Journal für Mathematik-Didaktik (2024) 45:7
https://doi.org/10.1007/s13138-024-00231-x

ORIGINALARBEIT/ORIGINAL ARTICLE

Urteilsverzerrungen beim Diagnostizieren von Fehlkonzepten bei Dezimalbrüchen

Andreas Rieu · Timo Leuders · Katharina Loibl

Eingegangen: 15. August 2022 / Angenommen: 2. April 2024 / Online publiziert: 26. April 2024

Zusammenfassung Fehlkonzepte von Lernenden zeigen sich als wiederkehrendes Muster bei der Lösung vergleichbarer Aufgaben. Dabei kann ein systematisch auftretender Fehler nicht immer direkt und eindeutig auf ein Fehlkonzept zurückgeführt werden. Diese akkurate Diagnose ist allerdings notwendig, wenn eine Lehrkraft adaptiven Unterricht durchführen möchte. Für eine akkurate Diagnose müssen diagnostisch relevante Informationen z. B. aus (fehlerhaften) Aufgabenlösungen verarbeitet werden. Bei der Informationsverarbeitung können kognitive Verzerrungen (sog. *biases*) auftreten; konkret kann die Mehrdeutigkeit der Situation unberücksichtigt bleiben und die nachfolgende Informationsverarbeitung (die Auswahl weiterer Aufgaben und die Interpretation ihres Diagnosepotenzials) nur im Sinne *eines* angenommen Fehlkonzeptes durchgeführt werden. Die vorliegende Studie untersucht diese Urteilsverzerrung bei diagnostischen Urteilen von angehenden Lehrkräften (N= 79) auf der Ebene der Informationsverarbeitungsprozesse. Zudem wird der Einfluss der Präferenz für Deliberation der urteilenden Person auf diese Informationsverarbeitungsprozesse und deren mögliche Verzerrung untersucht.

Die teilnehmenden Personen bearbeiteten fünf Fallvignetten, in denen sie jeweils mit einer fehlerhaften Lernendenlösung aus dem Bereich Dezimalbruchvergleiche konfrontiert wurden und aufgefordert waren, eine eindeutige Diagnose des vorliegenden Fehlkonzepts zu erstellen. Dazu sollten erste Diagnosehypothesen erstellt und anschließend weitere Aufgaben ausgewählt werden, welche die zu diagnostizierenden Lernenden lösen. Die zur Auswahl stehenden Aufgaben unterschieden sich in ihrer diagnostischen Relevanz. Auf der Grundlage der beschriebenen Modellierung der Urteilsprozesse konnten mit den erhobenen Daten Kategorien der Informationsverarbeitung und der kognitiven Verzerrung (*confirmation bias*) identifiziert und statistisch unterschieden werden.

✉ Andreas Rieu · Timo Leuders · Katharina Loibl
Institut für Mathematische Bildung, Pädagogische Hochschule Freiburg, Freiburg, Deutschland
E-Mail: andreas.rieu@ph-freiburg.de

Als Prädiktor für eine hohe Urteilsakkuratheit konnte die Verarbeitung relevanter Informationen im Laufe des Diagnoseprozesses, nicht aber die Wahrnehmung der Mehrdeutigkeit der Urteilssituation zu Beginn des Prozesses identifiziert werden. Eine Präferenz für deliberate, also bewusst informationsintegrierende Entscheidungen als Personenmerkmal wirkte sich positiv auf die Anzahl formulierter Mehrfachhypothesen aus, hatte allerdings keinen Einfluss auf die Informationssuche und die Akkuratheit der Enddiagnose.

Die Ergebnisse liefern erste Hinweise auf den Einfluss der Urteilsverzerrung bei Urteilen von angehenden Lehrkräften und geben Impulse für die weitere Forschung zum diagnostischen Denken. Daher werden abschließend mögliche Interventionen zur Reduktion von Urteilsverzerrungen bei angehenden Lehrkräften diskutiert.

Schlüsselwörter Diagnostische Kompetenz · Kognitive Prozesse · Urteilsverzerrungen · Fehlkonzepte bei Dezimalbrüchen

Judgment Bias in Diagnosing Misconceptions with Decimal Fractions

Abstract Learners' misconceptions often manifest as recurring patterns when solving comparable tasks. Yet, a systematic error cannot always be directly and unambiguously attributed to a misconception. However, such accurate diagnosis is necessary if a teacher wants to implement adaptive instruction. To achieve accurate diagnosis, diagnostically relevant information must be processed. During information processing, cognitive biases may arise. Specifically, individuals may overlook the ambiguity of the situation and proceed to select further tasks and interpret their diagnostic relevance based solely on one misconception. The present study examines the presence of these judgment biases in the diagnostic judgments of prospective teachers ($N = 79$) by analysing how they process the information. In addition, the influence of the person's preference for deliberation on information processing and possible judgment biases is examined.

The participants worked on five case vignettes in each of which they were confronted with an incorrect student solution in decimal fraction comparisons and were asked to make an unambiguous diagnosis of the misconception at hand. For this purpose, they generated initial diagnostic hypotheses and selected further tasks which the learner to be diagnosed would solve. The diagnostic relevance of the available tasks differed systematically. Based on the described model of judgment processes, it was possible to categorize information processing and cognitive bias (confirmation bias) with the collected data.

In our study, high judgment accuracy could be predicted through the processing of relevant information during the diagnostic process, but not through the perception of the ambiguity in the judgment situation at the beginning of the process. A preference for deliberate decision, i.e., a tendency to consciously integrate information, as a person characteristic had a positive effect on the number of formulated multiple hypotheses but had no effect on the selection of tasks and the accuracy of the final diagnosis.

The results provide initial evidence on the impact of judgment bias in teacher judgments and provide impetus for further research on diagnostic thinking. Therefore, we conclude with a discussion of possible interventions to reduce judgment bias in prospective teachers.

Keywords Diagnostic competence · Cognitive processes · Judgement bias · Misconceptions about decimal fractions

1 Einführung

Fehlkonzepte sind Vorstellungen oder Vorwissen von Lernenden, die sich vom Expertenwissen in dieser Domäne unterscheiden. Sie treten bei der Bearbeitung von Aufgaben als systematische Fehler zu Tage (Kuo et al. 2018; Nesher 1987; Smith et al. 1994). Dabei können Fehlkonzepte den Lernprozess von Schülerinnen und Schülern behindern. Aus Sicht der Lehrkraft bietet die Diagnose dieser Fehlkonzepte eine wichtige Gelegenheit, adaptive Maßnahmen einzuleiten (Klieme 2020), da die Fehlkonzepte das Denken und die Vorstellungen der Schülerinnen und Schüler aufzeigen. Da bei manchen Lerngegenständen verschiedene Fehlkonzepte auftreten können, ist eine möglichst eindeutige und akkurate Diagnose des einem Fehler zugrunde liegenden Fehlkonzeptes durch die Lehrkraft von entscheidender Bedeutung, um angemessene Hilfestellungen zu geben und systematische Förderung zu planen (Beck et al. 2008; Corno 2008; Klieme 2020; Tomlinson et al. 2003).

Auch wenn Urteile in pädagogischen Situationen und ihre Akkuratheit seit geraumer Zeit stärker ins Forschungsinteresse gerückt sind (Heitzman et al. 2019; Herppich et al. 2018; Leuders et al. 2018; Loibl et al. 2020; Südkamp et al. 2012), so mangelt es noch an belastbaren Erklärungen für die interindividuell stark variierende Urteilsakkuratheit von Lehrkräften bei der Einschätzung ihrer Schülerinnen und Schüler. Hier können Modelle zur Informationsverarbeitung bei der Genese der Urteile in spezifischen Situationen Erklärungen liefern (Ingenkamp und Lissmann 2008; Loibl et al. 2020). Durch die Fokussierung auf die Urteilsgenese, also beispielsweise die Untersuchung des Einflusses von Wissensfacetten oder von Merkmalen der Urteilssituation auf die kognitiven Urteilsprozesse, konnte in verschiedenen Studien gezeigt werden, dass wissensbasierte Prozesse pädagogischen Urteilen zugrunde liegen und die Vermittlung oder die Bereitstellung von spezifischem fachdidaktischem Wissen die Urteilsakkuratheit positiv beeinflusst (Binder et al. 2018; McElvany et al. 2009; Ostermann et al. 2018; Rieu et al. 2022). Neben (nicht) vorhandenem Wissen wird die Informationsverarbeitung als ein weiterer Aspekt bei der Genese diagnostischer Urteile fokussiert, um Erklärungen für individuelle Unterschiede in der Urteilsakkuratheit zu liefern: Diagnostische Urteile als Rückschluss von beobachtbarem Lernendenverhalten auf latente Eigenschaften der Lernenden sind unsicherheitsbehaftet, weil das Lernendenverhalten in der Regel nicht deterministisch oder nicht genügend informationshaltig ist (Leuders und Loibl 2021). Diagnose ist somit zu verstehen als Prozess der Unsicherheitsreduktion (Heitzman et al. 2019). Dabei – und in Anlehnung an Studien aus der Medizin, der Sozialpsychologie und der Wirtschaftswissenschaften (Golman et al. 2017) – konnte

auch bei schulischen Diagnosen und gerade in komplexen Urteilssituationen gezeigt werden, dass verschiedene Urteilsverzerrungen stattfinden, welche sich negativ auf die Urteilsakkuratheit auswirken (Moser Opitz und Nührenbörger 2015; Oswald und Grosjean 2004).

