of automatic adaptive feedback, $F(1,112) = 34.07$, $p < .001$, $\eta_p = 0.233$. The main effect of the social form of learning on the quality of justification in the learning process was not significant, $F(1,112) = 0.94$, $p = .336$, $\eta_p = 0.125$, neither was the interaction effect of social form of learning and type of feedback, $F(1,112) = 0.69$, $p = .407$, $\eta_p = 0.006$. These results are in support of Hypothesis 1b as we found a positive effect of adaptive feedback on the quality of justifications. Hypothesis 2b was not supported in our analysis as we found no interaction effect of social form of learning and type of feedback on the quality of justification in the learning process.

3.3. Effects on diagnostic reasoning skills in the post-test

To investigate effects of automatic adaptive feedback, the social form of learning as well as its interaction on diagnostic reasoning skills in the post-test we used a MANCOVA, controlling for log transformed time-on-task, (diagnostic accuracy, $F(1,112) = 13.57$, $p < .001$, $\eta_p = 0.108$, quality of justification, $F(1,112) = 16.42$, $p < .001$, $\eta_p = 0.128$), and prior conceptual knowledge, (diagnostic accuracy, $F(1,112) = 1.95$, $p = .166$, $\eta_p = 0.017$, quality of justification, $F(1,112) = 0.12$, $p = .732$, $\eta_p = 0.001$). The model explains 31% of variance in post-test diagnostic accuracy and 28% of variance in the post-test quality of justifications.

We found a significant medium-sized interaction effect of the type of feedback and the social form of learning on post-test *diagnostic accuracy*, $F(1,112) = 7.96$, $p = .006$, $\eta_p = 0.066$, supporting Hypothesis 4a. As for the learning process, also in the post-test, learners that collaborated in the learning phase and received static feedback reached significantly lower levels of diagnostic accuracy in the post-test than learners in all other conditions (see Table 2). Again, no other post hoc comparisons were significant (see Appendix 2). Because of the disordinal interaction, the main effects for type of feedback, $F(1,112) = 14.60$, $p < .001$, $\eta_p = 0.115$, and social form of learning, $F(1,112) = 17.66$, $p < .001$, $\eta_p = 0.136$, on post-test diagnostic accuracy cannot be meaningfully interpreted. Hypothesis 3a, in which a main effect of automatic adaptive feedback was hypothesised, was not supported.

Results regarding the post-test *quality of justifications* show a large significant main effect of automatic adaptive feedback, $F(1,112) = 21.02$, $p < .001$, $\eta_p = 0.158$. The social form of learning, $F(1,112) = 1.28$, $p = .260$, $\eta_p = 0.011$, and the interaction between type of feedback and social form of learning, $F(1,112) = 0.01$, $p = .923$, $\eta_p < 0.001$, were not significant. These results support Hypothesis 3b based on the main effect of automatic adaptive feedback. As we did not find an interaction for the post-test quality of justifications, Hypothesis 4b was not supported.

Table 2
Estimated marginal means (*M*) and standard errors (*SE*) for all dependent variables split by conditions of the 2X2 design.

Dependent Variable	Social form of learning	Type of feedback	*M*	*SE*
Diagnostic accuracy in the learning process	individual	static	.54	.05
		adaptive	.55	.05
	collaborative	static	.20	.05
		adaptive	.56	.05
Quality of justifications in the learning process	individual	static	.40	.02
		adaptive	.51	.02
	collaborative	static	.36	.02
		adaptive	.50	.02
Post-test diagnostic accuracy	individual	static	.74	.06
		adaptive	.80	.06
	collaborative	static	.32	.06
		adaptive	.71	.06
Post-test quality of justifications	individual	static	.33	.02
		adaptive	.43	.02
	collaborative	static	.35	.02
		adaptive	.45	.02

Note: Estimated means and standard errors with covariates on the following values: log transformed time-on-task = 10.16, prior conceptual knowledge = .70.

4. Discussion

In an experimental study, we investigated the effects of NLP-based adaptive feedback on diagnostic reasoning in individual and collaborative simulation-based learning. Methods of NLP using sophisticated algorithms of artificial neural networks allow for automatic analysis of written texts of learners, which facilitates providing process-oriented automatic feedback in real-time and without the need to involve a human corrector. However, to implement such NLP-based systems, training data is initially required: In this study, data from a prior study of 118 learners was used (see Bauer et al., 2020). These data were fully coded regarding the aspects we provided feedback for. Thus, developing and implementing NLP-based automatic adaptive feedback is a time-consuming and extensive task, which may, however, pay off when it comes to preparing many pre-service teachers for their upcoming challenge of diagnosing learning difficulties in pupils in their daily business as teachers.

4.1. Automatic adaptive feedback fosters learners' quality of justifications

Results showed that adaptive feedback fostered the pre-service teachers' quality of justifications in written assignments for both individual and collaborative learners. Adaptive feedback might have facilitated learners' comparison of their current performance to a desired goal performance in their diagnostic reasoning (e.g., Bimba et al., 2017; Narciss et al., 2014; Plass & Pawar, 2020). Compared with the static feedback, the demand on learners' cognitive resources for processing the feedback might have been reduced by adaptive feedback, which may have helped to improve the quality of justification (see Sweller et al., 2019). Even in simulations, which are designed to be less cognitively taxing than real-life situations, diagnostic reasoning is a complex task requiring a high amount of cognitive resources for information processing. Especially pre-service teachers, who lack prior knowledge and professional experience with respect to learning difficulties (Poznanski et al., 2021), are in danger of cognitive overload. However, for diagnostic reasoning and particularly for elaborated justifications, pre-service teachers need optimal conditions to be able to invest enough cognitive resources into actual learning processes (Sweller et al., 2019).

4.2. Effects of automatic adaptive feedback might Depend on the type of task

The results concerning diagnostic accuracy indicate that adaptive feedback did not outperform static feedback per se; instead, we found an interaction between feedback and the social form of learning on diagnostic accuracy. Collaborative learners receiving static feedback had a significantly lower diagnostic accuracy than learners from the other three experimental conditions. Collaborative learners receiving adaptive feedback and individual learners in both feedback conditions did not differ significantly in their diagnostic accuracy. Individual learners may

not require the same amount of adaptive support to relate their own diagnoses with a correct diagnosis as needed when considering various pieces of evidence to provide high-quality justification. Thus, the information provided in the static feedback may have already been sufficient to foster individual learners' diagnostic accuracy. A reason for this might be found in the complexity of the tasks underlying diagnostic accuracy and the quality of justifications: accuracy requires a diagnostic decision that might be fairly simple to derive in some cases – especially when a limited number of potential diagnoses (here: learning difficulties) is previously introduced to the learners, like in our study. The justification might be more cognitively demanding because of the broad range of evidence, for which learners have to evaluate relationships and interdependencies as relevant or irrelevant in a specific case. In the study, adaptive feedback was more effective for tasks which involved a variety of aspects that have to be considered at a time (such as multiple evidence for a high-quality justification), compared to tasks that require a single decision (such as concluding an accurate diagnosis) – especially when this decision is not too difficult.

4.3. Automatic adaptive feedback is helpful for collaborators' diagnostic decision making

For collaborative learning, the results indicated that compared with static feedback, adaptive feedback particularly facilitated collaborative learners' diagnostic accuracy. In the static condition, learners' cognitive capacities have already been more challenged than in the adaptive condition because learners had to compare their own performance with the expert solution. Additionally, transaction costs like communicating, social regulation, and coordinating might have induced additional collaboration load (see Janssen & Kirschner, 2020; Kirschner et al., 2009). This might especially affect dyads that collaborate for the first time and probably will not do so with the same collaboration partner in the future, such as the pre-service teachers in our laboratory study. In such situations, the chances for making use of the potentials of a collective working memory, which involves also knowledge about the other collaborators, are rather low (Janssen & Kirschner, 2020). Combined with evidence that collaborative learners tend to be less pragmatic in proposing solutions in reasoning tasks compared to individual learners (Csanadi et al., 2021), collaborative learners who received static feedback might have struggled most in proposing a solution in form of a diagnostic decision in their written explanations. Instead, they may have focused more strongly on evaluating evidence without finalising a diagnostic conclusion in explaining their diagnostic reasoning. For evaluating pieces of evidence to construct a high-quality justification, the benefits of collaboration, like the adoption of multiple perspectives (Okada & Simon, 1997), might balance the cognitive capacity disadvantages that occur particularly in the diagnostic accuracy outcomes. Further, compared to the static expert solution, the automatic adaptive feedback may have had a stronger compensating effect, possibly by explicitly scaffolding the learners to determine and explicate a shared diagnosis.

4.4. Limitations and future research

The results may be limited in their generalisability with respect to the diagnostic tasks involved. As there were relatively few possible diagnoses and the associated cues of these diagnoses were fairly similar, the decision task itself might not have been taxing the cognitive resources of the participating pre-service teachers to the extent that more complex decisions would. Further, effects on the two different outcomes, particularly diagnostic accuracy, might also be influenced by the choice of our sample: we included pre-service teachers in our study, who are typically not the ones making formal diagnoses about learning difficulties (i.e. diagnostic accuracy). Instead, they often provide evidence for or against certain differential diagnoses (i.e. quality of justification) to school psychologists or special education teachers. Thus, the pre-service teachers in our study might rather have focussed on the evaluation of different evidences than on a final diagnosis. Another possible limitation to generalisability may be that we used dyads to represent collaborative contexts (Jensen & Wiley, 2006); it will be interesting to investigate the effects of adaptive feedback for bigger groups. With respect to internal validity, there may have been issues because the instruction did not specifically emphasise the necessity to state the diagnosis, but asked the learners to justify their diagnosis, therefore only asking for the diagnosis implicitly. However, independently of inaccuracy or absence of the diagnosis, adaptive feedback – compared with static feedback – helped improve collaborative learners' performance in writing a congruent explanation of their diagnostic reasoning that includes a conclusion regarding the diagnosis. A further limitation relates to the relatively low reliabilities of the measurement of diagnostic accuracy. Potential reasons for that are the different areas of learning difficulties that we included in the study as well as the rather low number of categories.

Future studies might measure outcomes like diagnostic accuracy with a larger amount of codes or items. To do so, a larger number of cases in simulations that require less time for the learners to solve might help to get a more reliable measurement of the diagnostic accuracy outcomes. Future research might further investigate the interaction of the type of feedback and the social form of learning to explore different hypotheses that underlie the interaction effect found in this study. In this regard, the interaction of the type of feedback with other aspects of simulation-based learning may be further investigated, such as the complexity of the cases and tasks that need to be performed while processing the cases, which may also affect the need for specific types of feedback.

To address issues with external validity of the results, field studies with larger samples may be conducted to further specify relevant contextual conditions in which automatic adaptive feedback in simulation-based learning are most effective. Such studies could make use of existing groups, e.g., within university courses, and further investigate our findings of automatic adaptive feedback in collaborative settings as well as the acceptance of automatic adaptive feedback in practice. Acceptance of automatic adaptive feedback might critically depend on the trust of users in the feedback system as well as the transparency of the NLP and feedback system (see Shin, 2021). To ensure end-users' understanding of how our NLP and feedback system works, we implemented a transparent system that highlights detected components (see Fig. 2), making us optimistic about the field use of our simulation with respect to learners' acceptance and trust. In addition, in field studies, learners could be asked to utilise the feedback provided in order to revise their explanations and thus, to take an increasingly active role in their learning (see Zhu et al., 2020).

5. Conclusions

Automatic adaptive feedback in simulation-based learning can be used effectively to foster pre-service teachers' diagnostic reasoning in HE. Using methods of NLP like the algorithms related to artificial neural networks to automate adaptive feedback seems to provide particular benefits in terms of fostering learning more complex reasoning outcomes, such as justifying a diagnosis – even in short term interventions of the length of one single course session. Adaptive feedback is a promising instructional support to help pre-service teachers improve the quality of justifications in written assignments, independent of whether they learn together or alone. In collaborative learning contexts, adaptive feedback rather than static seems to effectively compensate for collaboration costs that lead to performance drops with respect to diagnostic accuracy. However, training and specialisation of artificial neural network-based NLP models is a time-consuming task, as it requires collection of data sets and elaborate manual coding before actual implementation of the automatic adaptive feedback. These efforts might be worthwhile where automatic adaptive feedback is subsequently

implemented in simulations in large programmes, such as teacher education or medical education. In such contexts, automatic adaptive feedback can offer a convenient solution for providing elaborate, process-oriented feedback in real-time to high numbers of students.

Funding

This research was supported by a grant of the German Federal Ministry of Research and Education (Grant No.: 16DHL1040) and by the Elite Network of Bavaria (K-GS-2012-209). We have no conflicts of interest to disclose.

CRediT authorship contribution statement

Michael Sailer: Conceptualization, Formal analysis, Visualization, Investigation, Methodology, Resources, Writing – original draft, Writing – review & editing. **Elisabeth Bauer:** Investigation, Methodology, Resources, Visualization, Software, Writing – original draft, Writing – review & editing. **Riikka Hofmann:** Validation, Writing – review & editing. **Jan Kiesewetter:** Conceptualization, Methodology, Writing – review & editing. **Julia Glas:** Formal analysis, Investigation, Software, Writing – original draft. **Iryna Gurevych:** Conceptualization, Methodology, Funding acquisition, Writing – review & editing. **Frank Fischer:** Conceptualization, Funding acquisition, Project administration, Writing – review & editing.

Acknowledgement

The authors would like to thank Gabrielle Arenge for the proofreading of the manuscript. Further, the authors would like to thank the whole FAMULUS team.

Appendix. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2022.101620.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). Flair: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics* (pp. 54–59). https://doi.org/10.18653/v1/N19-4010

Bauer, E., Fischer, F., Kiesewetter, J., Shaffer, D. W., Fischer, M. R., Zottmann, J. M., & Sailer, M. (2020). Diagnostic activities and diagnostic practices in medical education and teacher education. *An Interdisciplinary Comparison. Frontiers in Psychology, 11*, 2787. https://doi.org/10.3389/fpsyg.2020.562665

Bauer, E., Sailer, M., Kiesewetter, J., Schulz, C., Gurevych, I., Fischer, M. R., & Fischer, F. (2022). Learning to Diagnose Students' Behavioral, Developmental and Learning Disorders in a Simulation-Based Learning Environment for Pre-Service Teachers. In F. Fischer, & A. Opitz (Eds.), *Learning to Diagnose with Simulations* (pp. 97–107). Springer. https://doi.org/10.1007/978-3-030-89147-3_8.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26–55. https://doi.org/10.1002/sce.20286

Bimba, A. T., Idris, N., Al-Hunaiyyan, A., Mahmud, R. B., & Shuib, N. L. B. M. (2017). Adaptive feedback in computer-based learning environments: A review. *Adaptive Behavior, 25*(5), 217–234. https://doi.org/10.1177/1059712317727590

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems, 29*, 4349–4357.

Charlin, B., Lubarsky, S., Millette, B., Crevier, F., Audétat, M. C., Charbonneau, A., et al. (2012). Clinical reasoning processes: Unravelling complexity through graphical representation. *Medical Education, 46*(5), 454–463. https://doi.org/10.1111/j.1365-2923.2012.04242.x

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

Chuang, S. H., & O'Neil, H. F. (2012). Role of task-specific adapted feedback on a computer-based collaborative problem-solving task. In H. F. O'Neil, & R. S. Perez (Eds.), *Web-based learning: Theory, research, and practice* (pp. 239–254). New York: Routledge.

Codreanu, E., Sommerhoff, D., Huber, S., Ufer, S., & Seidel, T. (2020). Between authenticity and cognitive demand: Finding a balance in designing a video-based simulation in the context of mathematics teacher education. *Teaching and Teacher Education, 95*, Article 103146. https://doi.org/10.1016/j.tate.2020.103146

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

Cook, D. A., Hamstra, S. J., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., … Hatala, R. (2013). Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. *Medical Teacher, 35*(1), 867–898. https://doi.org/10.3109/0142159X.2012.714886

Csanadi, A., Kollar, I., & Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: Better together? *European Journal of Psychology of Education, 36*(1), 147–168. https://doi.org/10.1007/s10212-020-00467-4

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In P. A. Kirschner (Ed.), *Three worlds of CSCL. Can we support CSCL?* (pp. 61–91). Heerlen: Open Universiteit Nederland.

Graesser, A. C., Hu, X., & Sottilare, R. (2018). Intelligent tutoring systems. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 246–255). New York, NY: Routledge.

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

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

Henderson, M., Ryan, T., & Phillips, M. (2019). The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education, 44*(8), 1237–1252. https://doi.org/10.1080/02602938.2019.1599815

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

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 328–339).

Hsieh, I.-L. G., & O'Neil, H. F. (2002). Types of feedback in a computer-based collaborative problem-solving group task. *Computers in Human Behavior, 18*(6), 699–715. https://doi.org/10.1016/S0747-5632(02)00025-0

Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: Towards a research agenda. *Educational Technology Research & Development, 68*(2), 783–805. https://doi.org/10.1007/s11423-019-09729-5

Jensen, M. S., & Wiley, J. (2006). When three heads are better than two. In *Proceedings of the annual meeting of the cognitive science society, 28*. Retrieved from https://escholarship.org/uc/item/9160x87s.

Kiesewetter, J., Hege, I., Sailer, M., Bauer, E., Schulz, C., Platz, M., & Adler, M. (2022). A usability study for implementing remote collaboration in a virtual patient platform. *JMIR Medical Education*. https://doi.org/10.2196/24306

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

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

Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A.-K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education, 100*(4), Article 103298. https://doi.org/10.1016/j.tate.2021.103298

Lachner, A., Jarodzka, H., & Nückles, M. (2016). What makes an expert teacher? Investigating teachers' professional vision and discourse abilities. *Instructional Science, 44*(3), 197–203. https://doi.org/10.1007/s11251-016-9376-y

Manning, C. D., & Schütze, H. (2005). In *Foundations of statistical natural language processing* (8th ed.). Cambridge, MA: MIT Press.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge, Massachusetts: Harvard University Press. https://doi.org/10.4159/9780674977860

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., et al. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education, 71*, 56–76. https://doi.org/10.1016/j.compedu.2013.09.011

Okada, T., & Simon, H. A. (1997). Collaborative discovery in a scientific domain. *Cognitive Science, 21*(2), 109–146. https://doi.org/10.1016/S0364-0213(99)80020-2

Pfeiffer, J., Meyer, C. M., Schulz, C., Kiesewetter, J., Zottmann, J., Sailer, M., Bauer, E., Fischer, F., Fischer, M. R., & Gurevych, I. (2019). *FAMULUS: Interactive Annotation and Feedback Generation for Teaching Diagnostic Reasoning* (pp. 73–78). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-3013.