In der vorliegenden Studie fokussieren wir auf mehrdeutige Diagnosesituationen und auf die in diesen Situationen auftretende Urteilsverzerrungen bei der Informationsverarbeitung. Dies wird bei der Diagnose von Fehlkonzepten am Beispiel von Dezimalbrüchen untersucht, da hier in der Regel eine mehrdeutige Diagnosesituation vorliegt: Ein auftretender Lernendenfehler kann seine Ursache in unterschiedlichen Fehlkonzepten des Lernenden haben; die diagnostizierende Lehrkraft sollte also zu Beginn von mehreren Erklärungshypothesen ausgehen. Zur Bestimmung des vorliegenden Fehlkonzeptes müssen sodann durch die Auswahl geeigneter Aufgaben weitere Informationen gewonnen werden. Konkret bedeutet dies, dass beim Größenvergleich zweier Dezimalbrüche („Ist 3,92 größer, kleiner oder gleich 3,4813?“) eine richtige Lösung der Aufgabe nicht unbedingt bedeutet, dass ein echtes Verständnis für Dezimalbrüche vorliegt, sondern auch einer entsprechenden Fehlvorstellungen (LIK) geschuldet sein kann. Darüber hinaus kann eine fehlerhafte Lösung nicht eindeutig auf ein Fehlkonzept zurückzuführen ist, sondern das Ergebnis verschiedener Fehlkonzepte sein kann (hier: Komma-trennt- oder Kein-Komma-Strategie; Padberg und Wartha 2017; Stacey 2005). Zur eindeutigen Diagnose des vorliegenden Fehlkonzepts benötigt die Lehrkraft weitere Informationen. Diese kann sie in einem Gespräch mit den Lernenden erfahren (Padberg und Wartha 2017) oder die Lehrkraft wählt weitere Aufgaben aus einem Lehrbuch, einem Testbogen, etc. zur Bearbeitung aus, um auf Basis der erhaltenen Lösungen oder durch Beobachtung des Lösungsprozesses die nötigen Informationen für eine eindeutige Diagnose zu erhalten.

Diese Prozesse des Hypothesenbildens und der Informationssuche in mehrdeutigen Situationen können in Anlehnung an soziale Urteilsprozesse als soziales Hypothesentesten modelliert werden (Heitzman et al. 2019; Schons et al. 2022; Schulz-Hardt und Köhnken 2000; Trope und Liberman 1996), da auch hier Urteile über Personeneigenschaften auf der Grundlage von beobachtbaren Verhaltensweisen getroffen werden, welche a priori kein eindeutiges Urteil zulassen. So kann bspw. die Einschätzung der politischen Überzeugung einer Person anhand ihrer Handlungsmuster vorgenommen werden, wobei unterschiedliche Ausprägungen dieser politischen Überzeugungen denkbar sind und die beobachteten Handlungen nicht stets eindeutig einer einzigen politischen Überzeugung zugewiesen werden können. Ein konsequentes Ausschalten der Zimmerbeleuchtung kann beispielsweise sowohl für eine ökologische Überzeugung als auch für rein finanzielle Interessen einer Person mit anderen politischen Einstellungen stehen (Dreger 2012). Im spezifizierten Fall der Beobachtung von systematischen Fehlern bei Lernenden in unserer Studie geht die Lehrkraft zunächst vom Vorliegen von einem oder mehreren möglichen Fehlkonzepten aus und verwirft oder bestätigt diese Hypothese dann aufgrund der weiteren Informationen. Dabei ist anzunehmen, dass einige der dabei durchlaufenen kognitiven Prozesse angehender Lehrkräfte einer Verzerrung unterliegen: So könnte eine Lehrkraft zunächst die Mehrdeutigkeit der Situation verkennen und so nur *ein* mögliches Fehlkonzept als Ursache für die fehlerhafte Lösung in Betracht ziehen.

Die diagnostizierende Person würde dann im nächsten Schritt weitere Aufgaben auswählen, die lediglich die Bestätigung aber nicht die Prüfung der Hypothese ermöglichen, und sie könnte abschließend widersprüchliche Informationen konfirmatorisch als Beleg für die erste Hypothese interpretieren. Im konkreten Fall der Diagnose von Fehlkonzepten bei Dezimalbrüchen würde eine weitere Aufgabe, die in ihrer Struktur vergleichbar zur bereits fehlerhaft gelösten Aufgabe ist (z. B. verschiedene Zahlwerte, aber mit jeweils derselben Anzahl an Nachkommastellen in Verbindung mit demselben Verhältnis) nur die Bestätigung einer Hypothese (z. B. die größere Anzahl an Nachkommastellen wird als Beleg für einen kleineren Dezimalbruch verstanden) ermöglichen und nicht deren Prüfung, da eine solche Aufgabe auch nach der Lösung durch Lernende keine weiteren für die Diagnose relevanten Informationen liefert. Andere Aufgaben – zumindest bei konsistentem Lösungsverhalten der Lernenden – können aufgrund ihres diagnostischen Potentials hingegen weitere relevante Informationen liefern. Im Diagnoseprozess werden die Informationen, die durch die Bearbeitung verschiedener Aufgaben vorliegen, integriert. Eine deliberate, also bewusst informationsintegrierende – im Gegensatz zu einer intuitiven – Verarbeitung erkennt und berücksichtigt die Relevanz vorliegender Informationen bezüglich der Mehrdeutigkeit der Situation und des Aufgabengehaltes und führt zu einer akkuraten Diagnose des vorliegenden Fehlkonzeptes (s. Abschn. 2.2 und 2.4).

Obwohl die beschriebenen Annahmen über Urteilsverzerrungen plausibel erscheinen und im Einklang mit Ergebnissen der sozialen Urteilsbildung stehen, gibt es kaum Forschung, die die stattfindenden kognitiven Prozesse bei Lehrkräften konkret modelliert, empirisch untersucht und so (mögliche) Erklärungen für das bekannte Phänomen der variierenden Urteilsakkuratheit liefert. Ziel der vorliegenden Studie ist die Untersuchung des Einflusses von konfirmatorischen Urteilsverzerrungen auf die Urteilsprozesse bei der Wahrnehmung der Diagnosesituation und der Auswahl weiterer Aufgaben. Für diese empirische Untersuchung wurde ein kontrolliertes und simuliertes Setting gewählt, in welchem die formulierten Hypothesen zum vorliegenden Fehlkonzept und die ausgewählten Aufgaben als Prozessindikatoren erfasst werden. Es wird überprüft, ob diese Prozessindikatoren entsprechend der Modellannahmen zu den kognitiven Prozessen mit der Urteilsakkuratheit bei der Diagnose von Fehlkonzepten zusammenhängt. Die Ergebnisse sollen Rückschlüsse auf die Genese von diagnostischen Urteilen zulassen.

2 Theoretischer Rahmen und Forschungsstand

Das Erkennen von Lernendenfehlkonzepten durch Lehrkräfte wird als Teil ihres Professionswissens verstanden und stellt die Grundlage für adaptiven Unterricht dar (Krauss et al. 2008). Studien zur diagnostischen Kompetenz von Lehrkräften der letzten Jahre fokussierten vorrangig deren Urteilsakkuratheit – die Ergebnisse zeigen allerdings eine große Varianz (siehe die Meta-Analysen von Südkamp et al. 2012, Urhahne und Wijnia 2021). Die für die Meta-Analysen ausgewählten korrelativen Studien liefern keine Erklärungen für diesen Befund, da die Genese von Urteilen und die zugrundeliegenden kognitiven Prozesse bisher kaum explizit untersucht wurden. Aktuellere Studien fokussieren daher die Informationsverarbeitung bei der

Genese von Urteilen. Auf Basis von theoretischen Annahmen zu dem Einfluss von personalen und situativen Charakteristika auf die Informationsverarbeitung (Loibl et al. 2020) bietet diese Forschungsstrategie die Möglichkeit, die angenommene kognitive Verzerrung zu untersuchen. Dieser Forschungsstrategie folgend wird in der vorliegenden Studie eine komplexe, da mehrdeutige Urteilssituation aus dem Mathematikunterricht – die Diagnose von Fehlkonzepten bei Dezimalbrüchen – in einem kognitiven Modell dargestellt, theoretische Annahmen zur möglichen Urteilsverzerrung (bei der Wahrnehmung der Situation und bei der Informationsverarbeitung) erstellt und anschließend empirisch validiert.

2.1 Diagnose von Fehlkonzepten

Der Lernprozess von Schülerinnen und Schülern wird entscheidend von der Qualität ihres vorhandenen Vorwissens geprägt. Entspricht dieses Vorwissen der fachlichen Norm (Präkonzept), so können weitere Wissenselemente verknüpft werden; falls das Vorwissen allerdings im Widerspruch zu den gelehrten Inhalten steht und systematisch auftritt, wird es als Fehlerstrategie, systematischer Fehler oder Fehlkonzept bezeichnet (Nitsch 2015, Prediger und Wittmann 2009). Dabei liegen Fehlkonzepte zunächst als eine kognitive Abweichung von der fachlichen Norm vor und treten häufig als fehlerhafte Aussage in Erscheinung, wie z. B. die Übergeneralisierung der Vorstellung aus den natürlichen Zahlen in den Bereich der rationalen Zahlen, dass ein Produkt immer größer ist als die Zahlen, die miteinander multipliziert werden (Ay 2017; Baur 2018; Prediger 2008). Solche Aussagen von Lernenden geben Hinweise auf das Denken und die Vorstellungen der Lernenden, die aufgrund ihrer Erfahrungen in verschiedenen Kontexten entstanden sind (Fujii 2020). Da sich diese Vorstellungen für die Lernenden in anderen Bereichen oder Kontexten als nützlich oder funktionsfähig erwiesen haben (im Beispiel bei der Multiplikation natürlicher Zahlen), sind Fehlkonzepte keine trivialen und leicht behebbaren Fehler. Fehlkonzepte können allerdings den Lernprozess behindern und es ist somit wichtig für Lehrkräfte, diese Fehlkonzepte zu diagnostizieren, um eine gezielte Förderung einleiten zu können (Bradshaw und Templin 2014). Darüber hinaus ist die akkurate Diagnose der individuell vorliegenden Fehlkonzepte eine notwendige Bedingung, um adaptive Lernangebote für alle Lernenden zu schaffen, indem z. B. unterrichtliche Aktivitäten an die unterschiedlichen Lernbedürfnisse der einzelnen Schülerinnen und Schüler angepasst werden (Corno 2008; Tomlinson et al. 2003).

In der Mathematikdidaktik sind Fehlkonzepte in einigen Bereichen bereits gut erforscht (Confrey und Kazak 2006), wie z. B. die Annahme, Multiplikation vergrößere und Division verkleinere in der Bruchrechnung, der Graph-als-Bild-Fehler in der Algebra oder die Vorstellung, dass Dezimalbrüche kleiner werden, je mehr Nachkommastellen sie besitzen. Häufig auftretende Fehlkonzepte beim Vergleich von Dezimalbrüchen stehen im Fokus der hier vorgestellten Studie und werden daher nachfolgend vertieft dargestellt.

In Bezug auf den Größenvergleich von Dezimalbrüchen zeigt beispielsweise die Längsschnittstudie von Steinle und Stacey (1998), dass die exemplarisch ausgewählten und in Tab. 1 dargestellten Fehlkonzepte weit verbreitet sind, von Lernenden als Strategien bei der Lösung von Aufgaben mit Dezimalbrüchen eingesetzt werden

Tab. 1 Ausgewählte, häufig vorkommende Fehlkonzepte beim Dezimalbruchvergleich (nach Padberg und Wartha 2017) mit einer Übersicht diagnostischer Aufgaben inklusive der korrekten und fehlerhaften Lösungen. Diese Aufgaben wurden auch in der Studie verwendet

Fehlkonzept	Auftrittshäufigkeit in Jahrgangsstufe 5 (Desmet et al. 2010; Steinle und Stacey, 1998; Padberg und Wartha, 2017)	Aufgabenlösungen unter Verwendung der Strategie
Komma-trennt-Strategie (KT) Bei diesem Fehlkonzept werden Dezimalbrüche als Zusammensetzung zweier, durch ein Komma getrennter, natürlicher Zahlen (eine vor dem Komma, eine dahinter) interpretiert und entsprechend getrennt verrechnet. Hiernach würde 0,3 < 0,29 angenommen, da die Lernenden mit diesem Fehlkonzept 3 < 29 sehen	Ca. 30 %	0,03 < 0,029 ✗ 4,08 > 4,7 ✗ 1,006 < 1,53 ✓ 0,2 < 0,30 ✓
Kein-Komma-Strategie (KK): Bei diesem Fehlkonzept wird das Komma ignoriert und die Zahlen betrachtet, als ob es sich um natürliche Zahlen handelt. Hiernach würde 1,7 < 1,65 angenommen, da die Lernenden mit diesem Fehlkonzept 17 < 165 sehen	Ca. 30 %	0,03 < 0,029 ✗ 4,08 > 4,7 ✗ 1,006 > 1,53 ✗ 0,2 < 0,30 ✓
Länger-ist-kleiner-Strategie (LIK): Bei diesem Fehlkonzept wird nur die Stellenanzahl nach dem Komma betrachtet. Hiernach würde 0,375 < 0,25 angenommen, denn je mehr Dezimalstellen vorhanden sind, desto kleiner ist die Zahl. Dieses Fehlkonzept könnte aufgrund einer Übergeneralisierung des richtigen Gedankens zustande kommen, dass Tausendstel kleiner als Hundertstel sind. Eine weitere naheliegende Erläuterung könnte aus den Erfahrungen mit natürlichen Zahlen stammen, wobei die Hinzunahme einer Stelle die Zahl stets vergrößert. Übergeneralisiert und als Anpassung für die Dezimalbrüche könnten Lernende annehmen, dass jede weitere Stelle die Zahl stets verkleinert	Ca. 15 %	0,03 > 0,029 ✓ 4,08 < 4,7 ✓ 1,006 < 1,53 ✓ 0,2 > 0,30 ✗

Tab. 1 (Fortsetzung)

Fehlkonzept	Auftrittshäufigkeit in Jahrgangsstufe 5 (Desmet et al. 2010; Steinle und Stacey, 1998; Padberg und Wartha, 2017)	Aufgabenlösungen unter Verwendung der Strategie
Nullstrategie (NLS) Bei diesem Fehlkonzept werden Nullen direkt rechts des Dezimalkommas als Indikator für eine kleine Zahl angesehen. Hiernach würde 1,006 < 1,53 angenommen, da die Lernenden wissen, dass Dezimalbrüche mit einer Null nach dem Komma klein sind. Dieses Fehlkonzept liefert häufig richtige Ergebnisse, tritt oft in Kombination mit der KT-Strategie und dann vor allem bei Vergleichen von drei Dezimalbrüchen (0,009 < 0,03 < 0,029) auf. So ist die fehlerhafte Lösung 0,03 < 0,029 als Kombination von NLS- und KT-Strategie zu erklären	Ca. 12 %	0,03 < 0,029 ✗ 4,08 < 4,7 ✓ 1,006 < 1,53 ✓ 0,2 < 0,30 ✓

und zu Schwierigkeiten beim Aufbau eines validen Operationsverständnisses führen können (Mosandl und Sprenger 2014). Aufgrund ihrer weiten Verbreitung wurden diese Fehlkonzepte im Rahmen der hier vorliegenden Studie fokussiert. Die daraus entstehende Mehrdeutigkeit der Diagnosesituation – verschiedene Fehlkonzepte führen bei manchen Aufgaben zur gleichen Lösung – geht ebenfalls aus der Tabelle hervor.

Das Ziel von Diagnose und Förderung der Lernenden ist der Aufbau von Kompetenzen in einer Domäne. Dazu sollten Lehrkräfte aus inklusiver Sicht sowohl die Potenziale als auch Schwierigkeiten auf individueller und klassenübergreifender Ebene in den Blick nehmen (Prediger und Zindel 2017). Ein wichtiger Hinweis auf ein vorliegendes Fehlkonzept liefert ein konsistent über strukturell vergleichbare Aufgaben auftretender und somit systematischer Fehler. Dennoch gibt eine einzelne (fehlerhafte) Aufgabenlösung weder einen verlässlichen Hinweis auf einen systematischen Fehler noch bestimmt sie eindeutig das ursächliche Fehlkonzept. Zur eindeutigen Diagnose müssen daher idealerweise mehrere geeignete Aufgaben von den Lehrkräften ausgesucht und von den Lernenden bearbeitet werden, um aus den Lösungen die Informationen zur eindeutigen Diagnose von möglicherweise vorliegenden Fehlkonzepten zu gewinnen.

Die Entstehung von diagnostischen Urteilen und deren Akkuratheit in mehrdeutigen Situationen oder bei der Diagnose von Fehlkonzepten sind kaum erforscht. Die bisher vorliegenden Befunde stammen aus qualitativen Studien, in denen die Situation des diagnostischen Gespräches zwischen der Lehrkraft und den zu diagnostizierenden Lernenden untersucht wurde (z. B. Morris et al. 2009; Philipp 2018; Prediger und Zindel 2017). Ein Ziel der vorliegenden Studie ist es, die Entstehung des diagnostischen Urteils auf der Grundlage der durch die Lehrkraft verarbeiteten Informationen zu modellieren: Die Lehrkraft diagnostiziert in der beschriebenen Situation das Lernendendenken, welches durch das Auftreten des Fehlkonzepts cha-

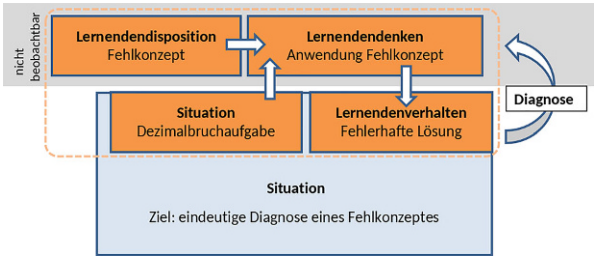

Abb. 1 Diagnose als Urteil über das Lernendendenken (hier: ein Fehlkonzept im Bereich der Dezimalbrüche) in Form eines Rückschlusses von manifestem Verhalten (Aufgabenlösungen) auf latente Eigenschaften oder Prozesse. (Nach Leuders und Loibl 2021)

rakterisiert wird. Dabei nimmt sie Informationen aus der Aufgabe und der fehlerhaften Lösung wahr und verarbeitet diese zu einer Diagnose (s. Abb. 1, in Anlehnung an Leuders und Loibl 2021).