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

Poznanski, B., Hart, K. C., & Graziano, P. A. (2021). What do preschool teachers know about attention-deficit/hyperactivity disorder (ADHD) and does it impact ratings of

child impairment? *School Mental Health*, *13*(1), 114–128. https://doi.org/10.1007/s12310-020-09395-6

Reinke, W. M., Stormont, M., Herman, K. C., Puri, R., & Goel, N. (2011). Supporting children's mental health in schools: Teacher perceptions of needs, roles, and barriers. *School Psychology Quarterly*, *26*(1), 1–13. https://doi.org/10.1037/a0022714

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, *38*(1), 1–37. https://doi.org/10.1111/cogs.12086

Scheuer, O., McLaren, B. M., Loll, F., & Pinkwart, N. (2012). Automated analysis and feedback techniques to support and teach argumentation: A survey. *Educational Technologies for Teaching Argumentation Skills*, 71–124. https://doi.org/10.2174/978160805015411201010071

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

Schulz, C., Meyer, C. M., & Gurevych, I. (2019). Challenges in the automatic analysis of students' diagnostic reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6974–6981. https://doi.org/10.1609/aaai.v33i01.33016974.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology*, *60*, Article 100832. https://doi.org/10.1016/j.newideapsych.2020.100832

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

Van Merriënboer, J. J. G. (2013). Perspectives on problem solving and instruction. *Computers & Education*, *64*, 153–160. https://doi.org/10.1016/j.compedu.2012.11.025

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, *47*, 513–539. https://doi.org/10.1146/annurev.psych.47.1.513

Van der Linden, W. J. (2016). Lognormal response-time model. In W. J. van der Linden (Ed.), *Handbook of item response theory*, *1* pp. 289–310). Boca Raton, FL: Chapman & Hall/CRC Press. models.

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

Weinberger, A., Stegmann, K., & Fischer, F. (2010). Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). *Computers in Human Behavior*, *26*(4), 506–515. https://doi.org/10.1016/j.chb.2009.08.007

Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648–1668. https://doi.org/10.1080/09500693.2017.1347303

Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*. https://doi.org/10.1016/j.compedu.2019.103668

5 General Discussion

This thesis aimed to (a) develop a cross-disciplinary research perspective on diagnostic reasoning, (b) integrate and refine the existing understanding of diagnostic reasoning skills, and (c) investigate approaches to facilitate the learning of diagnostic reasoning skills. The ideas underlying the aim of developing a cross-disciplinary research perspective on diagnostic reasoning in medical education and teacher education were discussed in the general introduction (see section 1). Based on these ideas, the first study (see section 2) made an empirical contribution to cross-disciplinary research on diagnostic reasoning by comparing diagnostic activities and diagnostic practices in medical education and teacher education. The second study (see section 3) focused on the second aim of refining the existing understanding of diagnostic reasoning skills and investigated the proposed distinction of diagnostic argumentation and diagnostic judgment as two different diagnostic reasoning skills in the context of teacher education. The third study (see section 4) focused on the third aim of investigating approaches to facilitating the learning of diagnostic reasoning skills and analyzed the effects of automatic adaptive feedback and collaborative learning on preservice teachers' simulation-based learning of diagnostic reasoning skills.

The results of all three studies are briefly summarized below (see section 5.1). Thereafter, the theoretical implications of the findings are discussed (see section 5.2). Subsequently, the practical implications are derived (see section 5.3), the relevant limitations of the findings are discussed (see section 5.4), and suggestions for future research are offered (see section 5.5).

5.1 Summary of the Results

5.1.1 Results of Study 1

The first study (Bauer et al., 2020; see section 2) compared the diagnostic activities (Heitzmann et al., 2019; see section 1.3.2) and diagnostic practices (Bauer et al., 2020; see section 1.3.3) in medical education and teacher education. Data from 142 learners from medical education and 122 learners from teacher education were analyzed. The learners diagnosed eight cases from their respective field in a simulation-based learning environment and wrote a report of their approach to diagnostic reasoning for each case.

According to an analysis of the reported diagnostic activities (RQ1 of study 1, see section 1.6.1), both medical students and preservice teachers focused mainly on the diagnostic activity of evaluating evidence. The medical students focused more on the activities of generating hypotheses and drawing conclusions, whereas the preservice teachers placed more emphasis on the activity of generating evidence. The results supported the prior assumption

that there are significant differences regarding the relative emphasis on each diagnostic activity between medical education and teacher education.

The differences in the use of diagnostic activities were also salient in the overall diagnostic practices (RQ2 of study 1, see section 1.6.1), which was depicted using the novel method of ENA (Shaffer, 2017). Overall, the networks depicting the diagnostic practices of the participating medical students and preservice teachers revealed a similar structure. Yet, the relative frequencies of the diagnostic activities' co-occurrences and thus the overall diagnostic practices were found to be significantly different between medical education and teacher education. The medical students demonstrated a more hypothesis-driven or explanation-driven approach (see Coderre et al., 2010; Kiesewetter et al., 2013; Seidel & Stürmer, 2014), whereas the preservice teachers demonstrated a more data-driven or description-driven approach (see Gräsel & Mandl, 1993; Kiesewetter et al., 2013; Norman et al., 2007; Seidel & Stürmer, 2014). Moreover, by comparison, the variability in the diagnostic practices in teacher education turned out to be higher than in medical education, which may indicate a lower standardization in diagnostic practices or lower knowledge of diagnostic standards in teacher education. The significant differences in the fields' diagnostic practices suggest that there are differences in their epistemic ideals and standards (see Chinn et al., 2011) regarding diagnostic reasoning. For example, medical education's higher emphasis on generating hypotheses might reflect its standards related to differential diagnosing (e.g., Kassirer, 2010; see section 5.2.2).

5.1.2 Results of Study 2

The second study (Bauer et al., 2022; see section 3) explored preservice teachers' diagnostic argumentation regarding the three suggested facets of justification, disconfirmation, and transparency (see section 1.4.4) and the proposed distinction of diagnostic argumentation and diagnostic judgment (e.g., Loibl et al., 2020) as two diagnostic reasoning skills (see section 1.4.3). For this purpose, the text data collected for the first study in teacher education was reanalyzed and complemented by data about the accuracy of preservice teachers' diagnostic judgments as well as of their prior conceptual and strategic diagnostic knowledge.

Analyzing the occurrences and relations of justification, disconfirmation, and transparency in preservice teachers' diagnostic argumentations (RQ1 of study 2, see section 1.6.2) revealed that justification may be an antecedent in preservice teachers' diagnostic argumentation, whereas disconfirmation or transparency might indicate a more advanced form of diagnostic argumentation. Disconfirmation, in particular, seemed to be unidirectional,

dependent on justification in diagnostic argumentation. Only a few diagnostic argumentations involved disconfirmation without justification (i.e., the listing of differential diagnoses without evaluating evidence or finalizing conclusions), whereas argumentations addressing different diagnoses usually made evidence-based conclusions. In contrast, preservice teachers offered one confirmatory and final diagnosis, without considering and disconfirming competing explanations. This unidirectional dependency explains the significant correlation between justification and disconfirmation, which was found in the correlational analysis. In contrast, transparency was not correlated with the other two facets. Overall, the findings suggested that justification, disconfirmation, and transparency are distinguishable facets of diagnostic argumentation and may even represent distinct reasoning subskills.

The three facets also differed in the patterns by which conceptual and strategic diagnostic knowledge predicted them (RQ2 of study 2, see section 1.6.2): Justification was predicted by both conceptual diagnostic knowledge about diagnoses and evidence as well as strategic diagnostic knowledge about diagnostic approaches. The disconfirmation of the differential diagnoses was only predicted by conceptual diagnostic knowledge about diagnoses. In contrast, transparency concerning the undertaken approaches to generate evidence was only predicted by strategic diagnostic knowledge about diagnostic approaches. The findings further supported that the three facets may represent distinct reasoning subskills. However, the amounts of variance in the three facets, which were explained by conceptual and strategic diagnostic knowledge, were generally rather low. Therefore, further research may address the role of other variables in explaining justification, disconfirmation, and transparency, such as knowledge about standards in diagnostic reasoning and argumentation (see section 5.2.3).

Investigating the relationship between making accurate diagnostic judgments and formulating justified, disconfirmatory, and transparent diagnostic argumentations (RQ3 of study 2, see section 1.6.2) supported the assumption that diagnostic judgment and diagnostic argumentation might be considered two different diagnostic reasoning skills. An ENA network comparison revealed the overall significant differences between argumentation texts addressing accurate judgments and argumentation texts addressing inaccurate judgments. Yet, a correlational analysis of justification, disconfirmation, transparency, and diagnostic accuracy indicated a significant correlation of justification in diagnostic argumentation and the accuracy of diagnostic judgments, which was not explained by a joint basis of conceptual and strategic diagnostic knowledge. Variables other than conceptual and strategic diagnostic knowledge may be relevant in explaining the relation between justification in diagnostic

argumentation and accurate diagnostic judgments, such as the proportion of controlled and intuitive information processing during diagnostic reasoning (see section 5.2.3).

5.1.3 Results of Study 3

The third study (Sailer et al., 2023; see section 4) investigated the effects of NLP-based automatic adaptive feedback (e.g., Bimba et al., 2017) and collaborative learning (e.g., Csanadi et al., 2021) in simulation-based learning environments (e.g., Chernikova, Heitzmann, Stadler et al., 2020) on the accuracy of preservice teachers' diagnostic judgments and their quality of justification in their diagnostic argumentations. The results indicated positive effects of adaptive feedback on the quality of justifications of both individual and collaborative learners (RQ1b & RQ2b of study 3, see section 1.6.3), suggesting that learners' comparison of their current performance to a desired goal performance might have been facilitated by adaptive feedback (e.g., Bimba et al., 2017; Narciss et al., 2014; Plass & Pawar, 2020).

However, concerning the achievement of accurate diagnostic judgments, adaptive feedback did not exhibit a higher benefit for learners per se (RQ1a of study 3, see section 1.6.3). Instead, the results indicated an interaction between adaptive feedback and collaborative learning (RQ2a of study 3, see section 1.6.3). The collaborative learners who received static feedback achieved significantly lower diagnostic accuracy compared to the collaborative learners who received adaptive feedback; in contrast, individual learners' diagnostic accuracy did not differ when they received static or adaptive feedback. The findings indicate that collaborative learners seemed to have a higher need for the additional support provided by the adaptive feedback for making accurate diagnostic judgments but not for formulating high-quality justifications. The interaction effect between the type of feedback and the social form of learning on making accurate diagnostic judgments requires further explanation (see section 5.2.4).

Summarizing the overall findings of the third study, the results confirmed the expected positive effect of the NLP-based automatic adaptive feedback compared to static feedback on preservice teachers' accuracy of diagnostic judgments and the quality of justification in their diagnostic argumentations.

5.2 Theoretical Implications

5.2.1 Advancing Cross-Disciplinary Research on Diagnostic Reasoning

One of the central aims of this thesis was to develop a cross-disciplinary research perspective on diagnostic reasoning, with a focus on medical education and teacher education. To integrate the theoretical and empirical accomplishments from both fields and advance the

existing understanding of the involved knowledge and skills as well as their learning, this thesis built on the work of Heitzmann et al. (2019) and discussed the commonalities and differences in diagnostic reasoning in medical education and teacher education. In doing so, the thesis further systematized the related arguments and elaborated on them with respect to diagnostic problems (see section 1.2), epistemic processing (see section 1.3), and cognitive processes (see section 1.4). The following section consolidates the conclusions made concerning the commonalities and differences in diagnostic reasoning in medical education and teacher education; the theoretical implications derived and integrated from the presented studies are discussed in subsequent sections (see sections 5.2.2, 5.2.3, and 5.2.4).

Diagnostic problems in medical education and teacher education can be analyzed with respect to different characteristics, such as content area, exemplarity, complexity, and required activities to solve diagnostic problems (see section 1.2). These characteristics can vary across diagnostic problems, not only across different fields but also within the same field. Cross-disciplinary research in diagnostic reasoning should consider variations in the characteristics of diagnostic problems because, as will be further explained in the following sections, such variations might systematically influence diagnostic reasoning.

The differences in the content areas of diagnostic problems imply that professionals need knowledge about problem-specific concepts and strategies to solve diagnostic problems (i.e., conceptual and strategic diagnostic knowledge; see Förtsch et al., 2018; Kolovou et al., 2021; Wimmers et al., 2007; see section 1.4.1). Because of the different content areas, diagnostic problems as well as problem-specific diagnostic knowledge in medical education and teacher education need to be considered content-specific, which generally limits their comparability.

There are, however, other characteristics of diagnostic problems with regard to which research can either investigate the effects of problem characteristics on diagnostic reasoning or consider them when investigating diagnostic reasoning with respect to epistemic processing and cognitive processes. A still content-related characteristic of diagnostic problems, which is yet better suited for systematically describing the differences of diagnostic problems, is the exemplarity of a diagnostic problem to a professional context. Exemplarity refers to how likely professionals will encounter the respective diagnostic problem in their professional environments (see Kolodner, 1992). Consequently, high exemplarity increases the likelihood that professionals have already collected experience regarding the respective problem. The problem characteristic of exemplarity is thus linked to certain cognitive processes in diagnostic reasoning, especially expertise development (i.e., knowledge encapsulation and

script formation; see Schmidt & Rikers, 2007): With repeated encounters of a diagnostic problem, diagnostic knowledge is increasingly encapsulated into higher-level concepts and integrated into episodic representations of diagnostic problems (i.e., script formation; e.g., Charlin et al., 2007). Expertise development, in turn, affects information processing. The option of recognizing familiar patterns facilitates fast and subconscious non-analytic information processing, which saves cognitive resources, as opposed to conscious, and thus effortful, analytic information processing (see Evans, 2008; Norman et al., 2017). Besides expertise development, the structural problem characteristic of complexity – that is, the amount and connectivity of information that needs to be processed – also affects the cognitive effort required to analyze a problem (see Campbell, 1988; Robinson, 2001; Stadler et al., 2019; Sweller, 2010). The cognitive processes involved in diagnostic reasoning may be considered non-specific to a field (Heitzmann et al., 2019; see Kirschner et al., 2017). Therefore, the respective research findings may be considered transferable across different fields, while considering the characteristics of the investigated individuals (e.g., level of expertise) and diagnostic problems (e.g., exemplarity or complexity) as well as the situational characteristics (e.g., collaborative diagnostic reasoning, which requires explicating reasons toward a collaborating professional).

Besides complexity, another structural characteristic of diagnostic problems is the required activities to solve the problem. For example, a diagnostic problem might be already given (e.g., stated by the patient or a colleague) or it might be necessary to initially identify the problem (e.g., during classroom observations). The required activities are especially relevant to be considered while researching the problem-solving process in diagnostic reasoning, either regarding the application of diagnostic reasoning skills (i.e., the judgment process or in-process argumentation) or epistemic processing in diagnostic reasoning. In solving diagnostic problems, individuals engage in diagnostic activities (such as generating hypotheses, generating evidence, evaluating evidence, and drawing conclusions; see section 1.3.2; Fischer et al., 2014; Heitzmann et al., 2019). Comparing the diagnostic activities in medical education and teacher education, the specific content of concrete hypotheses, evidence, and conclusions varies; for instance, a physician's hypothesis about the causes for a patient's back pain varies from a teacher's hypothesis about the causes for a pupil's writing difficulties. However, because the intended epistemic aim of achieving diagnostic accuracy is the same, it was assumed that the purpose of each of the diagnostic activities is conceptually transferable across different fields (see Hetmanek et al., 2018): Irrespective of the specific content of a hypothesis, the activity of generating hypotheses holds the purpose of identifying

potential explanations, which may require further investigation (see Appendix B for further information about the coding scheme for diagnostic activities). Because diagnostic activities – that is, epistemic activities – are established by and within epistemic communities, individuals within that community feature a shared understanding of when and how epistemic activities need to be performed (see Kelly, 2008; Leont'ev, 1978; Roth & Lee, 2006). This shared understanding shapes a collective pattern of epistemic activities, which is a community's epistemic practices (see Kelly, 2008; Roth & Lee, 2006). The epistemic practices of different epistemic communities can involve specifics that relate to their epistemic ideals and standards (criteria to assess the achievement of aims; e.g., whether a diagnosis is grounded on valid and convincing evidence) and processes that are considered reliable (e.g. how to generate valid and convincing evidence; see Duncan & Chinn, 2016). Comparing medical education and teacher education with respect to their diagnostic activities and diagnostic practices can indicate differences that might provide insights into field-specific ideals, standards, and processes. Systematically investigating the field-related specifics in diagnostic reasoning also contributes to advancing the cross-disciplinary research on diagnostic reasoning in medical education and teacher education.

5.2.2 Standards in Diagnostic Reasoning

To advance the outlined cross-disciplinary research perspective on diagnostic reasoning, the first study in this thesis compared diagnostic activities and diagnostic practices in medical education and teacher education. Medical education and teacher education were found to significantly differ in their relative emphasis on each diagnostic activity and their overall diagnostic practices, suggesting that there are also differences with respect to the fields' epistemic ideals and standards in diagnostic reasoning (see Chinn et al., 2011).

Medical students' stronger focus on generating and testing hypotheses (see Fischer et al., 2014) contributes to an overall rather hypothesis-driven or explanation-driven diagnostic practice (see Coderre et al., 2010; Kiesewetter et al., 2013; Seidel & Stürmer, 2014). The results pattern can be regarded as reflecting the agreed-upon standards in medical education related to differential diagnosing, which is considered ideal for ensuring a reliable process in medical education (see Chinn et al., 2011; Kassirer, 2010). The ideal is systematically put into practice by incorporating it on a systemic level, involving guidelines and university curricula (e.g., Weinstein & Pinto-Powell, 2016). Teaching differential diagnosing to future physicians in their medical programs introduces the standard to the level of individual learners, who internalize the standard with repeated practice and start to act accordingly in professional situations that require diagnostic reasoning (see Clark et al., 2008; Louie et al., 2007; Roth &

Lee, 2006). By adopting the approach of differential diagnosing, medical students in turn contribute to the overall collective pattern of diagnostic activities, which can be characterized as hypothesis-driven diagnostic practice.

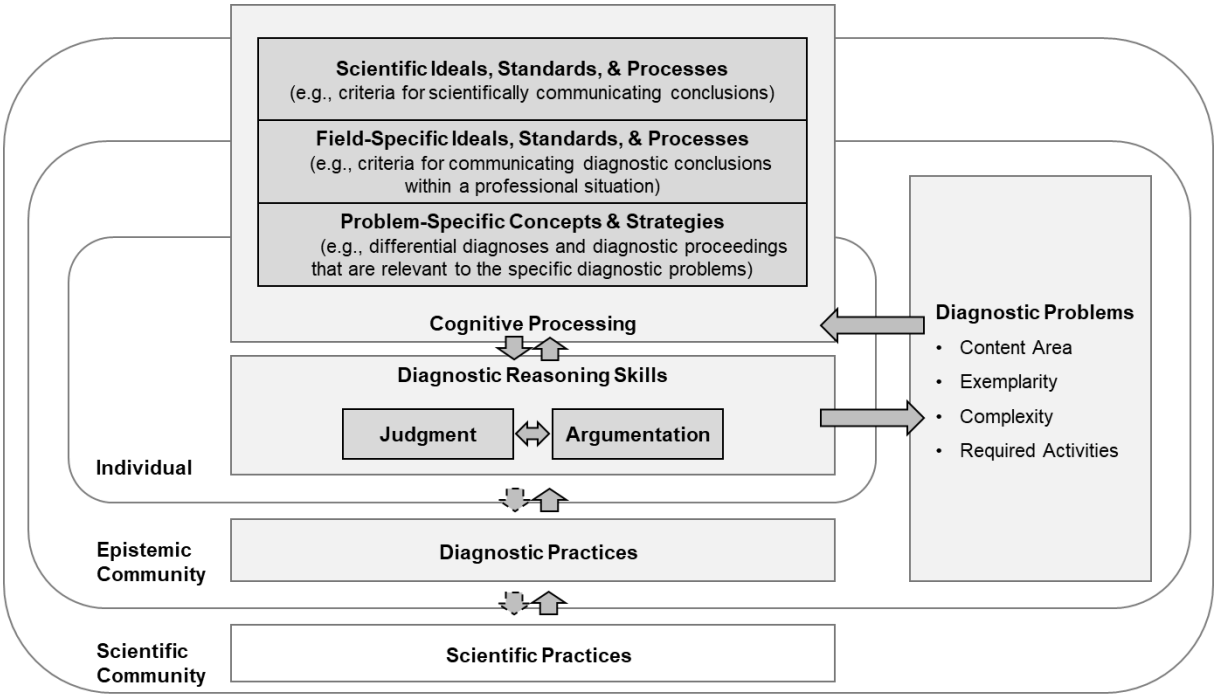
In teacher education, there seems to be no such widespread and specific standard for approaching diagnostic reasoning. However, there might be considerable variations between different national educational systems (Bauer & Prenzel, 2012). Research in teacher education was described as being "still a relatively young field" (Grossman & McDonald, 2008, p. 184). Therefore, the development of standards for diagnostic reasoning might be less progressed compared to medical education. The finding that, compared to medical students, preservice teachers seemed to exhibit greater variability in their individual realizations of an overall diagnostic practice supports the assumption of lower standardization in diagnostic reasoning in teacher education (see Scales et al., 2018). However, despite this greater variability, preservice teachers still had a tendency toward a data-driven or description-driven approach in their diagnostic reasoning. Their preference for a data-driven approach might be grounded in ideals, which has so far remained rather implicit in teacher education. After teacher education studies repeatedly reported about the biases in teachers' diagnostic reasoning (e.g., Südkamp et al., 2012; Urhahne & Wijnia, 2021), some teacher education programs reacted by orienting their teaching toward the concept of professional vision (Goodwin, 1994), which emphasizes the need to focus on describing observations before explaining them (see Seidel & Stürmer, 2014). This development might represent a critical progress in evolving and systematically teaching standards for diagnostic reasoning and, thus, further standardizing diagnostic practices in teacher education. Furthermore, there might be other rather implicit ideals complementing the development of standards for diagnostic reasoning in teaching and teacher education, such as to avoid being judgmental toward students (see Aalberts et al., 2012). Overall, the findings might therefore be considered to reflect the specific differences in epistemic ideals concerning diagnostic reasoning in medical education and teacher education, which are implemented in higher education and thus found in the fields' diagnostic practices (e.g., Kassirer, 2010; Weinstein & Pinto-Powell, 2016).

In summary, the first study generated evidence for the idea that comparing different fields with respect to diagnostic activities and diagnostic practices can provide insights into field-specific epistemic ideals, standards, and processes in diagnostic reasoning (see Chinn et al., 2011). In doing so, the study also contributed to advancing a cross-disciplinary research perspective on diagnostic reasoning in medical education and teacher education. The two fields can be considered reflecting epistemic communities: They seem to have developed

different approaches in performing diagnostic activities to collect and integrate information to reduce uncertainty and to make and communicate informed and justifiable decisions in professional situations that require diagnostic reasoning (see Heitzmann et al., 2019; Kelly, 2008). Individuals who become part of an epistemic community by attending a medical education program or a teacher education program internalize the ideals, standards, and processes of the respective field. In the course of the internalization process, they assimilate their understanding of suitable approaches in solving diagnostic problems (see Clark et al., 2008; Osberg, 2005). The individuals who internalize the ideals, standards, and processes of their field start to adapt their intuitive approaches and perform diagnostic activities in solving diagnostic problems accordingly (see Roth & Lee, 2006). Internalizing the field-specific epistemic ideals, standards, and processes may thus be considered part of the professional development and learning relevant to the situational performance of diagnostic reasoning skills (see Figure 2).

Figure 2

Framework for Cross-Disciplinary Research on Diagnostic Reasoning in Medical Education and Teacher Education



5.2.3 Diagnostic Argumentation as a Distinct Diagnostic Reasoning Skill

To refine the existing understanding of diagnostic reasoning skills, the second study investigated the proposed differentiation of diagnostic reasoning skills in terms of diagnostic judgment and diagnostic argumentation in the context of teacher education. The overall results pattern seems to support the assumption that diagnostic judgment (indicated by

diagnostic accuracy) and diagnostic argumentation (indicated by the three proposed facets of justification, disconfirmation, and transparency) represent different diagnostic reasoning skills. There was, however, a relation between accurate diagnostic judgment and justification in diagnostic argumentation, which was not explained by a joint basis of conceptual and strategic diagnostic knowledge (see Förtsch et al., 2018; Heitzmann et al., 2019). Another variable that could potentially explain the relation between accurate diagnostic judgment and justification in diagnostic argumentation might be the prevailing type of information processing during diagnostic reasoning (see Loibl et al., 2020). The literature about dual process theories (e.g., Kahneman, 2011; for an overview of several different theoretical accounts see Evans, 2008) suggests that controlled information processing results in more conscious and explicable reasoning compared to intuitive, unconscious information processing (i.e., pattern recognition; e.g., Norman et al., 2007). Whereas both types of information processing can result in making accurate judgments, diagnostic argumentation requires engaging in conscious and explicable controlled information processing (see section 1.4.3). Thus, controlled information processing could precede accurate diagnostic judgments (Coderre et al., 2010) and simultaneously facilitate justification in diagnostic argumentation.

Besides distinguishing between diagnostic judgment and diagnostic argumentation as two diagnostic reasoning skills, the findings also suggested distinguishing between justification, disconfirmation, and transparency as different subskills of diagnostic argumentation. In particular, this interpretation is grounded on the differences in the patterns by which the three facets were predicted by conceptual and strategic diagnostic knowledge (see Förtsch et al., 2018). Justification involves evaluating evidence, making warranted connections, and drawing conclusions about diagnoses, which seems to build on both conceptual diagnostic knowledge about diagnoses and evidences as well as strategic diagnostic knowledge about diagnostic proceedings to conclude and reject different diagnoses (see Fischer et al., 2014; Toulmin, 1958). By comparison, disconfirmation primarily seems to rely on conceptual diagnostic knowledge about differential diagnoses to address the differential diagnoses and thus show that alternative explanations have been considered in diagnostic reasoning (see Lawson, 2003; Toulmin, 1958). In contrast, transparency seems to require strategic diagnostic knowledge of which informational sources can deliver critical evidence when diagnosing a specific problem in order to describe the processes undertaken to generate evidence (see Chinn et al., 2011; Fischer et al., 2014).

Yet, significant amounts of variance in justification, disconfirmation, and transparency remained unexplained by conceptual and strategic diagnostic knowledge, raising the question

of which additional types of knowledge and skills might be relevant in explaining justification, disconfirmation, and transparency in diagnostic argumentation. One further explanatory variable might be knowledge about standards in diagnostic reasoning and, in particular, diagnostic argumentation (see Bauer et al., 2020; Chinn et al., 2011). Because of the limited explicit agreement about ideals and standards in diagnostic reasoning in the context of teacher education (as compared with some other fields, such as medical education; see Bauer et al., 2020) and the resulting lack of systematic teaching in teacher education programs, preservice teachers are not knowledgeable about any standards in diagnostic argumentation. Therefore, discussing and developing standards for diagnostic argumentation may advance teacher education and facilitate future teaching of diagnostic reasoning skills. To this end, field-specific ideals may be further researched and considered relevant for diagnostic reasoning. However, considering that field-specific standards in diagnostic argumentation should comply with fundamental norms and standards that have developed in the broader context of scientific argumentation (see Figure 2; see also section 1.4.4; e.g., Bricker & Bell, 2008), justification, disconfirmation, and transparency may serve as a starting point from which to systematize and teach standards for diagnostic argumentation in teacher education.

Furthermore, considering the performance differences that were found in justification, disconfirmation, and transparency, a second explanatory variable might be relevant to consider, which is argumentation skills that are transferrable across domains (Hetmanek et al., 2018). Hetmanek et al. (2018) suggested that cross-domain transferable skills could, to some extent, compensate for a lack of more specific knowledge (e.g., knowledge about standards in diagnostic argumentation). Comparing the prevalence of the three facets in diagnostic argumentation indicates a higher prevalence of justification as compared with disconfirmation and transparency. It is possible that justification is a subskill, that is easier to compensate for by cross-domain transferable argumentation skills, than disconfirmation and transparency. In summary, one may hypothesize that apart from having conceptual and strategic diagnostic knowledge, cross-domain transferable argumentation skills and knowledge about standards in diagnostic argumentation are additional variables that might be relevant for explaining justification, disconfirmation, and transparency in preservice teachers' diagnostic argumentation.

Because of the non-field-specific nature of cognitive processing (see section 1.4.2) and epistemic aims (see sections 1.3.1 and 1.4.3), the results pattern found in the context of teacher education should be replicable in the context of medical education. This requires further investigation. Replicating the results in medical education would further support the

interpretation of diagnostic judgment and diagnostic argumentation as two distinct diagnostic reasoning skills.

5.2.4 Facilitating Diagnostic Reasoning Skills

To investigate approaches to facilitate the learning of diagnostic reasoning skills, the third study compared the effects of automatic adaptive feedback vs. static feedback (i.e., expert solutions) in collaborative vs. individual simulation-based learning on the accuracy of diagnostic judgment and the quality of justification in diagnostic argumentation in the context of teacher education. Overall, the results suggested that adaptive feedback is particularly beneficial for fostering the quality of learners' justification in diagnostic argumentation, whereas collaborative learning seemed to pose an additional challenge for learners' achievement of diagnostic accuracy.

Although simulations are simplified representations of professional situations (Heitzmann et al., 2019; Sauvé et al., 2007), they still confront learners with high amounts of information and are associated with high demands on learners' cognitive resources (i.e., the working memory). To cope with the complexity of the simulated problems and still invest cognitive resources into the actual learning processes of learning diagnostic reasoning skills, learners need sufficient support and feedback (i.e., germane cognitive load; Sweller et al., 2019). By facilitating learners' comparison of their current performance to a desired goal performance (see Figure A3 in Appendix D), adaptive feedback might have freed cognitive capacities (i.e., the working memory) for the actual task performance and learning processes (see Moreno, 2004; Narciss, 2008; Sweller et al., 2019). Compared to adaptive feedback, the unsupported presentation of a large amount of information in the static feedback (see Figure A4 in Appendix D) condition might potentially have exceeded learners' working memory capacity and, thus, restrained their performance and learning (see Sweller et al., 2019).

More specifically, the additional support provided by adaptive feedback seemed to be particularly relevant for performing and thus learning justification in diagnostic argumentation. Diagnostic argumentation requires conscious, controlled information processing (see section 1.4.3; e.g., Kahneman, 2011) of the available evidences and their relations. Performing diagnostic argumentation can thus be considered limited by the available working memory capacity (see Evans, 2008; Kahneman, 2003; Kalyuga, 2011). Therefore, freeing working memory capacity by means of adaptive feedback might be particularly beneficial for performing and learning diagnostic argumentation. In contrast, regarding individual learners' achievement of diagnostic accuracy, adaptive feedback did not outperform static feedback. This finding may be explained by the assumption that diagnostic

judgment can not only be performed by consciously analyzing the causal relations between information but also by recognizing familiar patterns of information and making one single, intuitive decision for the seemingly most suitable diagnosis (see section 1.4.2 and 1.4.3; e.g., Evans, 2008; Norman et al., 2007; Schmidt & Rikers, 2007). Facilitating the making of intuitive judgments seems to depend less on elaborated feedback or processing information about different evidence and their relations respectively; it might be the case that the learners mainly referred to the information about the correct task solutions, thus requiring less support in processing the information, which is why the static feedback might have already been sufficient (see Hattie & Timperley, 2007; Narciss, 2008). Summarizing the effects of adaptive feedback as compared to static feedback, adaptive feedback was shown to be more effective for performing and learning skills, which involved processing a variety of information at a time (such as multiple evidence for high-quality justification in diagnostic argumentation), compared to tasks that required a single decision (such as making an accurate judgment).

However, in contrast to individual learners, adaptive feedback was revealed to be beneficial for collaborative learners' achievement of diagnostic accuracy. Collaborative learners seemed to have a higher need for the additional support provided by the adaptive feedback to foster their performance and learning of making diagnostic judgments. One reason for this may be that collaborative learning is associated with higher demands on learners' cognitive resources compared to individual learning (Janssen & Kirschner, 2020). Collaborative diagnostic reasoning involves explicating and exchanging reasons while engaged in the process of solving a diagnostic problem, meaning that learners are involved in in-process diagnostic argumentation (Berland & Reiser, 2009; see Rapanta & Felton, 2021). In doing so, collaborative learners analyze evidence as well as their relations with each other and with potential hypotheses before making a diagnostic judgment (Csanadi et al., 2021; Okada, 1997; Roscoe & Chi, 2008). This process requires collaborative learners to engage in controlled information processing while being engaged in the judgment process and is thus associated with high demands on their working memory. Therefore, in contrast to individual learners, collaborative learners might benefit less from the minimized cognitive effort associated with making one single, intuitive decision for the seemingly most suitable diagnosis (see Evans, 2008; Kahneman, 2003; Morewedge & Kahneman, 2010). In addition, collaboration involves costs for interacting, communicating, and coordinating, which induce collaboration load and further increase demands on collaborative learners' cognitive resources compared to individual learning (Janssen & Kirschner, 2020; Kirschner et al., 2009). This might especially affect the dyads that collaborate for the first time and probably will not do so

with the same collaboration partner in the future, such as the preservice teachers in the study. In such situations, the chances for making use of the potentials of a collective working memory, which also involves knowledge about the other collaborators, might be rather low (Janssen & Kirschner, 2020; Radkowitsch et al., 2021). With the higher cognitive demands induced by the collaborative learning situation, less working memory capacity is available for actual task performance and learning processes (i.e., germane load; see Sweller, 2010). Collaborative learners may thus have a higher need for the additional support provided by the adaptive feedback to foster their performance and learning of making diagnostic judgments.

A second reason for collaborative learners' higher need for additional support may be that collaborative learners struggled more in making a final diagnostic judgment than individual learners. Prior research found that collaborative learners are rather reflective in their reasoning processes but less straightforward in proposing solutions (Csanadi et al., 2021; Okada, 1997; Roscoe & Chi, 2008). If collaborative learners struggled in making a final diagnostic judgment, adaptive feedback would have better served as an explicit prompt to finalize and explicate a shared diagnostic judgment; in contrast, collaborative learners receiving static feedback needed to identify this option for improvement by themselves.

Summarizing the effects of the collaborative learning setting as compared to the individual learning setting on the performance and learning of diagnostic judgment, the costs associated with collaboration seemed to have outweighed its benefits. However, regarding the performance and learning of diagnostic argumentation (in terms of constructing a high-quality justification), the benefits of collaboration, such as the adoption of multiple perspectives (Csanadi et al., 2021; Okada, 1997), might have balanced the potential disadvantages concerning learners' cognitive load and task processing.

Furthermore, the results pattern found in the third study provides additional support for the conclusions made based on the second study (see section 5.2.3): The different effects of automatic adaptive feedback and collaborative learning in simulation-based learning environments on preservice teachers' diagnostic accuracy and quality of justification support the idea of differentiating between diagnostic judgment and diagnostic argumentation as two diagnostic reasoning skills. Moreover, the results underline that distinguishing between diagnostic judgment and diagnostic argumentation may be considered practically relevant with respect to teaching, learning, and measuring diagnostic reasoning skills.

5.3 Practical Implications

The theoretical considerations, empirical findings, and interpretations presented in this thesis involve several practical implications for medical education and teacher education.

These implications concern not only the teaching and learning of diagnostic reasoning skills but also the reflection, further development, and teaching of field-specific ideals, standards, and processes in diagnostic reasoning.

In terms of reflecting and further developing field-specific ideals, standards, and processes in diagnostic reasoning, an important finding of the first study is that the development of standards for diagnostic reasoning might be less progressed in teacher education compared to medical education. Therefore, the stakeholders in teacher education, such as educators, researchers, and in-service teachers as representatives of the professional practice, might actively discuss and reflect on the standards in diagnostic reasoning and the underlying epistemic ideals to exchange viewpoints and achieve a broad consensus. The stakeholders in medical education might as well reflect further on their standards in diagnostic reasoning and the underlying epistemic ideals to further increase the awareness of practitioners and systematization in teaching. In particular, research can support these advancements by further investigating the field-specific ideals, the already implemented standards, and the field-specifics in diagnostic practices. In all these efforts, it may be considered that the prior and further developments of field-specific ideals, standards, and processes are to be reflected with respect to their compliance with scientific ideals, standards, and processes. With respect to diagnostic argumentation, the further development of standards may, for example, consider the proposed, scientifically oriented facets of justification, disconfirmation, and transparency as a starting point for further reflecting and developing standards for situations that demand diagnostic argumentation.