2.2 Urteilsakkuratheit und mögliche Verzerrungen

Grundvoraussetzung für eine akkurate Diagnose – beim Vorhersagen von Aufgabenlösungen, bei der Beurteilung des Leistungsniveaus einer Lernendengruppe oder bei der Diagnose von individuellen Lernvoraussetzungen und -ständen (Südkamp und Praetorius 2017) – ist, dass die Lehrkraft ihr spezifisches Wissen über individuelle Eigenschaften der Lernenden anwendet.

Üblicherweise werden zur Untersuchung der Urteilsakkuratheit von Lehrkräften deren Einschätzungen mit Lernendentests (für Leistung) oder mit Selbstberichten der Schülerinnen und Schüler (für motivationale Merkmale) korreliert. In ihrer Meta-Analyse haben Südkamp et al. (2012) gezeigt, dass die Urteilsakkuratheit der Lehrkräfte bei Lernendenleistungen stark variiert und durchschnittlich eher ungenau ist (mittlere Korrelation zwischen vorausgesagter und tatsächlicher Leistung r= 0,63).

Die Urteilsakkuratheit gibt dabei zwar die Übereinstimmung der Lehrkrafteinschätzung mit den Leistungen einzelner Lernenden wieder, liefert aber keine Hinweise auf den Diagnoseprozess selbst (Südkamp und Praetorius 2017; Herppich et al. 2018). Neuere Studien verwenden die Urteilsakkuratheit weiterhin als Kenngröße, legen ihr Forschungsinteresse allerdings auf die beim Diagnostizieren stattfindenden Prozesse und generieren so ein umfassenderes Verständnis dessen, wie eine Diagnose zustande kommt und welche Einflüsse auf die Urteilsprozesse wirken. Hier wird untersucht, welche Informationen Lehrkräfte tatsächlich sammeln und verarbeiten, um ihr diagnostisches Urteil zu bilden (Becker et al. 2020; Brunner et al. 2021; Oudman et al. 2018; Rieu et al. 2022; Schreiter et al. 2021; van de Pol et al. 2021).

Loibl et al. (2020) konzeptualisieren diagnostische Urteile in pädagogischen Kontexten als Inferenzen einer Lehrkraft über Lernende (z. B. deren Fähigkeiten) oder Materialien (z. B. Aufgabenschwierigkeit), die auf den Informationen beruht, die explizit oder implizit in einer diagnostischen Situation vorhanden sind. Diese Definition verortet diagnostische Urteile innerhalb des größeren Bereichs des sozialen Urteilens und der Theorien der kognitiven Informationsverarbeitung. Die bereits beschriebene Varianz in der Urteilsakkuratheit von Lehrkräften wird in Studien auch auf Urteilsverzerrungen *(biases)* zurückgeführt, wie z. B. Halo- oder Pygmalion-

Effekte (Goldstein, 2002; Tobisch und Dresel 2017). Die meisten Belege stammen allerdings aus Studien zu Persönlichkeitsurteilen in den Bereichen der ethnischen Zugehörigkeit, des sozioökonomischen Status und des Geschlechts (McKown und Weinstein 2008; Rubie-Davies et al. 2006; Südkamp et al. 2012). Die genannten Verzerrungen finden auf der Ebene der Informationsverarbeitung statt: unter anderem werden vorhandene Informationen nur zur Verifizierung einer erstellten Annahme, nicht aber zu deren Falsifizierung gedeutet. Experimentelle Studien liefern Hinweise zu Unterschieden bei der Informationsverarbeitung in komplexen Situationen, allerdings im Vergleich von Novizen und Experten (Dünnebier et al. 2009, Van Ophuysen 2006). So zeigte z. B. Van Ophuysen (2006), dass erfahrene Lehrkräfte die Schullaufbahnempfehlung von Lernenden flexibler und differenzierter als Studierende treffen und dabei auch widersprüchliche Informationen berücksichtigen. Eine Ausdehnung dieser Ergebnisse auf den Einfluss kognitiver Verzerrungen bei der Informationsverarbeitung im Rahmen von diagnostischem Denken und Urteilen steht noch aus.

Herppich et al. (2018) postulieren, dass diagnostische Urteile im pädagogischen Kontext als soziales Hypothesentesten dargestellt werden können und die Informationssuche heuristische Tendenzen aufweisen kann. Dieser Rückgriff auf Heuristiken als vereinfachende Regeln kann das Auftreten einer Verzerrung bei der Wahrnehmung, der Suche und der Interpretation von Informationen beinhalten und wirkt sich auf die Urteilsakkuratheit aus (Tobisch und Dresel 2017; Westhoff und Kluck 2014).

Eine bereits gut untersuchte Urteilsverzerrung ist der c*onfirmation bias*, wonach Menschen Informationen selektiv verarbeiten. Der *confirmation bias* resultiert aus dem Versäumnis, verfügbare Informationen kritisch zu analysieren und entsteht vor allem, wenn sich Menschen in mehrdeutigen Situationen befinden, die entweder als konsistent oder inkonsistent mit der aktuell favorisierten Hypothese interpretiert werden können (Nickerson 1998; Oswald und Grosjean 2004). Dabei zeigte Nickerson (1998), dass Menschen sich vor allem mit Informationen befassen, die sie in ihrer ursprünglichen Meinung bestätigen. Dies hat zur Konsequenz, dass widersprüchliche Informationen weniger häufig ausgewählt und seltener verarbeitet und erinnert werden (Fiedler, 1997). Der negative Einfluss des *confirmation bias* auf die Urteilsfindung konnte in unterschiedlichen Bereichen der menschlichen Entscheidungsfindungen, wie beispielsweise bei militärischen Pilotenfehlern (Stewart 2008), bei medizinischen Diagnosen (Tschan et al. 2009) und bei Investitionsentscheidungen (Bashir et al. 2013) nachgewiesen werden. Aber auch bei alltäglichen Entscheidungen wie der Einschätzung von Personenmerkmalen oder politischen Entscheidungen konnten konfirmatorische Verzerrungen aufgezeigt werden (Gatlin et al. 2019; Oswald und Grosjean 2004). Da der *confirmation bias* daraus resultiert, dass verfügbare Informationen nicht kritisch analysiert werden, ist es naheliegend anzunehmen, dass das Auftreten des *confirmation biases* durch Personenmerkmale, wie die Präferenz für intuitive oder reflektive Entscheidungen, beeinflusst wird (Betsch 2004; Epstein et al. 1996).

Nach Betsch (2004) sind Intuition und Deliberation persönliche Entscheidungsmodi, die zwei voneinander unabhängige Dimensionen darstellen. Dabei wird Intuition als rein affektiver Modus verstanden, während Deliberation eine reflektive und kognitionsbasierte Entscheidungsfindung darstellt (s. auch Witteman et al.

2009). Diese Unterscheidung geht auf die kognitionspsychologischen Arbeiten im Bereich der dualen Entscheidungsfindungsprozesse (*dual process*) zurück (Epstein et al. 1996, Tversky und Kahneman 1983). Ein wichtiger Unterschied ist allerdings, dass Intuition als rein affektiver und nicht heuristischer Modus dargestellt wird. In durchgeführten Studien (Betsch 2004; Epstein et al. 1996) zeigte sich eine positive Korrelation zwischen dem Personenmerkmal der Intuitionspräferenz und schnellem Entscheiden. Eine Präferenz für Deliberation korreliert hingegen mit Gewissenhaftigkeit und Perfektionismus. In der genannten Studie von Betsch (2004) wird darüber hinaus noch eine Personengruppe identifiziert, die situationsabhängig, also je nach Situation entweder deliberat oder intuitiv entscheidet.

Die hier vorliegende Studie soll erste Erkenntnisse dazu liefern, ob eine kognitive Verzerrung wie der beschriebene *confirmation bias* auch im Bereich von diagnostischem Denken und Entscheiden vorkommen, ob spezifische Personenmerkmale die Urteilsprozesse beeinflussen und in welcher Weise sich diese mögliche kognitive Verzerrung auf die Informationsverarbeitung auswirkt. In der untersuchten Situation kann eine konfirmatorische Verzerrung dazu führen, dass (a) die Mehrdeutigkeit der Situation verkannt wird und (b) nur Aufgaben zur Bestätigung einer Hypothese gewählt werden.

2.3 Urteilsprozess bei der Diagnose von Fehlkonzepten bei Dezimalbrüchen

Die Diagnose von Fehlkonzepten beim Dezimalbruchvergleich zeichnet sich durch eine hohe Komplexität aufgrund der beschriebenen Mehrdeutigkeit aus. Es müssen Hypothesen zu den möglicherweise vorliegenden Fehlkonzepten (z. B. Kommatrennt-, Kein-Komma-, Länger-ist-kleiner- und Nullstrategie nach Padberg und Wartha, 2017 bzw. Stacey 2005) erstellt werden und diese dann anhand weiterer Informationen, also zusätzlich gelöster Aufgaben, belegt oder verworfen werden. So kann der Urteilsprozess der Lehrkräfte bei der Diagnose des vorliegenden Fehlkonzeptes in Anlehnung an das soziale Hypothesentesten (Trope und Liberman 1996) als wissensbasierter Prozess der Informationsverarbeitung (Loibl et al. 2020) beschrieben werden (s. Abb. 2).

Konkret kann die fehlerhafte Antwort von Schülerinnen und Schülern, dass 3,92 kleiner als 3,4813 sei, auf zwei unterschiedliche Fehlkonzepte zurückgeführt werden (s. Tab. 1), nämlich

- auf die Komma-trennt-Strategie, wobei die Lernenden Dezimalbrüche als zwei natürliche Zahlen betrachten, die durch das Komma getrennt werden und in dieser Situation $92 < 4813$ lösen.

oder

- auf die Kein-Komma-Strategie, wobei die Lernenden bei Dezimalbrüchen das Komma ignorieren und sie als ganze Zahlen, also $392 < 34813$, vergleichen.

Erkennt eine Lehrkraft vor dem Hintergrund ihres Wissens zu den verschiedenen Fehlkonzepten die Mehrdeutigkeit der Situation, kann sie als weitere Diagnoseauf-

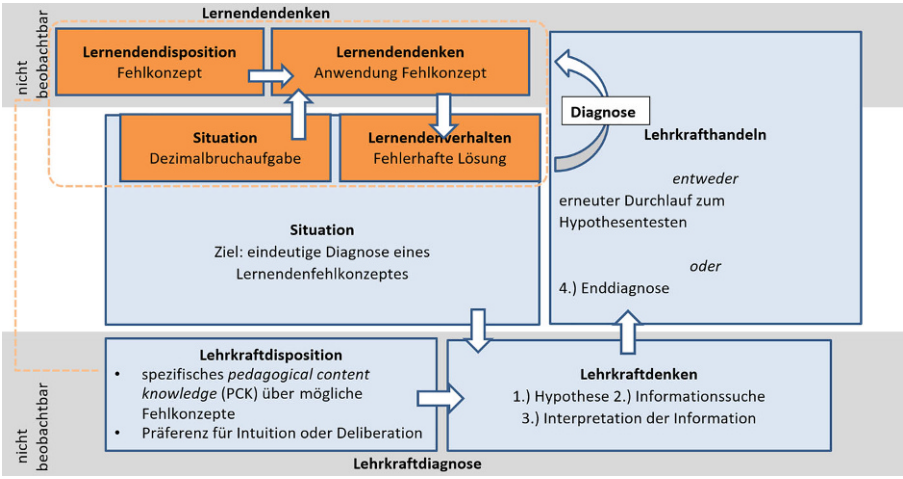

Abb. 2 Darstellung der informationsverarbeitenden Prozesse bei der Diagnose eines Personenmerkmals (hier: ein Fehlkonzept im Bereich der Dezimalbrüche). (Nach Leuders und Loibl (2021) in Anlehnung an das soziale Hypothesentesten (Trope und Liberman 1996). Nicht direkt beobachtbar und allenfalls erhebbar sind einerseits die Dispositionen und Denkprozesse der Lernenden sowie andererseits die Dispositionen und das Denken (hier: Urteilsprozesse) der Lehrkraft. Die Diagnose stellt dabei eine beobachtbare Beurteilung des Lernendendenkens innerhalb einer spezifischen Situation durch eine Lehrkraft dar)

gaben eine Aufgabe wählen, die zwischen diesen beiden Fehlkonzepten trennen kann. So würde beispielsweise der Vergleich von 1,006 und 1,53 mit der Komma-trennt-Strategie richtig gelöst (1,006 < 1,53) und mit der Kein-Komma-Strategie falsch gelöst (1,006 > 1,53). Die Mehrdeutigkeit der Diagnosesituation entsteht durch die fehlerhafte, aber nicht eindeutig einem Fehlkonzept zuzuordnenden Lernendenlösung. Durch die Auswahl diagnostisch relevanter Aufgaben kann die Lehrkraft jedoch eine eindeutige Diagnose erstellen.

Da ein Dezimalbruchvergleich aufgrund unterschiedlicher Fehlkonzepte fehlerhaft gelöst werden kann (Padberg und Wartha 2017; Stacey 2005), muss akkurates diagnostisches Denken als rekursiver Prozess stattfinden (Herppich et al. 2018): Zunächst werden die in dieser Situation möglichen Fehlkonzepte hypothetisch als Fehlerursache angenommen. Das Belegen oder Verwerfen von Hypothesen findet wissensbasiert unter Hinzunahme weiterer Informationen statt. Dazu kann die Lehrkraft den zu diagnostizierenden Lernenden weitere Aufgaben zur Bearbeitung vorlegen. Die nun zusätzlich erhaltenen Informationen (spezifische Aufgabenlösungen) können bereits ausreichend sein, um eine Hypothese zu validieren, also um das Fehlkonzept eindeutig zu diagnostizieren. Im Falle von bestehender Unsicherheit wird die Lehrkraft idealerweise noch weitere Aufgaben lösen lassen, um genügend Informationen zum Verwerfen oder Validieren einer (weiteren) Hypothese zu erhalten (in Anlehnung an Schulz-Hardt und Köhnken 2000).

Zusammengefasst kann festgestellt werden, dass in der mehrdeutigen Situation der Diagnose von Fehlkonzepten drei wissensbasierte Prozesse stattfinden: Zunächst das Analysieren der Situation zum Erkennen von möglichen Mehrdeutigkeiten, dann in mehreren Zyklen die Informationssuche über die Auswahl neuer Aufgaben und

die anschließende Interpretation der Informationen vor dem Hintergrund der erstellten Hypothesen (Trope und Liberman 1996). Nach dem Erkennen, dass mehrere Ursachen für ein auftretendes Fehlkonzept in Frage kommen, stellen die beiden letztgenannten Prozesse das eigentliche Testen der Hypothesen dar, wobei es sich hierbei um ein rekursives Vorgehen handeln kann. Abschließend wird durch die integrative Informationsverarbeitung ein *finales Urteil* gefällt. Erst wenn die diagnostizierende Lehrkraft aufgrund der verarbeiteten Informationen ausreichend sicher ist, wird sie das vorliegende Fehlkonzept diagnostizieren.

2.4 Konfirmatorische Verzerrungen bei der Diagnose von Fehlkonzepten bei Dezimalbrüchen

Eine mögliche Erklärung für die festgestellten, großen Variationen bei der Urteilsakkuratheit von Lehrkräften (Südkamp et al. 2012) können stattfindende Urteilsverzerrungen liefern (Tobisch und Dresel 2017), die in der vorliegenden Studie in der Situation der Diagnose von Fehlkonzepten untersucht werden. Eine Urteilsverzerrung kann angenommen werden, da die beschriebene, mehrdeutige Urteilssituation belastend für das Arbeitsgedächtnis sein kann (Hinson et al. 2003) und es daher plausibel scheint, dass Personen (a) die Mehrdeutigkeit dieser Situation nicht erkennen und (b) Informationen selektiv für die Erhaltung einer ersten Hypothese auswählen und verarbeiten (*confirmation bias*, Oswald und Grosjean 2004). In Anlehnung an den Prozess des sozialen Hypothesentestens werden im Folgenden die kognitiven Prozesse der Informationsverarbeitung der Lehrkraft und möglichen Verzerrungen postuliert (Schulz-Hardt und Köhnken 2000) und zusammenfassend in Abb. 3 dargestellt:

1. Ersthypothese

Lehrkräfte, die die Mehrdeutigkeit der diagnostischen Situation wahrnehmen, erstellen idealerweise unter Verwendung von oder unter Rückgriff auf *pedagogical content knowledge* (PCK) eine Mehrfachhypothese, da eine einzige fehlerhafte gelöste Aufgabe keine eindeutige Diagnose zulässt.

Bei verzerrter Wahrnehmung der Situation erstellen die Personen hingegen nur eine einzige oder eine nicht tragfähige Hypothese.

2. Informationssuche

Nach Formulierung einer Mehrfachhypothese werden weitere Aufgaben ausgewählt, deren Lösungen als Beleg oder als Widerspruch zu einem der angenommenen Fehlkonzepte dienen. Die (ungelösten) Aufgaben unterscheiden sich in der Anzahl der Ziffern, der Anzahl der Nachkommastellen, der Präsenz von Nachkomma-Nullen oder der Anzahl an Nachkomma-Nullen. Aufgrund der erstellten Hypothese(n) bezüglich des möglicherweise vorliegenden Fehlkonzeptes überprüft die Lehrkraft gedanklich, anhand welcher weiteren Aufgabe die Hypothese überprüft wird bzw. zwischen verschiedenen Hypothesen differenziert werden kann. Die zur Verfügung

stehenden Aufgaben sind – je nach vorhandenen Informationen und Diagnosekontext – relevant oder irrelevant für die Diagnose.

Falls die untersuchte Lehrkraft nur eine Hypothese aufgestellt hat und somit weitere mögliche Fehlkonzepte verkennt, wird sie bedingt durch diese Prämisse weitere Aufgaben auswählen, die nicht unbedingt Belege für das Vorliegen anderer möglicher Fehlkonzepte liefern. Es besteht für diesen Fall ebenfalls die Möglichkeit, dass weitere Aufgabenlösungen zum Verwerfen der Ersthypothese führen oder auf ein anderes Fehlkonzept hinweisen, wenn zufällig eine Aufgabe ausgewählt wird, die zwischen zwei möglichen Fehlkonzepten differenzieren kann.