Achieving a wide-reaching agreement on the ideals, standards, and processes in diagnostic reasoning in teacher education as well as medical education seems vital for further systematizing the implementation processes of standards in the respective fields. On a systemic level, the implementation of standards may involve developing guidelines and adapting university curricula (e.g., Weinstein & Pinto-Powell, 2016), which need to be realized at the level of higher education programs, especially by providing opportunities for practice (e.g., using simulation-based learning environments). With repeated practice, students can internalize the field-specific epistemic ideals, standards, and processes and start to perform diagnostic activities in solving diagnostic problems accordingly. Subsequently, by engaging in professional action, individuals contribute to the overall collective patterns of diagnostic activities in their fields, which accumulate to diagnostic practices that comply with previously implemented standards. Systematically implementing agreed-upon ideals, standards, and

processes in the teaching of diagnostic reasoning can thus facilitate the standardization of diagnostic practices.

Moreover, learning and internalizing the field-specific ideals, standards, and processes might be considered a part of students' professional development, enabling future professionals from medical education and teacher education to be acknowledged as members of their fields by helping them act accordingly in professional situations that demand diagnostic reasoning.

To solve diagnostic problems and act professionally in situations that require diagnostic reasoning, future professionals need to learn diagnostic reasoning skills in the course of their higher education. As suggested in this thesis, it seems not only important to teach future professionals to make accurate judgments – future physicians and teachers must also be able to explain their interpretations about diagnostic problems and resulting conclusions comprehensibly and persuasively. Therefore, they need to learn how to formulate diagnostic argumentations. Since the findings of the second and third studies suggest that diagnostic judgment and diagnostic argumentation have to be considered different diagnostic reasoning skills, they may be considered separately in designing curricula, setting learning objectives, providing learning opportunities, and conducting assessments in medical education and teacher education.

Regarding the learning of diagnostic argumentation, the results of the second study further suggested that justification, disconfirmation, and transparency might reflect different subskills that differ in their knowledge bases. Further research needs to replicate the findings and investigate which additional knowledge and skills are a relevant base for justification, disconfirmation, and transparency in diagnostic argumentation. The role of knowledge about standards and their internalization, in particular, requires further investigation. Moreover, when aiming to teach diagnostic argumentation – which involves the three proposed facets of justification, disconfirmation, and transparency – to preservice teachers and medical students, each of the three facets might require specific introduction, learning, and assessment.

In terms of learning diagnostic reasoning skills, future professionals are considered to require opportunities to reason about authentic diagnostic problems (Chernikova, Heitzmann, Fink et al., 2020), which can be done by using simulation-based learning environments. In accordance with prior research findings (Cook et al., 2010; Cook et al., 2013; Wisniewski et al., 2020), the results of the third study suggest that, compared to static feedback, adaptive feedback facilitates learners' performance and learning of diagnostic reasoning skills when solving simulated diagnostic problems. Therefore, adaptive feedback is recommended for

supporting the simulation-based learning of diagnostic reasoning skills. In addition, as the results suggest, implementing simulation-based learning in combination with collaboration is rather challenging for learners when it comes to making accurate diagnostic judgments. In light of this, further research on collaborative simulation-based learning of diagnostic reasoning skills may clarify the conditions under which the potential of collaborative learning (especially with respect to in-process dialogic argumentation) can be utilized effectively for facilitating the simulation-based learning of diagnostic reasoning skills.

Since offering adaptive feedback on written text can be resource-intensive for educators, the third study investigated the use of NLP methods. The employed algorithms of artificial neural networks allowed for the automatic analysis of learners' diagnostic argumentation texts, thus facilitating the automatic provision of elaborated adaptive feedback in real-time and without involving a human corrector. However, to implement such NLP-based systems initially requires training data, which needs to be coded based on the aspects to provide feedback for. Thus, developing and implementing NLP-based automatic adaptive feedback is a time-consuming and extensive task, but it is one that may pay off when it comes to preparing a significant number of future professionals for the challenge of performing diagnostic reasoning in real-life professional situations.

5.4 Limitations

The three presented studies involved several methodological aspects, which might be considered potential limitations concerning the preceding interpretations, conclusions, and implications.

The selection of the diagnostic problems for researching diagnostic reasoning in the context of teacher education focused on pupils' deficits in reading and writing as well as behavioral problems, which had to be distinguished as clinically relevant (e.g., because of dyslexia or ADHD) or clinically irrelevant (e.g., inattentiveness because of emotional stress induced by family conditions). These sets of diagnostic problems were selected and designed for the first study to achieve a comparatively high degree of matching with the diagnostic problems from medical education in terms of complexity and required activities (see sections 1.2 and 1.6.1), as well as accountability and responsibility, which determine the situational epistemic value of achieving diagnostic accuracy (see sections 1.3.1 and 1.6.1). Nevertheless, concerning the problem characteristic of exemplarity (see section 1.2), the clinical aspects in the diagnostic problems selected for teacher education may be considered less typical with respect to the frequency of the expected encounters in teachers' everyday routine as compared, for example, to assessing a pupil's level of skill. The selection of diagnostic

problems might therefore limit the generalizability of the findings of all three studies to other areas of diagnostic reasoning in teacher education. However, although the selected diagnostic problems are not the most frequent ones for teachers to reason about, teachers are still regularly confronted with pupils' clinical problems and behaviors. For example, the prevalence rates for ADHD are estimated to range from 5 to 10 percent of pupils, meaning that one to two out of twenty pupils is affected by the respective symptoms and functional impairment (Faraone et al., 2003; Scahill & Schwab-Stone, 2000) that demand specific support to avoid long-term consequences for their school career. Teachers are the first professionals who have the opportunity to identify an existing problem and initiate further action (Reinke et al., 2011; Rothì et al., 2008). However, studies have found that they often report a lack of required knowledge and skills, which is why they feel particularly challenged by identifying pupils' clinical problems, such as ADHD (Mohr-Jensen et al., 2019; Poznanski et al., 2021). Therefore, the selected diagnostic problems can still be considered relatively typical for teachers' everyday practice and teacher education.

Besides the matter of practical relevance, the selection of diagnostic problems for researching diagnostic reasoning in teacher education may relate to the particular concerns regarding the generalizability of specific study findings. With respect to the question of generalizability concerning the finding of a rather data-driven or description-driven diagnostic practice in teacher education, it is reasonable to assume that the findings may replicate in other content areas of teachers' diagnostic reasoning. This is because diagnostic practices and underlying ideals are considered field-specific. However, further research needs to address the replicability of the identified practices in different content areas, which seems relevant to both medical education and teacher education.

Moreover, the selected diagnostic problems could be considered to be limiting the generalizability of the finding that diagnostic judgment and diagnostic argumentation represent distinct diagnostic reasoning skills. However, this thesis considered the conceptualization of both diagnostic judgment (indicated by diagnostic accuracy) and diagnostic argumentation (indicated by justification, disconfirmation, and transparency) nonspecific to the content area of clinical problems. Thus, the result pattern is expected to be replicable in other areas where teachers have to make diagnoses. This could be investigated in further research.

Another limitation concerning the results of the first and second studies may be the study progress of the students from teacher education. The sample consisted of learners from a broad range of semesters. Those learners who were rather new to the field of study might

not have had sufficient time to internalize (explicitly taught or implicitly learned) the ideals and standards and adapt their individual approaches of applying diagnostic activities in accordance with their field's collective diagnostic practice. In contrast, the medical students were in their fifth or a higher semester and had completed significantly more semesters than the preservice teachers. Therefore, the potential effect of the preservice teachers' semester of study on their relative frequency in applying different diagnostic activities was tested. The number of semesters did not correlate with the proportion of the different diagnostic activities. Moreover, the analyses were repeated by comparing the medical students to a subsample of preservice teachers who were in the fifth or a higher semester. The subsample analyses revealed the same patterns of results as the analyses of the full sample. Hence, it seems unlikely that the *a priori* difference in the number of semesters would lead to substantial bias in the results (see supplement of Paper 1 in section 2).

The broad range of study semesters of the students from teacher education could have also affected their level of diagnostic reasoning skills as well as diagnostic knowledge that was assessed in the second study. In the local teacher education program, relevant courses about pupils' clinical problems were not compulsory or bound to a specific semester. Therefore, the sample was considered as representative for the local teacher education program. However, there might be considerable variations between different universities and, in particular, between different national educational systems (Bauer & Prenzel, 2012), which might be addressed by further research.

One limitation concerning the data analyzed in the first paper concerns the inter-rater reliabilities for *generating hypotheses* and *drawing conclusions*, which were relatively low in the teacher education data. This could limit the conclusions that can be drawn about the variability in diagnostic practices of teacher education learners in particular. The coding was done simultaneously with data segmentation and the identified segments were partially very fine-grained (see Appendix B for detailed information about the coding scheme). Most likely, the combined task decreased the overall inter-rater agreement compared to inter-rater agreements for exclusive coding tasks, which are usually reported in other studies. However, for our operationalization of diagnostic practices, we reduced the granularity of the original segments by accumulating the presence or absence of diagnostic activities within one sentence. Because of the higher abstraction level in the segmentation used in the analyses of diagnostic practices, the actual agreement of diagnostic activities on the segmentation level of sentences may be actually higher than the agreement of the individually coded diagnostic activities. Therefore, the inter-rater reliabilities for *generating hypotheses* and *drawing*

conclusions might actually be satisfactory to draw conclusions about the variability in preservice teachers' diagnostic practices.

In both the second and the third paper, there were issues with respect to low internal consistency of diagnostic accuracy. In the second study, the internal consistency limits the interpretability of the results of respective correlation analysis (RQ3), because it may hide further correlations with disconfirmation or transparency that were not observed in the results. In the third study, the low internal consistency of diagnostic accuracy may hide further effects of the interventions, would possibly affect the interpretation of the results pattern. One potential reason for the repeated issue with the internal consistency of diagnostic accuracy are the different areas of diagnostic problems (reading and writing plus behavioral problems) that we included in the study. A second potential reason is the small number of measurement items (simulated diagnostic problems), which is known to cause low internal consistency values (e.g., Monteiro et al., 2020). However, a higher number of measurement items would have overwhelmed the learners, because of the scope, amount of information, and processing time. Moreover, both studies found significant and meaningful results with respect to diagnostic accuracy, such as in study two that diagnostic accuracy correlated with the variables conceptual and strategic diagnostic knowledge, which was in accordance with the theoretically grounded expectation. Therefore, the low internal consistency of diagnostic accuracy might not be a major issue for the studies' interpretations. However, future studies might measure outcomes like diagnostic accuracy with a larger amount of measurement items. To do so, a larger number of simulated diagnostic problems that require less time for the learners to solve might help to get a more reliable measurement of diagnostic accuracy.

In the third study, the conclusions drawn with respect to the cognitive demands imposed by making diagnostic judgments might be limited in their generalizability with respect to the diagnostic tasks involved. As the number of possible diagnoses was relatively low and also the similarity of the diagnoses regarding their associated cues was relatively low, the decision task itself might not have been taxing the cognitive resources of the participating preservice teachers to the extent that more complex decisions would.

With respect to the internal validity of the quality of justifications in study three, there may have been issues because the instruction did not specifically emphasize the necessity to state the diagnosis, but asked the learners to justify their diagnosis, therefore only asking for the diagnosis implicitly. However, independently of inaccuracy or absence of the diagnosis, adaptive feedback – compared to static feedback – helped improve collaborative learners'

performance in writing a congruent explanation of their diagnostic reasoning that includes a conclusion regarding the diagnosis.

5.5 Directions for Future Research

Future research may further investigate the differences, commonalities, and the continuing development of diagnostic practices in and across different fields, such as medical education and teacher education. In particular, the finding of rather data-driven diagnostic practices in teacher education compared to hypothesis-driven diagnostic practices in medical education may be addressed by replication studies that involve diagnostic problems of other content areas. To this end, using matched study designs, as implemented in the first study, seems to be a beneficial research approach to maximize cross-disciplinary comparability. Moreover, using the novel method of ENA (Shaffer, 2017) is recommended as suitable analysis for further research on diagnostic practices. Analyzing co-occurrences of specific instances (such as diagnostic activities) as a basis for creating network graphs, ENA is specifically designed for exploring and comparing individual and collective patterns of epistemic processing, such as comparing diagnostic practices shown by groups of students from different fields. On a related note, there are other fields beyond medical education and teacher education that may be interesting to consider and research in terms of diagnostic reasoning and diagnostic practices. For example, Abele (2018) emphasized the commonalities of diagnostic reasoning in mechatronics education with diagnostic reasoning in medical education and teacher education because “technicians and engineers have to detect causes of malfunctioning machines” (Abele, 2018, p. 134). Researching diagnostic reasoning in additional fields might yield further insights and a deepened understanding of the fields’ commonalities, differences, and their ongoing development of diagnostic practices.

Moreover, the role of field-specific ideals and already implemented standards in diagnostic reasoning requires further investigation. The relation of ideals and standards with diagnostic activities and diagnostic practices, in particular, demands more in-depth research to better understand the commonalities, differences, and ongoing development of diagnostic reasoning and diagnostic practices in different fields. In doing so, research can facilitate the further discussion and development of field-specific standards and further the professionalization of diagnostic reasoning. In addition, future research may address how the knowledge and internalization of field-specific standards affect diagnostic reasoning in general and the performance of diagnostic judgment and diagnostic argumentation in particular.

The knowledge and internalization of field-specific standards, and the role of cross-domain transferable reasoning and argumentation skills in justification, disconfirmation, and transparency may also be investigated, considering the finding that large amounts of variance in diagnostic argumentation remained unexplained by conceptual and strategic diagnostic knowledge.

Future research is also necessary to further validate the ideas and findings that diagnostic argumentation is a diagnostic skill distinct from making accurate diagnostic judgments. In particular, the findings require replication research with respect to other diagnostic problems in teacher education, but the replicability of the results pattern in medical education is also of specific interest. Moreover, it needs to be clarified, which joint predictors might explain the relation between accurate diagnostic judgment and justification in diagnostic argumentation require clarification, because the relation that was found did not seem to be explained by conceptual and strategic diagnostic knowledge. Both accurate diagnostic judgment and justification in diagnostic argumentation might be predicted by the controlled processing of information in solving the cases. However, controlled vs. intuitive information processing is difficult to research. One possible approach to generate further and more specific evidence might be by experimentally manipulating the time pressure in solving brief diagnostic problems (see Evans, 2008; Loibl et al., 2020).

Further investigating the relation of information processing types and diagnostic reasoning skills is also relevant to clarify the proposed mechanisms underlying the interaction effect of adaptive feedback and collaborative learning on diagnostic judgment and diagnostic argumentation. Beyond manipulating time pressure, other experimental variations might also offer interesting insights into the underlying mechanisms, such as the experimental manipulation of diagnostic problem characteristics (e.g., complexity).

In general, the proposed characteristics of diagnostic problems require further investigation to make functional adaptations of problem characteristics in simulations for learning, assessment, and research. Moreover, the effects of the variations of the problem characteristics on diagnostic reasoning and the performance of diagnostic reasoning skills in particular need to be explored in more detail.

Research on collaborative simulation-based learning of diagnostic reasoning skills may clarify the conditions under which the potential of collaborative learning can be effectively utilized for facilitating the simulation-based learning of diagnostic reasoning skills. For example, collaboration scripts (Radkowitsch et al., 2021) may be useful to structure in-

process dialogic argumentation, which could be beneficial for achieving accurate diagnostic judgments and facilitating post-hoc diagnostic argumentation.

Finally, considering the promising results of NLP-based automatic adaptive feedback, the use of NLP methods for providing adaptive feedback and other adaptive learning support may be further investigated. In particular, the transferability of the employed algorithms to problems other than the used diagnostic ones is of particular interest. Achieving transferability of the employed NLP algorithms would allow the implementation of NLP-based automatic adaptive feedback in simulation-based learning of diagnostic reasoning skills at a larger scale in both medical education and teacher education.

6 Conclusion

Diagnostic reasoning has been researched and taught in medical and teacher education. To advance a cross-disciplinary research perspective on diagnostic reasoning, this thesis discussed diagnostic reasoning regarding the similarities and differences between the two fields of medical and teacher education. Respective research needs to consider and acknowledge the differences between diagnostic reasoning in medical and teacher education: In particular, differences in the content areas of diagnostic problems result in content specificity concerning the diagnostic knowledge (see Förtsch et al., 2018) that is required to achieve diagnostic accuracy. However, diagnostic problems in varying fields and within one single field can be analyzed regarding other characteristics, namely, exemplarity, complexity, and required activities. These problem characteristics can systematically affect diagnostic reasoning processes. Therefore, research should consider the problem characteristics when investigating for example epistemic or cognitive processing in diagnostic reasoning.

In both medical and teacher education, achieving diagnostic accuracy is the central epistemic aim in diagnostic reasoning (see Chinn et al., 2011) to initiate action that helps patients and pupils. However, fields can vary concerning the relative emphasis on diagnostic activities in their diagnostic practices (see Bauer et al., 2020; Fischer et al., 2014; Heitzmann et al., 2019), which they have developed to achieve diagnostic accuracy. Using the novel method of ENA (Shaffer, 2017) to compare the epistemic processing in diagnostic reasoning in medical and teacher education, the findings of the first paper suggested that medical students showed a more hypothesis-driven approach in their diagnostic practices (see Coderre et al., 2010; Kiesewetter et al., 2013; Seidel & Stürmer, 2014). By comparison, preservice teachers demonstrated a more data-driven approach in their diagnostic practices (see Gräsel & Mandl, 1993; Kiesewetter et al., 2013; Norman et al., 2007; Seidel & Stürmer, 2014). Therefore, medical education and teacher education might be considered as representing different epistemic communities (see Kelly, 2008), which have developed specific approaches to diagnostic reasoning that relate to field-specific ideals and standards (see Chinn et al., 2011). The internalization of these field-specific ideals and standards is part of medical students' and preservice teachers' professional development and learning. In addition to content-specific diagnostic knowledge (see Förtsch et al., 2018), the knowledge of field-specific standards may be considered an additional part of the knowledge base on which professionals perform diagnostic reasoning.

The situational application of professional knowledge in solving diagnostic problems can be subsumed under diagnostic reasoning skills. I suggested distinguishing diagnostic reasoning skills concerning diagnostic judgment (see Loibl et al., 2020) and a novel

conceptualization of diagnostic argumentation, including the three facets of justification, disconfirmation, and transparency. The second paper presented in this thesis provided evidence for the assumption that preservice teachers do not necessarily seem equally capable of making accurate diagnostic judgments and formulating justified, disconfirmatory, and transparent diagnostic argumentations. Moreover, the results supported the notion that justification, disconfirmation, and transparency represent distinct subskills of diagnostic argumentation because of the disparities found in the underlying knowledge bases.