3. Interpretation der Information

Lehrkräfte, die eine Mehrfachhypothese formuliert haben, interpretieren bei diagnostischen Aufgaben die Lernendenlösungen als Beleg für eine der Hypothesen und als Widerspruch zur anderen Hypothese. Bei der Verarbeitung von nicht-diagnostischen Aufgaben kann weiterhin nicht zwischen den Hypothesen entschieden werden. Fällt die Lehrkraft auf dieser Basis ein abschließendes Urteil, ist dies zufällig akkurat oder falsch.

Nach der Erstellung einer einzelnen Hypothese können vier Fälle unterschieden werden, drei bei Auswahl diagnostisch relevanter Aufgaben, einer bei Auswahl diagnostisch irrelevanter Aufgaben:

- Falls die erste Einzelhypothese zufällig dem gesuchten Fehlkonzept entspricht und diagnostisch relevante Aufgaben ausgewählt werden, werden die Lösungen der weiteren Aufgaben als Beleg für die Hypothese interpretiert.
- Falls die erste Einzelhypothese dem gesuchten Fehlkonzept nicht entspricht und diagnostisch relevante Aufgaben ausgewählt werden, widersprechen die Aufgabenlösungen der Hypothese. Interpretiert die Lehrkraft diese Informationen auch als Widerspruch, verwirft sie ihre Ersthypothese und erstellt eine besser an die Situation angepasste Hypothese.
- Alternativ kann die die Lehrkraft die Lösungen konfirmatorisch verzerrt als Beleg für die Ersthypothese interpretieren oder ignorieren.
- Werden diagnostisch irrelevante Aufgaben ausgewählt, widersprechen die Aufgabenlösungen der Ersthypothese nicht – unabhängig von deren Richtigkeit. Die Lehrkraft interpretiert die Aufgabenlösungen als Beleg für die zufällig richtige oder falsche Hypothese.

Die Prozesse der Informationssuche und der Interpretation der Informationen müssen nicht einmalig stattfinden, sondern können sequenziell mehrfach durchlaufen werden, insbesondere bei der Überprüfung mehrerer Hypothesen.

4. Enddiagnose

Die Enddiagnose ist das beobachtbare Verhalten der untersuchten Lehrkraft. Sie kann unabhängig von den stattfindenden kognitiven Prozessen als akkurat oder falsch eingeteilt werden.

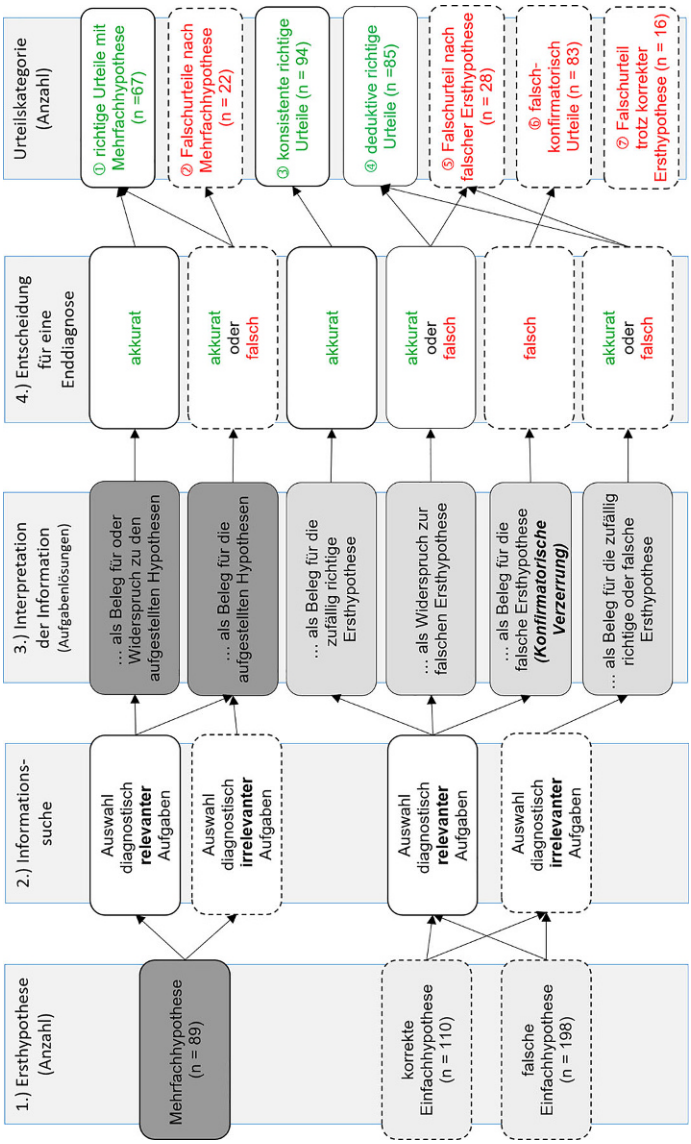

Abb. 3 Darstellung der möglichen Informationsverarbeitung bei der Diagnose von Fehlkonzepten bei Dezimalbrüchen ausgehend von der Formulierung einer Mehrfachhypothesen (*dunkelgrau*) oder einer Einfachhypothese (*hellgrau*). (Die *durchgezogenen Ränder* der *Kästen* stehen für die normative Informationsverarbeitung, die *gestrichelten Ränder* stehen für erwartbare, aber nicht optimale Prozesse. Darüber hinaus stehen die *dicken Ränder* der *Kästchen* für beobachtbares Verhalten, während *dünne Ränder* interne Prozesse darstellen)

Die hier vorliegende Studie untersucht, ob sich interindividuelle Unterschiede bei der Diagnose von Fehlkonzepten auf die beschriebene kognitive Urteilsverzerrung im Sinne des *confirmation bias* zurückführen lassen. Zusammenfassend können die Annahmen zu kognitiven Prozessen und der Einfluss der Urteilsverzerrung bei der Diagnose von Fehlkonzepten bei Dezimalbrüchen folgendermaßen illustriert werden (Abb. 3).

Abb. 3 zeigt die möglichen Verhaltensweisen von Lehrkräften bei der Wahrnehmung, Informationsverarbeitung und Diagnose von Fehlkonzepten bei Dezimalbrüchen. Darüber hinaus dient die Darstellung zur Ableitung der Forschungsfragen, der aufgestellten Forschungshypothesen und der vorgenommenen Kategorisierung der Urteilsprozesse.

3 Forschungsfragen

Die diagnostische Kompetenz ist eine wichtige Facette der professionellen Kompetenz von Lehrkräften und kann bereits während der Ausbildung gestärkt werden, allerdings variiert sie interindividuell stark (Heitzman et al. 2019; Herppich et al. 2018; Loibl et al. 2020; Südkamp et al. 2012). Die vorliegende Studie untersucht Urteilsprozesse von angehenden Lehrkräften, um Erklärungswissen zu dieser Varianz zu generieren (Loibl et al. 2020; Rieu et al. 2022). Theoretische Grundlage ist die Modellierung der kognitiven Prozesse, die bei der oben beschriebenen mehrdeutigen Situation stattfinden. So kann das diagnostische Denken in mehrdeutigen Unterrichtssituationen theoretisch beschrieben werden und durch die Erhebung externer Indikatoren belegt werden (Loibl et al. 2020).

Konkret wird untersucht, ob bei der mehrdeutigen diagnostischen Situation der Feststellung von Fehlkonzepten beim Dezimalbruchvergleich im Sinne des sozialen Hypothesentestens und bei der Auswahl weiterer Aufgaben zur Bearbeitung eine konfirmatorische Verzerrung auftritt, die eine akkurate Diagnose verhindert. Aufgrund der modellierten Urteilssituation (Abb. 2) und den theoretischen Annahmen (Abb. 3) können folgende Forschungsfragen abgeleitet werden:

> „Lässt sich die Urteilsakkuratheit durch die Formulierung einer Mehrfachhypothese bei der ersten Hypothesenbildung und durch die Auswahl von diagnostisch relevanten Aufgaben bei der Informationssuche vorhersagen?“

Aufgrund der in Abb. 3 dargestellten Annahmen zu den Urteilsprozessen wird davon ausgegangen, dass bei der Informationsverarbeitung beim Diagnostizieren von Fehlkonzepten im Bereich der Dezimalbrüche eine Urteilsverzerrung stattfindet und diese die Urteilsakkuratheit reduziert. Konkret wird angenommen, dass (H1) die Formulierung einer Mehrfachhypothese einerseits und (H2) die Anzahl von Aufgaben und die Auswahl von diagnostisch relevanten Aufgaben andererseits Prädiktoren für die Urteilsakkuratheit sind. Außerdem kann vermutet werden, dass die Urteilssicherheit je nach Urteilsprozess variiert und wird daher zusätzlich zur Urteilsakkuratheit untersucht.

Da der Einfluss von Personenmerkmalen in der sozialen Urteilsforschung bereits nachgewiesen werden konnte (Betsch 2004; Epstein et al. 1996) und die vorliegen-

de Studie in Anlehnung an das soziale Hypothesentesten modelliert wurde, soll der Einfluss von Personenmerkmalen untersucht werden. Dabei wird fokussiert, welchen Einfluss die Präferenz einer beurteilenden Person für schnelle, intuitive bzw. deliberate und somit informationsintegrierende Entscheidungen in einer komplexen pädagogischen Urteilssituation auf die Wahrnehmung der Situation und die Informationsverarbeitung ausübt.

> „Inwiefern beeinflusst die Präferenz einer beurteilenden Person für intuitive oder deliberate Urteile die Wahrnehmung der Urteilssituation als mehrdeutig, die Informationsverarbeitung und die Urteilsakkuratheit?“

Zur Beantwortung dieser Forschungsfrage wird die Wahrnehmung der Urteilssituation und die Informationssuche (d. h. die Auswahl diagnostisch relevanter oder irrelevanter Aufgaben) in Abhängigkeit der Ergebnisse des PID-Fragenbogens (Präferenz für Intuition oder Deliberation; Betsch 2004) untersucht.

Zusammenfassend dienen einerseits die erhobene PID-Typisierung der teilnehmenden Personen und andererseits die erstellte(n) Ersthypothese(n), die Anzahl und die Art der verarbeiteten Informationen im Rahmen der Aufgabenauswahl (Informationssuche) und die Urteilsakkuratheit der Enddiagnose als Grundlage für die Kategorisierung der kognitiven Urteilsprozesse. Ergebnisse, die den hier formulierten Annahmen folgen, werden als empirischer Beleg der Plausibilität der angenommenen Urteilsprozesse verstanden.

4 Materialien und Methode

Zur Überprüfung dieser theoretischen Annahmen wurde ein Design gewählt, in welchem angehende Lehrkräfte Vignetten mit je einer fehlerhaft gelösten Aufgabe einer Schülerin oder eines Schülers aus dem Bereich des Dezimalbruchvergleichs vorgelegt bekommen (vgl. Abb. 4). Der Kontext ist an einer alltäglichen Diagnosesituation im Klassenverband oder in einer Gruppenarbeit angelehnt, in welcher die Lehrkraft einen systematischen Fehler feststellt und zur Diagnose auf weitere Aufgaben zurückgreift, die aus dem Lehrbuch oder einem Testbogen stammen. In der hier untersuchten Situation sollen weder eigene, diagnostische Aufgaben erstellt noch die Möglichkeit eines vertiefenden Diagnosegesprächs operationalisiert werden, um das diagnostische Aufgabenpotenzial (Kron et al. 2021) kontrollieren zu können. Zur Erfassung der diagnostischen Kompetenz der angehenden Lehrkräfte in einer mehrdeutigen Urteilssituation und zu deren empirischer Untersuchung wurde daher ein simuliertes und kontrolliertes Setting gewählt.

Das Diagnoseziel war, das zugrundeliegende Fehlkonzept eindeutig zu bestimmen. Da die angezeigte, fehlerhaft gelöste Aufgabe aufgrund mehrerer Fehlkonzepte zustande kommen kann, konnten die teilnehmenden Personen weitere Aufgaben auswählen, welche die zu diagnostizierenden Lernenden konsistent nach dem

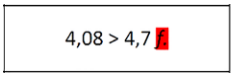

Abb. 4 Fehlerhaft gelöste Aufgabe aus dem Bereich des Dezimalbruchvergleichs

ihnen zugrundeliegende Fehlkonzept lösten. Die teilnehmenden Personen wurden darüber informiert, dass die ihnen gezeigten Diagnosefälle auf virtuelle Lernende zurückzuführen sind und diese die Aufgaben immer unter Verwendung des gleichen Fehlkonzeptes lösen. Bei realen Schülerinnen und Schülern müssten auch zufällige Lösungen bedacht werden, was hier aus methodischen Gründen ausgeschlossen wurde.

Bei der in Abb. 4 gezeigten Aufgabe führt sowohl die Kein-Komma-Strategie als auch die Komma-trennt-Strategie zur falschen Lösung. Lernende, die die Nullstrategie oder die Länger-ist-Kleiner-Strategie anwenden, kommen hier trotz bzw. wegen ihres Fehlkonzeptes zur richtigen Lösung, das vorliegende Fehlkonzept bliebe eventuell unentdeckt (Padberg und Wartha 2017). Dieses Beispiel zeigt die Mehrdeutigkeit und somit die Komplexität der Diagnosesituation für Lehrkräfte und unterstreicht die Notwendigkeit des spezifischen fachdidaktischen Wissens (PCK).

4.1 Stichprobe

Die Erhebung wurde mit Studierenden des Primarstufenlehramts an der Pädagogischen Hochschule Freiburg durchgeführt. Alle teilnehmenden Personen studierten Mathematik als Hauptfach. Das durchschnittliche Alter betrug 21,43 Jahre (SD = 2,70) und zum Zeitpunkt der Erhebung befanden sich die Studierenden im 2. Semester (SD = 1,03), also zu Beginn ihres Studiums. Diese Personengruppe wurde bewusst ausgewählt, um mögliche Störvariablen zu kontrollieren: Es kann davon ausgegangen werden, dass den teilnehmenden Studierenden weder durch ihr Studium noch aufgrund von Praxiserfahrungen fachwissenschaftliches oder fachdidaktisches Wissen im Bereich der Diagnose zu Fehlkonzepten bei Dezimalbrüchen vorliegt, da Dezimalbrüche erst im Mathematikunterricht der Sekundarstufe behandelt werden. Durch die Übersicht über Fehlkonzepte, also die Darstellung des spezifischen, notwendigen PCK, soll dazu beigetragen werden, dass allen teilnehmenden Personen während der Intervention dieselbe spezifisch fachdidaktische Wissensbasis zur Verfügung steht. Bei der untersuchten Kohorte ist von einer geringen Varianz (z. B. durch Nachhilfeunterricht) auszugehen.

4.2 Durchführung der Studie und Fragebogendesign

Die Datenerhebung erfolgte computerbasiert. Wie in Abb. 5 ersichtlich wurde vor der eigentlichen Datenerhebung der Fragebogen zur Präferenz für Intuition oder Deliberation (PID) von Betsch (2004) durchgeführt. Er misst die Präferenz von Menschen für Intuition und Deliberation mit zwei Skalen: der PID-I für Intuition und die PID-D für Deliberation. Die 19 Items mit einer 5-stufigen Likert-Skala (1 = stimme nicht zu, 5 = stimme voll zu) wurden in zufälliger Reihenfolge vorgegeben. Es handelt sich dabei um Aussagen mit Alltagsbezug (z. B. „Ich ziehe Schlussfolgerungen lieber aufgrund meiner Gefühle, Menschenkenntnis und Lebenserfahrung.“, „Ich denke erst nach, bevor ich handle.“).

Anschließend waren standardisierte aber in unterschiedlicher Reihenfolge dargebotene Fälle ($N = 7$) zu diagnostizieren. In den Fällen lösen virtuelle Lernende einen Dezimalbruchvergleich fehlerhaft. Durch Anklicken von weiteren Aufgaben lösen

1.Teil

Allgemeine Erhebung – Fragebogen PID

Bitte antworten Sie auf die folgenden Fragen möglichst ehrlich.

	Stimme überhaupt nicht zu	Stimme eher nicht zu	weder noch	Stimme eher zu	Stimme voll und ganz zu
Bevor ich Entscheidungen treffe, denke ich meistens erst mal gründlich nach.	○	○	○	○	○
Ich beobachte sorgfältig meine innersten Gefühle.	○	○	○	○	○
Bevor ich Entscheidungen treffe, denke ich meistens erst mal über meine Ziele nach, die ich erreichen will.	○	○	○	○	○
Bei den meisten Entscheidungen ist es sinnvoll, sich ganz auf sein Gefühl zu verlassen.	○	○	○	○	○
Ich mag Situationen nicht, in denen ich mich auf meine Intuition verlassen muss.	○	○	○	○	○
Ich denke über mich nach.	○	○	○	○	○
Ich schmiede lieber ausgefeilte Pläne, als etwas dem Zufall zu überlassen.	○	○	○	○	○

2. Teil

Diagnose von Fehlkonzepten anhand von Fallvignetten

Aufgaben im Dezimalbruchvergleich können aufgrund fehlerhafter Schülerlösungen die vorliegenden Fehlkonzepte aufzeigen. Grundsätzlich werden 4 Fehlerstrategien unterschieden (Wartha & Padberg, 2017):

Komma-trennt-Strategie (KT-Strategie): Dezimalbrüche werden als Zusammensetzung zweier Zahlen (eine vor dem Komma, eine dahinter) interpretiert und entsprechend getrennt verrechnet. Hiernach würde 0,3 < 0,29 angenommen, da die Lernenden mit dieser Fehlerstrategie 3 < 29 sehen.

Kein-Komma-Strategie (KK-Strategie): Bei dieser Fehlerstrategie wird das Komma ignoriert und die Zahlen betrachtet, als ob es sich um natürliche Zahlen handelt. Hiernach würde 1,7 < 1,65 angenommen, da die Lernenden mit dieser Fehlerstrategie 17 < 165 sehen.

Länger-ist-kleiner-Strategie (LIK-Strategie): Diese Fehlerstrategie ist eine falsche Übergeneralisierung eines richtigen Gedankens, denn

Allgemeine Intervention: stets sichtbare, schriftliche Darstellung der typischen Fehlkonzepte bei Dezimalbrüchen (spezifisches PCK) – Übersicht über Fehlkonzepte

fehlerhaft gelöste Aufgabe

4,08 > 4,7

Textfeld Ersthypothese

Welche der Fehlerstrategien könnte bei Bente vorliegen?

Schieberegler Urteilssicherheit

Wie sicher sind Sie sich dabei?
Ziehen Sie den Regler mit der Maus an die gewünschte Stelle.