The distinction between diagnostic judgment and diagnostic argumentation seems practically relevant in supporting the learning of diagnostic reasoning skills. In simulation-based learning (Chernikova, Heitzmann, Stadler et al., 2020), the implemented automatic adaptive feedback (see Bimba et al., 2017; Pfeiffer et al., 2019) facilitated the justification of diagnostic argumentation compared to a static feedback condition. For the presumably less cognitively demanding skill of making an accurate diagnostic judgment, the adaptive feedback only outperformed static feedback in collaborative, simulation-based learning. Adaptive feedback seemed to effectively compensate for collaboration costs compared to the static feedback condition, in which collaborative learners experienced performance drops in diagnostic accuracy. Moreover, NLP methods, such as the implemented algorithms of artificial neural networks (see Pfeiffer et al., 2019) to automate adaptive feedback, provide particular benefits in fostering learning cognitively demanding reasoning skills, such as diagnostic argumentation, even in short-term interventions. Therefore, providing adaptive feedback is a promising instructional support to help preservice teachers improve their diagnostic reasoning skills and their diagnostic argumentation.

Given that the suggestion to differentiate diagnostic reasoning skills is theoretically grounded by non-field-specific reasons relating to cognitive processing and epistemic aims, I assume that the distinction between diagnostic judgment and diagnostic argumentation, as well as the effects of adaptive feedback and collaboration in simulation-based learning, are also relevant to and replicable in medical education. Research in medical education and teacher education needs to replicate the findings and further investigate the relationships, commonalities, and differences of diagnostic problems, epistemic and cognitive processes in diagnostic reasoning, and the nature and facilitation of diagnostic reasoning skills.

Building on the research and insights presented in this thesis, further cross-disciplinary research in analyzing and facilitating diagnostic reasoning may contribute to preparing medical students and preservice teachers to perform diagnostic reasoning in real-life professional situations.

7 References

Aalberts, J., Koster, E., & Boschhuizen, R. (2012). From prejudice to reasonable judgement: integrating (moral) value discussions in university courses. *Journal of Moral Education*, *41*(4), 437–455. https://doi.org/10.1080/03057240.2012.677600

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

Asterhan, C. S. C., & Schwarz, B. B. (2009). Argumentation and Explanation in Conceptual Change: Indications From Protocol Analyses of Peer-to-Peer Dialog. *Cognitive Science*, *33*(3), 374–400. https://doi.org/10.1111/j.1551-6709.2009.01017.x

Barrows, H. S., & Feltovich, P. J. (1987). The clinical reasoning process. *Medical Education*, *21*(2), 86–91. https://doi.org/10.1111/j.1365-2923.1987.tb00671.x

Barrows, H. S., & Pickell, G. C. (1991). *Developing clinical problem-solving skills: A guide to more effective diagnosis and treatment* (1. ed.). *Norton medical books*. Norton.

Bauer, E., Fischer, F., Kiesewetter, J., Shaffer, D. W., Fischer, M. R., Zottmann, J. M., & Sailer, M. (2020). Diagnostic Activities and Diagnostic Practices in Medical Education and Teacher Education: An Interdisciplinary Comparison. *Frontiers in Psychology*, *11*, Article 562665. https://doi.org/10.3389/fpsyg.2020.562665

Bauer, E., Sailer, M., Kiesewetter, J., Fischer, M. R., & Fischer, F. (2022). Diagnostic Argumentation in Teacher Education: Making the Case for Justification, Disconfirmation, and Transparency. *Frontiers in Education*, *7*, Article 977631. https://doi.org/10.3389/feduc.2022.977631

Bauer, J., & Prenzel, M. (2012). Science education. European teacher training reforms. *Science (New York, N.Y.)*, *336*(6089), 1642–1643. https://doi.org/10.1126/science.1218387

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55. https://doi.org/10.1002/sce.20286

Bimba, A. T., Idris, N., Al-Hunaiyyan, A., Mahmud, R. B., & Shuib, N. L. B. M. (2017). Adaptive feedback in computer-based learning environments: a review. *Adaptive Behavior*, *25*(5), 217–234. https://doi.org/10.1177/1059712317727590

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond Dichotomies. *Zeitschrift Für Psychologie*, *223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194

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

Bradley, P. (2006). The history of simulation in medical education and possible future directions. *Medical Education*, *40*(3), 254–262. https://doi.org/10.1111/j.1365-2929.2006.02394.x

Braun, L. T., Lenzer, B., Kiesewetter, J., Fischer, M. R., & Schmidmaier, R. (2018). How case representations of medical students change during case processing - Results of a qualitative study. *GMS Journal for Medical Education*, *35*(3), Doc41. https://doi.org/10.3205/zma001187

Bricker, L. A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. *Science Education*, *92*(3), 473–498. https://doi.org/10.1002/sce.20278

Bromme, R., Stadtler, M., & Scharrer, L. (2018). The provenance of certainty: Multiple source use and the public engagement with science. In J. L. G. Braasch, I. Bråten, & M. McCrudden (Eds.), *Educational Psychology Handbook. Handbook of Multiple Source Use* (pp. 269–284). Taylor and Francis.

Campbell, D. J. (1988). Task Complexity: A Review and Analysis. *Academy of Management Review*, *13*(1), 40–52. https://doi.org/10.5465/amr.1988.4306775

Charlin, B., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. (2000). The Script Concordance test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine*, *12*(4), 189–195. https://doi.org/10.1207/S15328015TLM1204_5

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

Charlin, B., Lubarsky, S., Millette, B., Crevier, F., Audétat, M.-C., Charbonneau, A., Caire Fon, N., Hoff, L., & Bourdy, C. (2012). Clinical reasoning processes: Unravelling complexity through graphical representation. *Medical Education*, *46*(5), 454–463. https://doi.org/10.1111/j.1365-2923.2012.04242.x

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., & Fischer, F. (2020). Facilitating Diagnostic Competences in Higher Education—a Meta-Analysis in Medical and Teacher Education. *Educational Psychology Review*, *32*(1), 157–196. https://doi.org/10.1007/s10648-019-09492-2

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-Based Learning in Higher Education: A Meta-Analysis. *Review of Educational Research*, *90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*(1), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Chinn, C. A., Buckland, L. A., & Samarapungavan, A. (2011). Expanding the Dimensions of Epistemic Cognition: Arguments From Philosophy and Psychology. *Educational Psychologist*, *46*(3), 141–167. https://doi.org/10.1080/00461520.2011.587722

Chinn, C. A., & Rinehart, R. W. (2016). Commentary: Advances in research on sourcing - source credibility and reliable processes for producing knowledge claims. *Reading and Writing*, *29*(8), 1701–1717. https://doi.org/10.1007/s11145-016-9675-3

Chinn, C. A., Rinehart, R. W., & Buckland, L. A. (2014). Epistemic cognition and evaluating information: Applying the AIR model of epistemic cognition. *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*, 425–453.

Chuang, S.-h., & O'Neil, H. F. (2013). Role of Task-Specific Adapted Feedback on a Computer-Based Collaborative Problem-Solving Task. In H. F. O'Neil & R. S. Perez (Eds.), *Web-Based Learning* (p. 239). Routledge.

Clark, J., Dodd, D., & Coll, R. K. (2008). Border crossing and enculturation into higher education science and engineering learning communities. *Research in Science & Technological Education*, *26*(3), 323–334. https://doi.org/10.1080/02635140802276793

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

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

Cook, D. A., Erwin, P. J., & Triola, M. M. (2010). Computerized virtual patients in health professions education: A systematic review and meta-analysis. *Academic Medicine:*

Journal of the Association of American Medical Colleges, *85*(10), 1589–1602. https://doi.org/10.1097/ACM.0b013e3181edfe13

Cook, D. A., Hamstra, S. J., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., Erwin, P. J., & Hatala, R. (2013). Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. *Medical Teacher*, *35*(1), e867-98. https://doi.org/10.3109/0142159X.2012.714886

Cook, D. A., Hatala, R., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., Erwin, P. J., & Hamstra, S. J. (2011). Technology-enhanced simulation for health professions education: A systematic review and meta-analysis. *JAMA*, *306*(9), 978–988. https://doi.org/10.1001/jama.2011.1234

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

Csanadi, A., Kollar, I., & Fischer, F. (2021). Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: better together? *European Journal of Psychology of Education*, *36*(1), 147–168. https://doi.org/10.1007/s10212-020-00467-4

De Coninck, K., Valcke, M., Ophalvens, I., & Vanderlinde, R. (2019). Bridging the theory-practice gap in teacher education: The design and construction of simulation-based learning environments. In K. Hellmann, J. Kreutz, M. Schwichow, & K. Zaki (Eds.), *Kohärenz in der Lehrerbildung* (Vol. 5, pp. 263–280). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-23940-4_17

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The Evolution of Research on Computer-Supported Collaborative Learning. In N. Balacheff, S. Ludvigsen, T. de Jong, A. Lazonder, & S. Barnes (Eds.), *Technology-Enhanced Learning* (Vol. 29, pp. 3–19). Springer Netherlands. https://doi.org/10.1007/978-1-4020-9827-7_1

Duncan, R. G., & Chinn, C. A. (2016). New Directions for Research on Argumentation: Insights from the AIR Framework for Epistemic Cognition. *Zeitschrift Für Pädagogische Psychologie*, *30*(2-3), 155–161. https://doi.org/10.1024/1010-0652/a000178

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring behavioural patterns during complex problem-solving. *Journal of Computer Assisted Learning*, *36*(6), 933–956. https://doi.org/10.1111/jcal.12451

Elhoweris, H. (2008). Teacher Judgment in Identifying Gifted/Talented Students. *Multicultural Education*, *15*(3), 35–38.

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

Faraone, S. V., Sergeant, J., Gillberg, C., & Biederman, J. (2003). The worldwide prevalence of ADHD: Is it an American condition? *World Psychiatry : Official Journal of the World Psychiatric Association (WPA)*, *2*(2), 104–113.

Fink, M. C., Heitzmann, N., Siebeck, M., Fischer, F., & Fischer, M. R. (2021). Learning to diagnose accurately through virtual patients: Do reflection phases have an added benefit? *BMC Medical Education*, *21*(1), 523. https://doi.org/10.1186/s12909-021-02937-9

Fink, M. C., Radkowitsch, A., Bauer, E., Sailer, M., Kiesewetter, J., Schmidmaier, R., Siebeck, M., Fischer, F., & Fischer, M. R. (2021). Simulation research and design: a dual-level framework for multi-project research programs. *Educational Technology Research and Development*, *69*(2), 809–841. https://doi.org/10.1007/s11423-020-09876-0

Fink, M. C., Reitmeier, V., Stadler, M., Siebeck, M., Fischer, F., & Fischer, M. R. (2021). Assessment of Diagnostic Competences With Standardized Patients Versus Virtual Patients: Experimental Study in the Context of History Taking. *Journal of Medical Internet Research*, *23*(3), e21196. https://doi.org/10.2196/21196

Fischer, F., Clark A., C., Engelmann, K., & Osborne, J. (Eds.). (2018). *Scientific Reasoning and Argumentation*. Routledge. https://doi.org/10.4324/9780203731826

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., & Fischer, M. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, *2*(3), 28–45.

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

Gagné, R. M., & Merrill, M. D. (1990). Integrative goals for instructional design. *Educational Technology Research and Development*, *38*(1), 23–30. https://doi.org/10.1007/BF02298245

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

Ghanem, C., Kollar, I., Fischer, F., Lawson, T. R., & Pankofer, S. (2018). How do social work novices and experts solve professional problems? A micro-analysis of epistemic activities and the use of evidence. *European Journal of Social Work*, *21*(1), 3–19. https://doi.org/10.1080/13691457.2016.1255931

Goodwin, C. (1994). Professional vision. *American Anthropologist*, *96*(3), 606–633. http://www.jstor.org/stable/682303

Gorman, M. E., Gorman, M. E., Latta, R. M., & Cunningham, G. (1984). How disconfirmatory, confirmatory and combined strategies affect group problem solving. *British Journal of Psychology*, *75*(1), 65–79. https://doi.org/10.1111/j.2044-8295.1984.tb02790.x

Graesser, A. C., Hu, X., & Sottilare, R. (2018). Intelligent tutoring systems. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International handbook of the learning sciences* (pp. 246–255). Routledge.

Gräsel, C., & Mandl, H. (1993). Förderung des Erwerbs diagnostischer Strategien in fallbasierten Lernumgebungen. Advance online publication. https://doi.org/10.25656/01:8195

Greene, J. A., Azevedo, R., & Torney-Purta, J. (2008). Modeling Epistemic and Ontological Cognition: Philosophical Perspectives and Methodological Directions. *Educational Psychologist*, *43*(3), 142–160. https://doi.org/10.1080/00461520802178458

Grossman, P., & McDonald, M. (2008). Back to the Future: Directions for Research in Teaching and Teacher Education. *American Educational Research Journal*, *45*(1), 184–205. https://doi.org/10.3102/0002831207312906

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

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

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

Higher Education A Framework and a Research Agenda. *Frontline Learning Research*, 1–24. https://doi.org/10.14786/flr.v7i4.384

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., Klug, J., Hetmanek, A., Ohle, A., Böhmer, I., Karing, C., Kaiser, J., & Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education, 76*(4), 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Hetmanek, A., Engelmann, K., Opitz, A., & Fischer, F. (2018). Beyond intelligence and domain knowledge: Scientific reasoning and argumentation as a set of cross-domain skills. In F. Fischer, C. Clark A., K. Engelmann, & J. Osborne (Eds.), *Scientific Reasoning and Argumentation* (pp. 203–226). Routledge.

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

Hmelo-Silver, L. C. E., & DeSimone, C. (2013). Problem-based learning: An instructional model of collaborative learning. In C. E. Hmelo-Silver (Ed.), *Educational psychology handbook series. The International Handbook of Collaborative Learning* (pp. 382–398). Routlege.

Hoth, J., Döhrmann, M., Kaiser, G., Busse, A., König, J., & Blömeke, S. (2016). Diagnostic competence of primary school mathematics teachers during classroom situations. *ZDM, 48*(1-2), 41–53. https://doi.org/10.1007/s11858-016-0759-y

Hsieh, I.-L. G., & O'Neil, H. F. (2002). Types of feedback in a computer-based collaborative problem-solving group task. *Computers in Human Behavior, 18*(6), 699–715. https://doi.org/10.1016/S0747-5632(02)00025-0

Ilgen, J. S., Humbert, A. J., Kuhn, G., Hansen, M. L., Norman, G. R., Eva, K. W., Charlin, B., & Sherbino, J. (2012). Assessing diagnostic reasoning: A consensus statement summarizing theory, practice, and future needs. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine, 19*(12), 1454–1461. https://doi.org/10.1111/acem.12034

Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: towards a research agenda. *Educational Technology Research and Development*, *68*(2), 783–805. https://doi.org/10.1007/s11423-019-09729-5

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*(1), 100284. https://doi.org/10.1016/j.edurev.2019.100284

Jiménez-Aleixandre, M. P., & Crujeiras, B. (2017). Epistemic Practices and Scientific Practices in Science Education. In K. S. Taber & B. Akpan (Eds.), *Science Education* (Vol. 79, pp. 69–80). SensePublishers. https://doi.org/10.1007/978-94-6300-749-8_5

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *The American Psychologist*, *58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D. (2011). *Thinking, fast and slow* (First edition). Farrar Straus and Giroux.

Kalyuga, S. (2011). Informing: A Cognitive Load Perspective. *Informing Science: The International Journal of an Emerging Transdiscipline*, *14*, 33–45. https://doi.org/10.28945/1349

Kassirer, J. P. (2010). Teaching clinical reasoning: Case-based and coached. *Academic Medicine : Journal of the Association of American Medical Colleges*, *85*(7), 1118–1124. https://doi.org/10.1097/ACM.0b013e3181d5dd0d

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

Kelly, G. J. (2008). Inquiry, Activity and Epistemic Practice. In R. A. Duschl & R. E. Grandy (Eds.), *Teaching Scientific Inquiry* (pp. 99–117). BRILL. https://doi.org/10.1163/9789460911453_009

Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Science Education*, *86*(3), 314–342. https://doi.org/10.1002/sce.10024

Kiesewetter, J., Ebersbach, R., Görlitz, A., Holzer, M., Fischer, M. R., & Schmidmaier, R. (2013). Cognitive problem solving patterns of medical students correlate with success in diagnostic case solutions. *PloS One*, *8*(8), e71486. https://doi.org/10.1371/journal.pone.0071486

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

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., Verschaffel, L., Star, J., & van Dooren, W. (2017). There is more variation within than across domains: an interview with Paul A. Kirschner about applying cognitive psychology-based instructional design principles in mathematics teaching and learning. *ZDM*, *49*(4), 637–643. https://doi.org/10.1007/s11858-017-0875-3

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

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

Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A.-K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education*, *100*(4), 103298. https://doi.org/10.1016/j.tate.2021.103298

Kopp, V., Stark, R., Kühne-Eversmann, L., & Fischer, M. R. (2009). Do worked examples foster medical students' diagnostic knowledge of hyperthyroidism? *Medical Education*, *43*(12), 1210–1217. https://doi.org/10.1111/j.1365-2923.2009.03531.x

Kramer, M., Förtsch, C., Boone, W. J., Seidel, T., & Neuhaus, B. J. (2021). Investigating Pre-Service Biology Teachers' Diagnostic Competences: Relationships between Professional Knowledge, Diagnostic Activities, and Diagnostic Accuracy. *Education Sciences*, *11*(3), 89. https://doi.org/10.3390/educsci11030089

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

Kron, S., Sommerhoff, D., Achtner, M., & Ufer, S. (2021). Selecting Mathematical Tasks for Assessing Student's Understanding: Pre-Service Teachers' Sensitivity to and Adaptive Use of Diagnostic Task Potential in Simulated Diagnostic One-To-One Interviews. *Frontiers in Education*, *6*, Article 604568, 738. https://doi.org/10.3389/feduc.2021.604568

Lachner, A., Jarodzka, H., & Nückles, M. (2016). What makes an expert teacher? Investigating teachers' professional vision and discourse abilities. *Instructional Science*, *44*(3), 197–203. https://doi.org/10.1007/s11251-016-9376-y

Lawson, A. (2003). The nature and development of hypothetico-predictive argumentation with implications for science teaching. *International Journal of Science Education*, *25*(11), 1387–1408. https://doi.org/10.1080/0950069032000052117

Lenzer, B., Ghanem, C., Weidenbusch, M., Fischer, M. R., & Zottmann, J. (Eds.) (2017). *Scientific reasoning in medical education: a novel approach for the analysis of epistemic activities in clinical case discussions*.

Leont'ev, A. N. (1978). *Activity, consciousness, and personality*. Prentice-Hall.

Li, H. (2018). Deep learning for natural language processing: advantages and challenges. *National Science Review*, *5*(1), 24–26. https://doi.org/10.1093/nsr/nwx110

Loibl, K., Leuders, T., & Dörfler, T. (2020). A Framework for Explaining Teachers' Diagnostic Judgements by Cognitive Modeling (DiaCoM). *Teaching and Teacher Education*, *91*(3), 103059. https://doi.org/10.1016/j.tate.2020.103059

Louie, A. K., Roberts, L. W., & Coverdale, J. (2007). The enculturation of medical students and residents. *Academic Psychiatry : The Journal of the American Association of Directors of Psychiatric Residency Training and the Association for Academic Psychiatry*, *31*(4), 253–257. https://doi.org/10.1176/appi.ap.31.4.253

Mamede, S., Schmidt, H. G., Rikers, R. M. J. P., Penaforte, J. C., & Coelho-Filho, J. M. (2007). Breaking down automaticity: Case ambiguity and the shift to reflective approaches in clinical reasoning. *Medical Education*, *41*(12), 1185–1192. https://doi.org/10.1111/j.1365-2923.2007.02921.x

Manning, C. D., & Schuetze, H. (2005). *Foundations of statistical natural language processing, 8. Ausgabe*. MIT Press.

Mayer, R. E. (2010). Applying the science of learning to medical education. *Medical Education*, *44*(6), 543–549. https://doi.org/10.1111/j.1365-2923.2010.03624.x

Mercier, H., & Heintz, C. (2014). Scientists' Argumentative Reasoning. *Topoi*, *33*(2), 513–524. https://doi.org/10.1007/s11245-013-9217-4

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press. https://doi.org/10.4159/9780674977860

Mohr-Jensen, C., Steen-Jensen, T., Bang-Schnack, M., & Thingvad, H. (2019). What Do Primary and Secondary School Teachers Know About ADHD in Children? Findings From a Systematic Review and a Representative, Nationwide Sample of Danish Teachers. *Journal of Attention Disorders*, *23*(3), 206–219. https://doi.org/10.1177/1087054715599206

Monteiro, S. D., Sherbino, J., Schmidt, H., Mamede, S., Ilgen, J., & Norman, G. (2020). It's the destination: Diagnostic accuracy and reasoning. *Advances in Health Sciences Education : Theory and Practice*, *25*(1), 19–29. https://doi.org/10.1007/s10459-019-09903-7

Moreno, R. (2004). Decreasing Cognitive Load for Novice Students: Effects of Explanatory versus Corrective Feedback in Discovery-Based Multimedia. *Instructional Science*, *32*(1/2), 99–113. https://doi.org/10.1023/B:TRUC.0000021811.66966.1d

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, *14*(10), 435–440. https://doi.org/10.1016/j.tics.2010.07.004

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector (Ed.), *Handbook of research on educational communications and technology* (3rd ed., Vol. 3, pp. 125–144). Lawrence Erlbaum Associates.

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, *71*(2), 56–76. https://doi.org/10.1016/j.compedu.2013.09.011

Norman, G. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education: Theory and Practice*, *14 Suppl 1*, 37–49. https://doi.org/10.1007/s10459-009-9179-x

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education*, *41*(12), 1140–1145. https://doi.org/10.1111/j.1365-2923.2007.02914.x

Norman, G. R., Monteiro, S. D., Sherbino, J., Ilgen, J. S., Schmidt, H. G., & Mamede, S. (2017). The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. *Academic Medicine: Journal of the Association of American Medical Colleges*, *92*(1), 23–30. https://doi.org/10.1097/ACM.0000000000001421

O'Donnell, A. M., & Hmelo-Silver, C. E. (2013). Introduction what is collaborative learning? An overview. In C. E. Hmelo-Silver (Ed.), *Educational psychology handbook series. The International Handbook of Collaborative Learning* (pp. 13–28). Routlege.

Okada, T. (1997). Collaborative discovery in a scientific domain. *Cognitive Science*, *21*(2), 109–146. https://doi.org/10.1016/S0364-0213(99)80020-2

Osberg, D. (2005). Redescribing ‘Education’ in Complex Terms. *Complicity: An International Journal of Complexity and Education*, *2*(1). https://doi.org/10.29173/cmplct8731

Osborne, J. (2014). Teaching Scientific Practices: Meeting the Challenge of Change. *Journal of Science Teacher Education*, *25*(2), 177–196. https://doi.org/10.1007/s10972-014-9384-1

Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine : Journal of the Association of American Medical Colleges*, *70*(3), 194–201. https://doi.org/10.1097/00001888-199503000-00009

Papa, F. J., Stone, R. C., & Aldrich, D. G. (1996). Further evidence of the relationship between case typicality and diagnostic performance: Implications for medical education. *Academic Medicine : Journal of the Association of American Medical Colleges*, *71*(1 Suppl), S10-2. https://doi.org/10.1097/00001888-199601000-00028.

Papa, F. J. (2016). A Dual Processing Theory Based Approach to Instruction and Assessment of Diagnostic Competencies. *Medical Science Educator*, *26*(4), 787–795. https://doi.org/10.1007/s40670-016-0326-8

Pfeiffer, J., Meyer, C. M., Schulz, C., Kiesewetter, J., Zottmann, J., Sailer, M., Bauer, E., Fischer, F., Fischer, M. R., & Gurevych, I. (2019). Famulus: Interactive annotation and feedback generation for teaching diagnostic reasoning. In S. Padó (Ed.), *The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing - proceedings of system demonstrations: November 3-7, 2019, Hong Kong, China : Emnlp-IJCNLP 2019* (pp. 73–78). Association for Computational Linguistics (ACL). https://aclanthology.org/D19-3013.pdf

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

Poznanski, B., Hart, K. C., & Graziano, P. A. (2021). What Do Preschool Teachers Know About Attention-Deficit/Hyperactivity Disorder (ADHD) and Does It Impact Ratings of Child Impairment? *School Mental Health*, *13*(1), 114–128. https://doi.org/10.1007/s12310-020-09395-6

Praetorius, A.-K., Berner, V.-D., Zeinz, H., Scheunpflug, A., & Dresel, M. (2013). Judgment Confidence and Judgment Accuracy of Teachers in Judging Self-Concepts of

Students. *The Journal of Educational Research*, *106*(1), 64–76. https://doi.org/10.1080/00220671.2012.667010

Praetorius, A.-K., Koch, T., Scheunpflug, A., Zeinz, H., & Dresel, M. (2017). Identifying determinants of teachers' judgment (in)accuracy regarding students' school-related motivations using a Bayesian cross-classified multi-level model. *Learning and Instruction*, *52*(4), 148–160. https://doi.org/10.1016/j.learninstruc.2017.06.003

Putnam, R. T. (1987). Structuring and Adjusting Content for Students: A Study of Live and Simulated Tutoring of Addition. *American Educational Research Journal*, *24*(1), 13–48. https://doi.org/10.3102/00028312024001013

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., 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*(5), 101487. https://doi.org/10.1016/j.learninstruc.2021.101487

Rapanta, C., & Felton, M. K. (2021). Learning to Argue Through Dialogue: a Review of Instructional Approaches. *Educational Psychology Review*, *38*(1), 67. https://doi.org/10.1007/s10648-021-09637-2

Reinke, W. M., Stormont, M., Herman, K. C., Puri, R., & Goel, N. (2011). Supporting children's mental health in schools: Teacher perceptions of needs, roles, and barriers. *School Psychology Quarterly*, *26*(1), 1–13. https://doi.org/10.1037/a0022714

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, *38*(1), 1–37. https://doi.org/10.1111/cogs.12086

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57. https://doi.org/10.1093/applin/22.1.27

Roschelle, J., & Teasley, S. D. (1995). The Construction of Shared Knowledge in Collaborative Problem Solving. In C. O'Malley (Ed.), *Computer Supported Collaborative Learning* (Vol. 13, pp. 69–97). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-85098-1_5

Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: the role of explaining and responding to questions. *Instructional Science*, *36*(4), 321–350. https://doi.org/10.1007/s11251-007-9034-5

Roth, W.-M., & Lee, Y.-J. (2006). Contradictions in theorizing and implementing communities in education. *Educational Research Review*, *1*(1), 27–40. https://doi.org/10.1016/j.edurev.2006.01.002

Rothì, D. M., Leavey, G., & Best, R. (2008). On the front-line: Teachers as active observers of pupils' mental health. *Teaching and Teacher Education*, *24*(5), 1217–1231. https://doi.org/10.1016/j.tate.2007.09.011

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction*, *83*, Article 101620. https://doi.org/10.1016/j.learninstruc.2022.101620

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, *92*(3), 447–472. https://doi.org/10.1002/sce.20276

Sandoval, W. A., & Millwood, K. A. (2005). The Quality of Students' Use of Evidence in Written Scientific Explanations. *Cognition and Instruction*, *23*(1), 23–55. https://doi.org/10.1207/s1532690xci2301_2

Sauvé, L., Renaud, L., Kaufman, D., & Marquis, J.-S. (2007). Distinguishing between games and simulations: A systematic review. *Journal of Educational Technology & Society*, *10*(3), 247–256.

Scahill, L., & Schwab-Stone, M. (2000). Epidemiology of Adhd in School-Age Children. *Child and Adolescent Psychiatric Clinics of North America*, *9*(3), 541–555. https://doi.org/10.1016/S1056-4993(18)30106-8

Scales, R. Q., Wolsey, T. D., Lenski, S., Smetana, L., Yoder, K. K., Dobler, E., Grisham, D. L., & Young, J. R. (2018). Are We Preparing or Training Teachers? Developing Professional Judgment in and Beyond Teacher Preparation Programs. *Journal of Teacher Education*, *69*(1), 7–21. https://doi.org/10.1177/0022487117702584

Scheuer, O., McLaren, B. M., Loll, F., & Pinkwart, N. (2012). Automated analysis and feedback techniques to support and teach argumentation: A survey. *Educational Technologies for Teaching Argumentation Skills*, *10*(3), 71–124.

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

Schulz, C., Meyer, C. M., & Gurevych, I. (2019). Challenges in the Automatic Analysis of Students' Diagnostic Reasoning. In A. Korhonen, D. Traum, & L. Màrquez (Eds.),

The 57th Annual Meeting of the Association for Computational Linguistics - proceedings of the conference: July 28-August 2, 2019, Florence, Italy (pp. 6974–6981). Association for Computational Linguistics.

Schwartz, A., & Elstein, A. S. (2008). Clinical Problem Solving and Diagnostic Decision Making: A Selective Review of the Cognitive Research Literature. In J. A. Knottnerus & F. Buntinx (Eds.), *The Evidence Base of Clinical Diagnosis* (pp. 237–255). Wiley-Blackwell. https://doi.org/10.1002/9781444300574.ch12

Schwarz, B. B., Neuman, Y., Gil, J., & Ilya, M. (2003). Construction of Collective and Individual Knowledge in Argumentative Activity. *Journal of the Learning Sciences*, *12*(2), 219–256. https://doi.org/10.1207/S15327809JLS1202_3

Seidel, T., & Prenzel, M. (2008). Wie Lehrpersonen Unterricht wahrnehmen und einschätzen - Erfassung pädagogisch-psychologischer Kompetenzen mit Videosequenzen [How teachers observe and interpret classroom situations]. In M. Prenzel, I. Gogolin, & H.-H. Krüger (Eds.), *Kompetenzdiagnostik* (pp. 201–216). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90865-6_12

Seidel, T., & Stürmer, K. (2014). Modeling and Measuring the Structure of Professional Vision in Preservice Teachers. *American Educational Research Journal*, *51*(4), 739–771. https://doi.org/10.3102/0002831214531321

Shaffer, D. W. (2017). *Quantitative ethnography* (First printing). Cathcart Press.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Shulman, L. S. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review*, *57*(1), 1–23. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Siebert-Evenstone, A. L., Irgens, G. A., Collier, W., Swiecki, Z., Ruis, A. R., & Shaffer, D. W. (2017). In search of conversational grain size: modelling semantic structure using moving stanza windows. *Journal of Learning Analytics*, *4*(3), 123–139.

Stadler, M., Niepel, C., & Greiff, S. (2019). Differentiating between static and complex problems: A theoretical framework and its empirical validation. *Intelligence*, *72*(185), 1–12. https://doi.org/10.1016/j.intell.2018.11.003

Stürmer, K., Seidel, T., & Holzberger, D. (2016). Intra-individual differences in developing professional vision: preservice teachers' changes in the course of an innovative teacher education program. *Instructional Science*, *44*(3), 293–309. https://doi.org/10.1007/s11251-016-9373-1

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. https://doi.org/10.1037/a0027627

Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review*, *22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Thomassen, A. O., & Stentoft, D. (2020). Educating Students for a Complex Future: Why Integrating a Problem Analysis in Problem-Based Learning Has Something to Offer. *Interdisciplinary Journal of Problem-Based Learning*, *14*(2). https://doi.org/10.14434/ijpbl.v14i2.28804

Toulmin, S. E. (1958). *The uses of argument*. Cambridge university press.

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*(4), 100374. https://doi.org/10.1016/j.edurev.2020.100374

van Eemeren, F. H., & Grootendorst, R. (2004). *A Systematic Theory of Argumentation: The pragma-dialectical approach*. Cambridge university press.

Vazire, S. (2017). Quality Uncertainty Erodes Trust in Science. *Collabra: Psychology*, *3*(1), Article 1, 488. https://doi.org/10.1525/collabra.74

Victor-Chmil, J. (2013). Critical Thinking Versus Clinical Reasoning Versus Clinical Judgment: Differential Diagnosis. *Nurse Educator*, *38*(1), 34–36. https://doi.org/10.1097/NNE.0b013e318276dfbe

Volpe, R. J., DuPaul, G. J., DiPerna, J. C., Jitendra, A. K., Lutz, J. G., Tresco, K., & Junod, R. V. (2006). Attention Deficit Hyperactivity Disorder and Scholastic Achievement: A Model of Mediation via Academic Enablers. *School Psychology Review*, *35*(1), 47–61. https://doi.org/10.1080/02796015.2006.12088001

Walton, D. N. (1990). What is Reasoning? What Is an Argument? *The Journal of Philosophy*, *87*(8), 399. https://doi.org/10.2307/2026735

Weinberger, A., Stegmann, K., & Fischer, F. (2010). Learning to argue online: Scripted groups surpass individuals (unscripted groups do not). *Computers in Human Behavior*, *26*(4), 506–515. https://doi.org/10.1016/j.chb.2009.08.007

Weinstein, A., & Pinto-Powell, R. (2016). Introductory Clinical Reasoning Curriculum. *MedEdPORTAL*, Article 10370. Advance online publication. https://doi.org/10.15766/mep_2374-8265.10370

Westwood, P. S. (2008). *What teachers need to know about reading and writing difficulties*. *What teachers need to know about*. ACER Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=244168

Wildgans-Lang, A., Scheuerer, S., Obersteiner, A., Fischer, F., & Reiss, K. (2020). Analyzing prospective mathematics teachers' diagnostic processes in a simulated environment. *ZDM*, *52*(2), 241–254. https://doi.org/10.1007/s11858-020-01139-9

Wimmers, P. F., Splinter, T. A. W., Hancock, G. R., & Schmidt, H. G. (2007). Clinical competence: General ability or case-specific? *Advances in Health Sciences Education: Theory and Practice*, *12*(3), 299–314. https://doi.org/10.1007/s10459-006-9002-x

Winch, C. (2004). What Do Teachers Need To Know About Teaching? A Critical Examination Of The Occupational Knowledge Of Teachers. *British Journal of Educational Studies*, *52*(2), 180–196. https://doi.org/10.1111/j.1467-8527.2004.00262.x

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648–1668. https://doi.org/10.1080/09500693.2017.1347303

Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*(2), 103668. https://doi.org/10.1016/j.compedu.2019.103668

Ziv, A., Wolpe, P. R., Small, S. D., & Glick, S. (2003). Simulation-based medical education: An ethical imperative. *Academic Medicine: Journal of the Association of American Medical Colleges*, *78*(8), 783–788. https://doi.org/10.1097/00001888-200308000-00006

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35–62. https://doi.org/10.1002/tea.10008

8 Appendices

Appendix A

Case Materials

Case Materials from Teacher Education

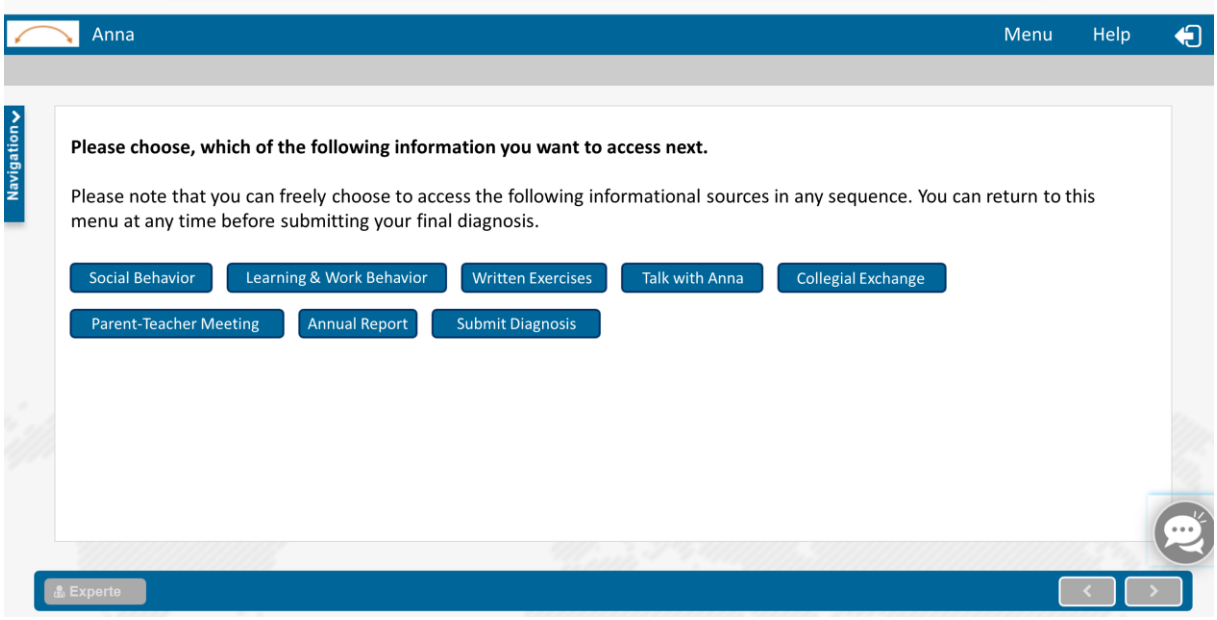
See Supplementary material of study 1 (Bauer et al., 2020): In the teacher education cases, preservice teachers were asked to imagine themselves in the position of a teacher who was encountering a student with some initial performance-related or behavioral problems that might even be clinically relevant, e.g. ADHD or dyslexia.

One example is the case of a secondary student named Anna who is displaying symptoms of an attention-deficit disorder. The leaners are asked to put themselves into the role of Anna's class teacher, who teaches German classes and music lessons. The initial problem statement for the case describes Anna as a 5th grade student, eleven years old, who constantly needs to be pushed to finish her tasks and who has bad grades in many subjects, especially the main subjects. The learners could examine written observations of Anna's in-class and out-of-class behavior, read recordings of conversations with Anna, or with her parents and several teachers, or look at Anna's last annual report and an example of a written exercise (see Figure A1). Her behavior is described as very calm and distracted. She is slow in reading and it is difficult for her to answer questions about a text that she just read. She often fails to fulfill the exact instruction of a task or fails to fully complete a task. Moreover, she often does not bring all required school supplies or comes late in the mornings. In a parent-teacher meeting, Anna's mother backs up the impression of a disorganized and slow learning behavior when talking about the homework situation. Anna's last annual report and the conversations with the other teachers show that her grades are mostly affected by her inattentiveness as well, with the exception of artistic subjects and gym classes. Anna mostly interacts with her one friend and is rather distanced from the other students. Anna herself points out that it is hard for her to concentrate since she feels easily distracted. However, at home, where there are fewer ambient noises, she can focus on and enjoy reading, drawing and painting. Overall, the case information was designed in a way that, the diagnosis of an attention-deficit disorder is the most likely clinical diagnosis, despite several differential diagnoses being relevant.

All case materials are publicly available in German language in a repository (https://osf.io/hn7wm/).

Figure A1

Screenshot of user interface for the teacher education case in the CASUS learning environment (see supplementary material of study 1; Bauer et al., 2020)



Case Materials from Medical Education

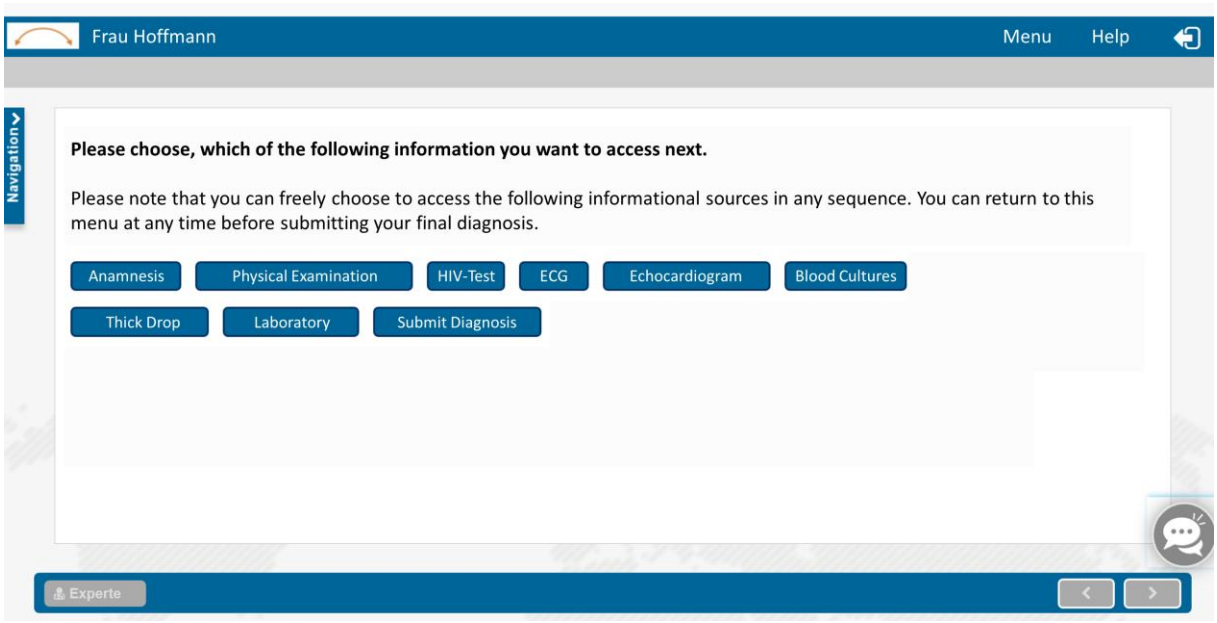
See Supplementary material of study 1 (Bauer et al., 2020): The medical education cases presented virtual patients with symptoms of fever and back pain and medical students were asked to take over the role of a general practitioner.

One exemplary case was about a 36 year-old female, Mrs. Hoffmann, who had a febrile and flu-like infection for almost a week before seeing the doctor. In addition, she experienced fatigue, loss of appetite, sickness and diarrhea. One month earlier, she returned from a trip to Costa Rica, for which she received the recommended vaccinations prior to departure. The anamnesis provided the information that no other persons in her surrounding had the same symptoms; that she does not know about any pre-existing illnesses and does not consume any prescribed drugs, apart from occasionally using homeopathic globules; and that, she is allergic to penicillin and nickel; moreover, she is a non-smoker, occasionally consumes alcohol, and excluded the option of pregnancy. To gather more information, learners could access the patient's history and had the option to access different tests and test results, e.g. physical examination, laboratory, x-ray, ECG, HIV test, and others (see Figure A2). Overall, the symptoms and test results point to an acute hepatitis A infection. Typical symptoms of the patient's current stage of the infection are fatigue, limb pain, fever, sickness, diarrhea and joint pain.

All case materials are publicly available in German language in a repository (https://osf.io/92nyv/).

Figure A2

Screenshot of user interface for the medical education case in the CASUS learning environment (see supplementary material of study 1; Bauer et al., 2020)



Appendix B

Coding Schemes

Coding Scheme for Epistemic Diagnostic Activities

For study 1 (Bauer et al., 2020), we coded two sets of written explanations of learners' diagnostic reasoning (i.e., justificatory reports), one set from medical education and one set for teacher education, on four diagnostic activities: *generating hypotheses*, *generating evidence*, *evaluating evidence*, and *drawing conclusions*. We developed a coding scheme applicable for medical education and teacher education. Coding and segmentation were done simultaneously to account for overlap in the activities as well. The text data collected and coded for the first study in teacher education was reanalyzed as part of the second study (Bauer et al., 2022).

Procedure

Annotation time shall not exceed a maximum of 1.5 hours without taking a break. You should then take a break of at least 15 minutes. The annotation takes place in several steps, whereby a text is always annotated at once, i.e. without a break. A text is first read completely before starting its annotation. Then segments and categories are annotated. Finally, the whole text is checked again to make sure that all segments and categories are correctly annotated.

Please pay attention to annotate precisely to ensure a high level of agreement between the annotators.

Segmentation

The texts are not pre-segmented. Raters see entire texts (these are usually between two and twenty sentences long), in which they then freely determine segments that they annotate. These can be single words up to cross-sentence segments. Punctuation marks (full stop, comma, etc.) at the end of a segment are not annotated, punctuation marks within an annotated segment are (of course) included into the annotation. If two segments are separated by "and" or a punctuation mark, this is not annotated as part of the segments.

General Annotation Rules

The following four epistemic diagnostic activities are annotated: 1) Hypotheses Generation, 2) Evidence Generation, 3) Evidence Evaluation, and 4) Drawing Conclusion. All activities can generally appear anywhere in the text. To distinguish between different epistemic-diagnostic activities, the main focus is on the function of the respective segment in relation to the diagnosis or thought process expressed in the text. The content, especially its correctness and relevance, does not matter. The annotation is based on linguistic references to the function. An example is "Decreased eyesight indicates that he does not see the boundaries in the mandala clearly". Here the whole sentence is a conclusion that also contains evidence. Normally, the decreased eyesight should be inferred from the failure to recognize the boundary lines, so the failure to recognize should be the evidence. Here, however, it is linguistically the other way around: from the reduced vision, it is concluded that the boundary lines are not correctly recognized (X indicates that Y). The annotation here follows what is expressed linguistically: the first part is also annotated as an evidence evaluation.

One versus Multiple Activities. An activity (i.e., a segment) can contain several content elements if these elements are linguistically related and a division into two segments is therefore not linguistically meaningful. This means that there is only one "element of action", e.g., a verb. For example, “I performed Test X and Test Y” counts as one generation of evidence, although it contains two elements (Test X, Test Y) in terms of content. Listings without a verb also count as an "element of action" and therefore as a coherent segment. In “The patient shows symptoms of X: insomnia, poor appetite, high blood pressure.” The list “Insomnia, poor appetite, high blood pressure” is a segment (evidence evaluation). In addition, an epistemic diagnostic activity (a segment) can comprise several sentences if the facts of the first sentence are explained in more detail in the following sentences.

Examples: “Markus seems to have a general problem with aggression. Everything just seems to make him aggressive. Whether it's school, his hobby, his friends, classmates, parents

or teachers.", "In addition, he has difficulties not only in German, but also in other subjects, such as art in particular. In this subject he once had to color in a mandala and he was unable to paint the fields properly."

In summary, the following rule applies:

1) A related activity is indicated by presence of only one verb ("element of action");

2) Several verbs ("element of action ") in combination with a more detailed description and explanation of a matter indicate one coherent activity (especially if it would not be clear what it was about if it were divided into several activities, since one activity depends on another in terms of content)

3) Several verbs ("element of action ") in combination with a description of various issues indicate several individual activities.

Examples: "His typeface is unclean and careless mistakes can be seen in exams." (2 facts), "The fact that she has no friends and is very opinionated is due to the fact that she is not really a "child" at home, but that the siblings take care of themselves and thus have to stand up for themselves, as none of the parents approve them seems to regulate and correct." (three issues, the last is explained in detail), "Since hypothyroidism is actually present, she takes L-thyroxine." (two issues).

In general, insertions in brackets are annotated as part of the same segment. For example, "This is confirmed by the hepatitis serology (IgM stands for an acute infection, with a previous vaccination IgG would be positive)." is a segment (evidence evaluation). In rare cases, an activity may not have a verb if it is still clear what type of activity it is.

Examples: "31-year-old patient with acute pain in the lumbar spine area after carrying heavy loads and suspected lumbago.", "It follows that she developed a weakness in reading and writing due to the lack of attention." (Evidence Evaluation).

Insertions and Overlaps. If an activity is inserted into another activity, the insertion is annotated as part of the outer segment and annotated as a separate segment. For example, in the case of "I suspect that diagnosis Y applies despite my observation X", the entire segment is annotated as hypotheses generation / conclusion (depending on the context) and the inset "despite my observation X", i.e. Evidence Evaluation, is annotated. However, epistemic diagnostic activities of the same type (e.g. two evidence evaluations) cannot overlap.

Content Errors. When annotating epistemic diagnostic activities, it is not important that the content is correct, but only that it is an activity from a linguistic point of view. For example, segments with incorrect content such as "Since the moon is made of cheese, the

patient has X" or "The blood test showed that the hand was broken" are nevertheless annotated as epistemic diagnostic activities (conclusion and evidence evaluation).

Spelling and Grammar. No spelling or grammar improvements are made prior to annotation. As far as the text can be interpreted, spelling and grammatical errors in the annotation are not annotated differently. In cases in which the error gives rise to different possibilities for interpretation, the annotators choose the most likely one and take a note for later comparison (name of the text and a short note on the problematic piece of text). For such cases, a piece of paper and a pen are kept ready during the annotation.

Categories of Epistemic Diagnostic Activities

Hypotheses Generation. The function of generating a hypothesis is to initiate the diagnostic process, for example to stimulate the generation and evaluation of evidence. Therefore, a potential diagnosis of the problem is often named, which guides the further diagnostic process. Generating hypotheses often appear at the beginning of a text.

Examples: "I would initially suspect that Thomas has a visual impairment." (first sentence in the text), "Maria does not seem very mature emotionally. (first sentence in the text).

However, a generation of hypotheses can also appear in other places in the text, for example before the evidence generation and evaluation. It can also appear at the end of a text after various evidence evaluations, the following could be written, "although I first thought that X", which is annotated as hypotheses generation, since the hypothesis is thus placed before the aforementioned evidence generation and evaluation.

Segmentation: Part of a hypothesis is anything that suggests that it is a hypothesis, especially linguistic clues. The segment annotated as Hypothesis Generation should also be recognizable as a hypothesis if it is on its own (i.e. without further sub-clauses).

Examples: "Young patients with acute back pain after exertion (lifting heavy loads) initially suggest an acute trauma." (first sentence in the text), "The patient shows typical symptoms of X: insomnia, poor appetite, high blood pressure." (first sentence in the text), "The patient's increased blood pressure indicates disease X.", "The physical examination reveals pain in the right upper abdomen, this could be pain in the liver capsule." (first sentence in the text).

Evidence Generation. This activity refers to elements that describe how evidence was obtained, for example, "I looked at Test X". Most of the time, this activity includes references to materials from the case simulation.

Example: “First I read through all the information about Annika thoroughly and looked at the relevant material.”

However, generating evidence is also available if, for example, one's own knowledge retrieval is explicitly mentioned, as in “I remembered course X” or “I've seen something similar before”.

Since activities are annotated here, mentioning an investigation as part of the evidence evaluation does not count as an evidence generation, as it does not include any activity on the part of the writer. So "in test X ..." or "in examination X ..." is not an evidence generation and usually counts as part of the evidence evaluation. However, a generation of evidence does not necessarily have to contain a verb. If an action that generates evidence is expressed through a noun, it is evidence generation.

Examples: “An infection was detected when looking at the laboratory results.”, “An infection was shown in the laboratory results.” (No action element, therefore no evidence generation), “While reading the case, it was noticed right from the start that she had problems reading.”

Evidence Evaluation. This involves referring to information that can (but does not have to) be used to support or exclude hypotheses or conclusions. It is thus an evaluation of the relevance of given information or a selection of relevant information. Evidence evaluation often includes the results of tests or investigations. Here, linguistic elements that express an active selection of information or reflective thinking are annotated as part of the evidence evaluation, e.g. “It is also important to note that X” or “It is noticeable that X” or “I think that X is important is".

Examples: “It was also noticed that she had problems in reading samples, which was confirmed on the reading sample.”, “No travel vaccination was given.”, “She is very much looking for the attention of teachers or educators.”, “This was also shown with the grade 2 in the certificate in the subject of writing.”, “This can be justified with the fact that suddenly his performance has dropped so much compared to the previous school years.”, “His parents also note that he has great problems doing his homework in one piece and with concentration.”

In addition, own knowledge that is linked to the present case counts as evidence evaluation, e.g., "the normal blood value is Y". This also includes the evaluation of the correctness of the information given, for example the correctness of test results (e.g. doubting), as in "Blood tests are inaccurate" or "You cannot trust in X-ray image because ...“

Source of Evidence: If the source of the evidence is not explicitly generated (in the sense of the definition of evidence generation), it is part of the evidence evaluation, such as in "Test Y notices that X" or "Examination Y shows that X ".

Segmentation: Evidence evaluations usually include complete (partial) sentences. Words at the beginning of a sentence which, for example, form a transition from the previous sentence (such as “also”, “there”, “but” ...) are annotated as part of the evidence evaluation. This is especially the case when it comes to an evidence evaluation that forms an insertion in a conclusion, e.g. "[There evidence], it is disease X" is a conclusion, where "Because of [evidence]" is also used as an evidence evaluation annotated (see the “Conclusions” section for more information).

Examples: “This can be justified by the fact that suddenly his performance has dropped so much compared to the previous school years.”, “Since he is otherwise very good, especially in sports, and enthusiastically participates ...”, “... because the mistakes appeared so suddenly and Thomas used to be a good student.”

Drawing Conclusions. This activity includes any kind of integration of information that generates new information related to the diagnosis of the problem. It is unimportant how much information is integrated or how much evidence the conclusion refers to. Frequently occurring signal words are verbs (and corresponding nouns) such as “conclude, diagnose” and conjunctions such as “there, therefore, therefore, therefore, consequently”. However, a conclusion does not have to explicitly refer to evidence. For example, if several pieces of evidence are mentioned and at the end of the text it says “The patient has X”, this counts as a conclusion, as the sequence in the text implicitly makes it clear that this follows from the previous evidence (even without conclusive signal words).

Examples: “Markus has a lot of power that he has to let out, but no ADS.” (only one verb, so one segment), “Markus has a lot of power that he has to let out, but he has no ADS.” (here there are two verbs, thus two action elements, and since there are two facts, there are two segments)

It is important that a conclusion is an activity, so it can be recognized by its inferential function. A conclusion therefore does not necessarily have to contain a diagnosis.

Examples: “The diagnosis could be made from the anamnesis, physical examination and laboratory-proven positive HLA-27.” (Evidence Evaluation), ”Otherwise she shows no symptoms that could suggest another disease.” (Evidence Evaluation)

The word "diagnosis" also suggests a conclusion as it indicates the completion of the diagnostic process. This means, for example, “Diagnosis: ADHD.” is a conclusion, regardless of where it is in the text.

Segmentation: It is important that with a conclusion, the linguistic elements that indicate a conclusion (if any) are always annotated. The segment annotated as a conclusion should therefore be recognizable as a conclusion if it stands alone. In addition, evidence that is in the same sentence and supports the inferring part is annotated as part of the conclusion. The evidence evaluation (or evidence generation) is also (separately) annotated as such. Not every conclusion has to contain an evidence evaluation.

Examples: “The value X tells me that he has a reading disorder.”, “For example, I would rule out autism in X.”, “Because X cannot actually be a reading disorder.”, “Since in my opinion, as described above, everything points to a visual impairment, Thomas might just need some glasses.”

Coding Scheme for Differential Diagnoses

In study 2 (Bauer et al., 2022), we used the written explanations of learners’ diagnostic reasoning as the data source to determine the differential diagnoses that learners considered in their written explanations in the six included cases. We developed a coding scheme including a set differential diagnoses, which was used for the coding of all six included cases. We used one category per case, which represented the presence (coded as 1) or absence (coded as 0) of at least two or more differential diagnoses in the learners’ explanation.

List of Differential Diagnoses

Social problems with peers or family, Behaviors related to emotional stress (no depression), Puberty-related behaviors, Visual impairment, Hearing impairment, Intellectual giftedness (behaviors related to insufficient intellectual challenges), Mental retardation, Anxiety disorder, Depression, Autism, Developmental disorder, Language development disorder, Isolated reading disorder, Isolated spelling disorder, Reading and spelling disorder / dyslexia, Arithmetic disorder, General learning disorder / disorder of scholastic skills, ADD, ADHD / Hyperkinetic Disorder, Hyperkinetic conduct disorder, Conduct disorder.

Coding Scheme for Diagnostic Accuracy

Coding Scheme for Diagnostic Accuracy in Study 2

In study 2 (Bauer et al., 2022), we used the written explanations of learners’ diagnostic reasoning as the data source to determine leaners’ diagnostic accuracy in the six included cases. We developed a coding scheme to operationalise diagnostic accuracy for each learning

case. To measure diagnostic accuracy, we coded all the written diagnoses as accurate (1 point), partially accurate (0.5 points), or inaccurate (0 points).

Example from learning case 4 (Anna).

Category 1: Diagnosis "ADHD - predominantly inattentive type". Accurate answer; 1 point. Subject response contains the diagnosis "ADHD - predominantly inattentive type" or a synonym ("ADS"). Not included: "ADHD", "Combined type".

Category 2: Diagnosis "ADHD - predominantly inattentive type". Partially accurate answer; 0.5 points. Subject response contains one of the diagnoses "ADHD", "ADHD - Combined type".

Coding Scheme for Diagnostic Accuracy in Study 3

In study 3 (Sailer et al., 2023), we used the written explanations of learners' diagnostic reasoning as the data source to determine leaners' diagnostic accuracy in each case. We developed a coding scheme to operationalise diagnostic accuracy for each learning case and each post-test case. We used one category per case, which represented the presence (coded as 1) or absence (coded as 0) of the correct diagnosis in the learners' explanation.

Diagnostic accuracy in the learning process consisted of five categories, each indicating the presence of the correct diagnoses in the written explanation of learning cases numbered two to six. The first learning case was not included in the learning process measurement as the learners received the first feedback after the completion of the cases.

Post-test diagnostic accuracy consisted of two categories, each indicating the presence of the correct diagnoses in the written explanation of both unsupported post-test cases. All participants solved the two post-test cases individually. To obtain the post-test diagnostic accuracy for the dyads, we calculated the mean value for every category for every dyad.

Example from learning case 4 (Anna).

Category 1: Diagnosis "ADHD - predominantly inattentive type". Subject response contains the diagnosis "ADHD - predominantly inattentive type" or a synonym ("ADS"). Not included: "ADHD", "Combined type".

Coding Scheme for the Quality of Justification

In study 3 (Sailer et al., 2023), to determine learners' quality of justifications in each case, we used the written explanations of learners' diagnostic reasoning as the data source. We developed a coding scheme, which is based on expert solutions, to operationalise the quality of diagnostic justifications. We used six categories for each case, which indicated the presence (coded as 1) or absence (coded as 0) of the six primary supporting pieces of

evidence for the correct diagnosis. Based on experts' solutions, employing all six pieces of evidence in a case is considered a high-quality justification.

The quality of justification in the learning process was operationalised by 30 categories. We used learners' explanations from five learning cases (learning cases two to six), in which we used six categories representing the presence or absence of the primary supporting pieces of evidence for the correct diagnosis in each learning case. As the learners received the first feedback after completing the first learning cases, we excluded the first case for the measurement of the quality of justification in the learning process.

Post-test quality of justification consisted of 12 categories, which were assigned in the learners' explanations from two unsupported post-test cases that all students completed individually at the end of the study. In each of these two post-test cases, we used six categories, which were coded for the presence or absence of supporting evidence in the corresponding case. To obtain the quality of justification values for the dyads, we calculated the mean value of each of the 12 categories of every dyad.

List of Categories for an Example Case (Case 4 Anna)

Category 1: Inattention. Subject response contains the cardinal ADD symptom of inattentiveness. Examples: "dreamy, careless, comes too late", "forgets things that she has already learned", "looks out the window", "she is disorganized, her workplace is always chaotic".

Category 2: No hyperactivity and impulsiveness. The test person's response contains that Anna's ADHD symptoms hyperactivity and / or impulsivity are not excessively pronounced. Examples: "She is a nice girl and not impulsive", "Anna is a very calm child". Note: Code "1" is also assigned if only one of the two symptoms (hyperactivity or impulsivity) is explicitly excluded and the other symptom is not mentioned.

Category 3: Reading speed and comprehension slow. Subject answer contains that Anna reads slowly and has difficulty reading comprehension (does not understand the content of what is read). Example: "reads slowly and does not understand the content".

Category 4: No problems with reading accuracy. Subject answer contains that Anna shows no deficits in reading accuracy. Examples: "The reading accuracy is good", "She does not make many mistakes when reading".

Category 5: Selectively good concentration performance with interesting content. Subject answer contains that Anna can concentrate better when she is busy with something interesting. Examples: "is very interested in art and can delve into this area", "she can only

focus for a long time while painting", "she is very talented in art and is completely at it". Not included: "is gifted in art", "art is her passion".

Category 6: Cross-disciplinary performance problems. Subject answer contains that Anna has difficulties in several subjects. Examples: "I already needed tuition for the transfer", "Can't keep up in German and also has problems in English", "She says that she would like to have better grades".

Appendix C

Knowledge Tests

Conceptual Diagnostic Knowledge Test

Description

In study 2 (Bauer et al., 2022), we assessed prior conceptual diagnostic knowledge with a pretest that was administered after the theoretical input. We used 14 single-choice items about diagnosing ADHD and dyslexia, with four answer options each (one correct answer and three distractors). Participants received 1 point per correct answer and, thus, were able to achieve a total of 0 to 14 points of conceptual diagnostic knowledge on the pretest.

Instruction for the Participants

On the following cards, you will be asked several questions. Please read these questions carefully. For every question, there is exactly one right answer. You will not receive any feedback on the correctness of your answer. To save your answer, please make sure to click the button "Abschicken". Otherwise, your answer will not be saved. After submitting your answer, click on the arrow pointing to the right, on the downright side, and you will be forwarded to the next question.

Example Item

Which of the following is <u>not</u> one of the cardinal symptoms of ADHD?

Answer options

(a) Inattentiveness, (b) Hyperactivity, (c) Impulsivity, (d) Impatience

Strategic Diagnostic Knowledge Test

Description

In study 2 (Bauer et al., 2022), we measured prior strategic diagnostic knowledge by using the format of key feature cases (Page et al., 1995). Each item presented the key features of a case as a brief description consisting of a few sentences before asking about the strategic approach used to diagnose the case. We included four key feature cases in total. Two key feature cases were about ADHD and two about dyslexia. The key features presented in the

case information described a school student's behavior and other observations or background information. Two multiple-choice questions were asked per case. One example key feature case introduced the fourth-grader Luis, who has always been a rather poor reader but has begun to fall behind his classmates even more during the last few months and just recently again received the lowest grade in the class on a reading test. He cannot summarize the contents of a short text even right after reading it and can only read aloud very slowly. Apart from his performance issues, he has a chronic disease due to which he cannot regularly attend school for several weeks.

After reading this brief description of a case, the first multiple choice-question asked participants to choose all differential diagnoses that were relevant, given the information presented in the key feature description. As an answer, participants had to pick the relevant differential diagnoses out of a list of seven to 10 options that included clinical as well as nonclinical differential diagnoses. Both the answer options and the number of correct answer options varied across the key feature cases (Case 1 = one correct out of eight options; Case 2 = one correct out of seven options; Case 3 = two correct out of nine options; Case 4 = three correct out of eight options). Participants received 1 point per answer option if they correctly chose a relevant differential diagnosis and correctly did not choose an irrelevant differential diagnosis. Participants received 0 points per answer option if they incorrectly did not choose a relevant differential diagnosis or incorrectly chose an irrelevant differential diagnosis. Accordingly, participants were able to achieve a minimum of 0 points in this first key feature question and a maximum of the number of answer options that each key feature case had (Case 1 = max. 8 points; Case 2 = max. 7 points; Case 3 = max. 9 points; Case 4 = max. 8 points). A mean score was calculated across all answer options of each key feature case, resulting in a diagnosis score of 0 to 1 for the first question for each key feature case.

For every key feature case, the second multiple-choice question asked participants to choose from a list of further approaches and resources that could be used to confirm and/or disconfirm a given set of differential diagnoses for the presented key feature case. Participants had to select the relevant further approaches and resources out of a list of seven to 10 options. Both the answer options and the number of correct answer options varied across the key feature cases (Case 1 = five correct out of eight options; Case 2 = four correct out of seven options; Case 3 = three correct out of seven options; Case 4 = six correct out of 10 options). Participants were again awarded points for correctly choosing relevant options and correctly not choosing irrelevant options and were awarded no points for incorrectly not choosing relevant options or incorrectly choosing irrelevant options. Thus, participants could achieve a

minimum of 0 points on the second question and a maximum of the number of answer options that each key feature case had (Case 1 = max. 8 points; Case 2 = max. 7 points; Case 3 = max. 7 points; Case 4 = max. 10 points). Across all options, we calculated one mean score per key feature case for the second question, resulting in a resources score of 0 to 1 for the second question for each key feature case.

Therefore, with two questions per four key feature cases being scored 0 or 1, participants were able to achieve a range of 0 to 8 overall points for strategic diagnostic knowledge on the pretest.

Instruction for the Participants

On the following cards, you will get to know several pupils. Please read the case descriptions carefully. You will be asked to answer two questions per case. At least one answer is right and there can be multiple right answers. You will not receive any feedback on the correctness of your answer. To save your answer, please make sure to click the button "Abschicken". Otherwise, your answer will not be saved. After submitting your answer, click on the arrow pointing to the right, on the downright side, and you will be forwarded to the next question.

Example Item

At the beginning of the third school year were having a meeting with the parents of your pupil Anja. Lately, her reading and writing performance is very much below the performance of her classmates. Her spelling is often inaccurate and her reading is very slow. The parents mention about the death of Anja's grandfather some weeks ago, which seemed to have thrown the girl off track.

Question 1

Considering the available information, which of the following options may be relevant to consider as possibly applying to this case?

Answer Options Question 1

(a) Specific Reading Disorder, (b) Emotional Stress, (c) Test Anxiety, (d) Specific Spelling Disorder, (e) General Learning Disorder, (f) Motivational Problems, (g) ADHD, (h) Vision Impairment.

Question 2

Which investigations would you initiate to falsify or confirm the diagnoses specific reading disorder, emotional stress, specific spelling disorder and motivation problems?

Answer Options Question 2

(a) Ask her about her current emotional situation, (b) Send her to an oculist, (c) Send her to a school psychologist, (d) Let her take a reading and spelling test with the school psychologist, (e) Reduce her school task demands to see if something changes, (f) Look at her motoric behavior in class, (g) Analyze her past reading and writing competences, (h) Offer after school classes for her in which you can do further investigations.

Appendix D

Feedback Intervention

Adaptive Feedback

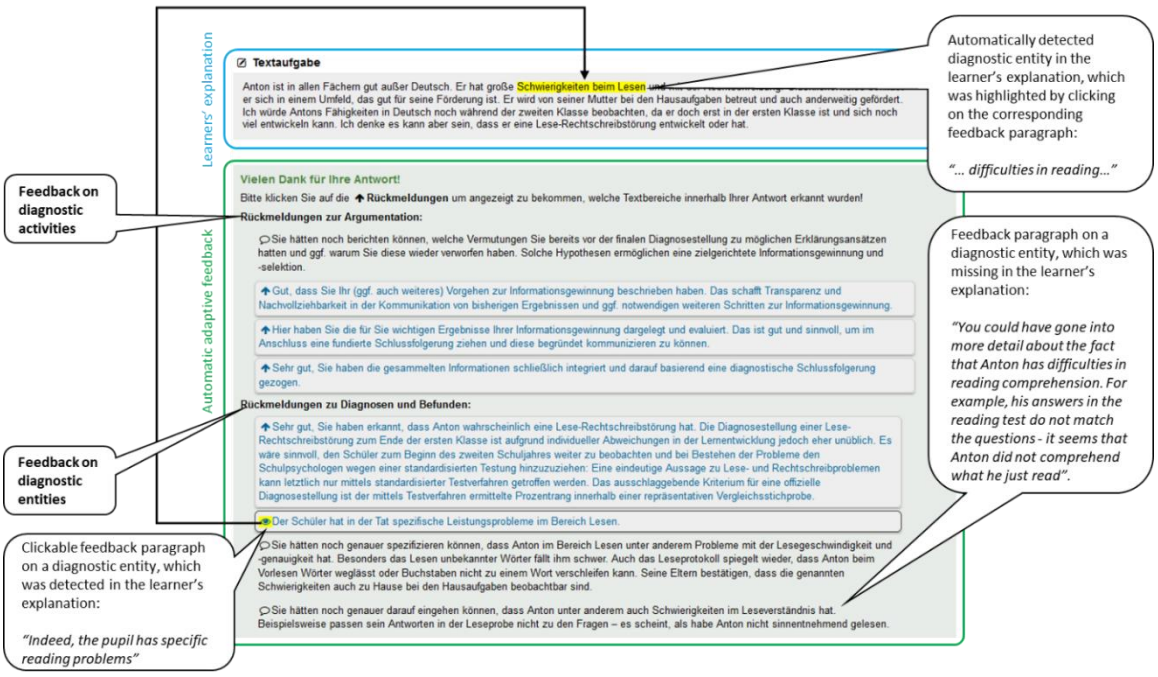
In study 3 (Sailer et al., 2023), learners' diagnostic explanations were analyzed in real-time using NLP. A system consisting of three components (NeuralWeb, INCEpTION, and CASUS) was implemented to automatically and adaptively provide feedback on learners' current diagnostic reasoning (for in-depth explanation see Pfeiffer et al., 2019). In a "cold-start" phase, domain experts coded explanations written by learners of a prior study with N = 118 preservice teachers, who worked on the same simulations in CASUS (see Bauer et al., 2020). The experts used the annotation platform INCEpTION and coded the data according to diagnostic entities (e.g., reading problems, hyperactivity) and epistemic activities (hypothesis generation, evidence generation, evidence evaluation, and drawing conclusions; for details see Schulz et al., 2019). The coded data was used to initially train a predictive model in NeuralWeb, a Python-based web service.

Written explanations of new learners are processed through the NeuralWeb model to output a label-set of the discrete diagnostic classes (diagnostic entities and epistemic activities). Depending on the automatically identified classes, specific paragraphs of predefined feedback text are adaptively activated. The feedback paragraphs were created by domain experts and validated by independent domain experts prior to the study. Experts created approximately 40 feedback paragraphs for every case. For example, these feedback paragraphs inform the learner that a specific symptom was correctly identified in the simulated pupil case. When the corresponding element of a pupil's profile is not detected in a learner's written explanation, the feedback informs the learner that they missed mentioning that symptom. The identified classes and the automatic adaptive feedback, consisting of a range of different feedback paragraphs, are sent back to CASUS. CASUS then presents this adaptive feedback to the user.

The automatic adaptive feedback targets two levels of the learner's written explanation of their diagnostic reasoning: diagnostic activities (whether appropriate reasoning activities were applied or missing) and diagnostic entities (whether the chosen diagnosis and its justification are correct, incorrect, or missing in terms of the domain-specific and case-specific content). By clicking on feedback paragraphs, the relating text part, which was identified in their answer is highlighted (see Figure A3).

Figure A3

Automatic Adaptive Feedback in Casus (see Sailer et al., 2023)

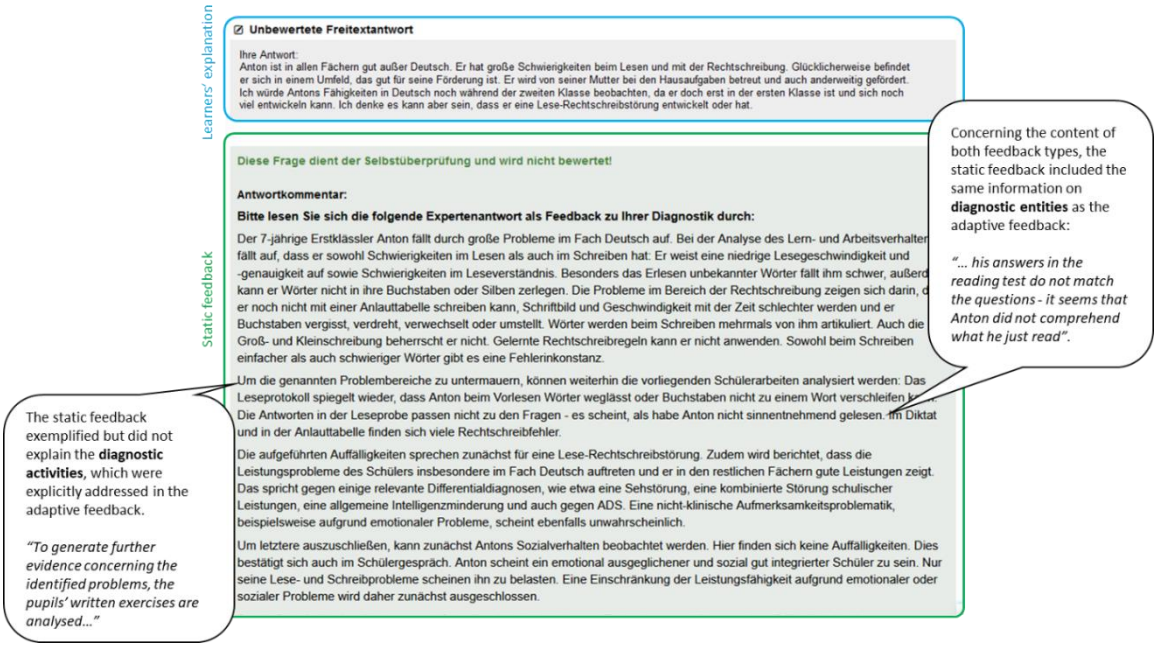


Static Feedback

Learners in the comparison group in study 3 received an expert solution of the case (see Figure A4), after they had entered and justified their diagnosis. Learners were asked to compare it with their own solution. Two independent domain experts validated the expert solutions prior to their use in the study.

Figure A4

Static Feedback in Casus (see Sailer et al., 2023)



Appendix E

Ethical Approval

The studies were approved by the Ethics Committee of the Medical Faculty of LMU Munich (no. 17-249).

Eidesstattliche Versicherung

Statement of Scientific Integrity

Bauer, Anna Elisabeth

Name, Vorname

Last name, first name

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

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

Diagnostic Reasoning and Argumentation: Analysis and Facilitation Using Simulation-Based Learning Environments in Medical Education and Teacher Education

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

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

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

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

München, 03.12.2021

Ort, Datum

Place, Date

Elisabeth Bauer

Unterschrift Doktorandin/Doktorand

Signature of the doctoral candidate