Sicherheit der Hypothese

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Auswahl weiterer Aufgaben aus 4er-Block zur Bearbeitung für Lernenden

Sie können Bente nun weitere Aufgaben lösen lassen.

0,3 □ 0,4	3,72 □ 3,073	3,27 □ 3,003	1,006 □ 1,53
Diese Aufgabe lösen lassen	Diese Aufgabe lösen lassen	Diese Aufgabe lösen lassen	Diese Aufgabe lösen lassen

Können Sie bereits das vorliegende Fehlkonzept diagnostizieren? Falls Sie „nein" auswählen werden Ihnen weitere Aufgaben für die Diagnose zur Verfügung gestellt.

○ ja
○ nein

Textfeld Enddiagnose

Welche Fehlerstrategie liegt bei Bente vor?

Notieren Sie Ihre Diagnose:

Schieberegler Urteilssicherheit

Wie sicher ist Ihre Enddiagnose?
Ziehen Sie den Regler mit der Maus an die gewünschte Stelle.

Sicherheit der Enddiagnose

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Abb. 5 Darstellung des Ablaufs der Studie

die Lernenden diese ebenfalls – und zwar immer konsistent nach dem ihnen zugrundeliegenden Fehlkonzept. Das zur Diagnose notwendige fachdidaktische Wissen (Rieu et al. 2020) über die typischen Fehlkonzepte bei Dezimalbrüchen wurde den teilnehmenden Personen anhand eines Textes zu den typischen Fehlkonzepten zu jedem Zeitpunkt des Urteilsprozesses eingeblendet (Übersicht über Fehlkonzepte). In allen Diagnosesituationen lässt sich die initiale fehlerhafte Lösung auf den ersten Blick nicht eindeutig auf ein Fehlkonzept zurückführen, da bei den ausgewählten Aufgaben jeweils verschiedene Fehlkonzepte zur fehlerhaften Lösung führen. Es handelt sich daher um eine mehrdeutige Diagnosesituation. Die verwendeten, fehlerhaft gelösten Aufgaben der Diagnosesituationen können im Anhang eingesehen werden.

Die angehenden Lehrkräfte formulierten – anhand des möglichst offen formulierten Auftrages *Welche der Fehlerstrategien könnte bei (Name des Lernenden) vorliegen?* – zunächst eine erste Hypothese zum vorliegenden Fehlkonzept, wobei eine freie Texteingabe die Nennung von einer oder mehrerer Hypothesen als Ersthypothese ermöglichte. Außerdem gaben sie über einen Schieberegler ihre prozentuale Hypothesensicherheit an. Anschließend konnten sie durch Anklicken weitere Aufgaben auswählen und erhielten zu diesen Aufgaben die Lösungen der zu diagnostizierenden Schülerinnen oder Schüler (Informationssuche).

Die zur Auswahl angebotenen Aufgaben wurden jeweils zu viert abgebildet (Abb. 6). Die vier Aufgaben unterschieden sich hinsichtlich der Relevanz ihrer Informationen für die Diagnose. In einem 4er-Block von auszuwählenden Aufgaben befanden sich immer zwei Aufgaben mit diagnostisch relevanten Informationen für die Diagnosesituation. Aufgaben mit relevanten Informationen für die Diagnose liefern in der mehrdeutigen Diagnosesituation Belege für ein mögliches Fehlkonzept und stehen im Widerspruch mit einem anderen möglichen Fehlkonzept. Zwei weitere Aufgaben erhielten keine zusätzlich relevanten Informationen: Diese Aufgaben werden entweder mit jedem Fehlkonzept (im hier vorliegenden Untersuchungssetting stets) korrekt gelöst oder die Aufgaben entsprechen in der Struktur und somit in dem Informationsgehalt der bereits zu Beginn der Vignette gezeigten Aufgabe. Die Platzierung der relevanten und irrelevanten Aufgaben änderte sich von 4er-Block zu 4er-Block.

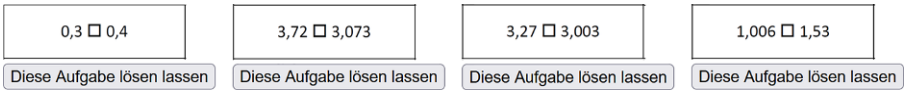

Abb. 6 Darstellung der zur Auswahl angebotenen Aufgaben, welche sich in ihrem diagnostischen Potenzial unterscheiden. (Die beiden *linken Aufgaben* liefern bei der gegebenen fehlerhaft gelösten Aufgaben 4,08 > 4,7 keine zusätzlichen diagnostischen Informationen, da sie entweder unabhängig vom vorhandenen Fehlkonzept der Lernenden richtig gelöst werden (*1. Aufgabe von links*) oder vom Aufgabentyp der bereits angezeigten Aufgabe entsprechen (*2. Aufgabe von links*, jeweils gleiche Anzahl bzw. gleiches Verhältnis an Nachkomma-Nullen bzw. an Nachkommastellen). Die beiden *Aufgaben rechts* liefern zusätzliche diagnostische Informationen, da sie zwischen verschiedenen möglichen Fehlkonzepten (Komma-trennt- bzw. Kein-Komma-Strategie) aufgrund der jeweils unterschiedlichen Anzahl und unterschiedlicher Verhältnisse der Nachkommastellen trennen können. Die Anordnung der irrelevanten und der relevanten Informationen wurde im Laufe der Erhebung variiert)

Die teilnehmenden Personen entschieden frei über die Anzahl weiterer gestellter Aufgaben, indem sie (a) Aufgaben aus einem 4er-Block frei auswählten und (b) nach der Sichtung der Lösungen weitere Aufgaben zur Bearbeitung in weiteren 4er-Blöcken ansehen oder direkt zur Abgabe der Enddiagnose schreiten konnten. Falls sich die Personen für einen weiteren 4er-Block an Aufgaben entschieden, wurde ein neuer Block angezeigt, in welchem sich wiederum zwei Aufgaben mit diagnostischem Potenzial und zwei Aufgaben ohne zusätzliche diagnostische Informationen befanden (s. Anhang). Auch bei der Enddiagnose wurde die Diagnosesicherheit anhand eines Schiebereglers mit Prozentangaben erhoben. Es wurden keinerlei zeitliche Restriktionen formuliert, die teilnehmenden Personen wurden lediglich aufgefordert eine akkurate Diagnose zu treffen.

4.3 Datenerfassung

Als soziodemographische Daten und mögliche Kovariaten wurden das Alter, das Geschlecht und die Praxiserfahrung im schulischen Kontext wie auch in Nachhilfesettings der teilnehmenden Personen erfasst. Die Erhebung der Präferenz für Intuition oder Deliberation (PID) geschah über den oben abgebildeten Fragebogen (Abb. 5). Bei der Auswertung des PID gelten Personen mit Intuitionswerten über dem Median und Deliberationswerten unter dem Median als intuitive Urteilende (Typ-I), Personen mit hohen Deliberations- und niedrigen Intuitionswerten werden als deliberate Typen (Typ-D) charakterisiert. Falls die Scores auf beiden Skalen hoch oder niedrig ausfallen, werden diese Personen als situationsabhängiger Typus (Typ-S) bezeichnet (Betsch 2004).

In Anlehnung an die erstellen Modellannahmen und die angenommene konfirmatorische Urteilsverzerrung werden folgende Variablen erfasst.

Bei der Erstellung der Ersthypothese:

Welche Art der Ersthypothese – Einfach- oder Mehrfachhypothese – wird formuliert? Entspricht die Ersthypothese dem vorliegenden Fehlkonzept (bei Formulierung einer Einzelhypothese)?

Während der Informationssuche:

Hierbei wird die Annahme verfolgt, dass die untersuchten angehenden Lehrkräfte aus den ihnen angebotenen Aufgaben solche Aufgaben auswählen, von deren Lösung sie bedeutsame Informationen erwarten. Daher werden angeklickte Aufgaben gleichgesetzt mit verarbeiteten Informationen.

- Wie viele Aufgaben werden von den teilnehmenden Personen angeklickt und damit verarbeitet?
- Welche Art (diagnostisch relevant oder irrelevant) der Informationen verarbeiten die teilnehmenden Personen anteilig an der Anzahl angeklickter Aufgaben?

Zur Kategorisierung der Urteilsprozesse werden die Anzahl der verwendeten Aufgaben als absoluten Wert angegeben, während der Anteil der ausgewählten diagnostischen Aufgaben als relativer Wert gemessen an der Anzahl aller verarbeiteter Aufgaben angegeben wird. In der hier vorliegenden Studie sollen Aussagen über auftretende Urteilsprozesse gemacht werden; der Anteil diagnostischer Aufgaben

gemessen an allen verwendeten Aufgaben ist ein Indikator für systematische oder verzerrte Urteilsprozesse.

Bezüglich der formulierten Enddiagnose:

- Ist die Enddiagnose akkurat?
- Entspricht die Enddiagnose der Ersthypothese (bei Formulierung einer Einzelhypothese)?
- Wie hoch schätzen die teilnehmenden Personen ihre Urteilssicherheit bei Erstellung der Enddiagnose ein?

4.4 Datenauswertung

Zur Beantwortung der übergreifenden Forschungsfrage, ob bei der mehrdeutigen diagnostischen Situation der Feststellung von Fehlkonzepten beim Dezimalbruchvergleich eine kognitive Verzerrung auftritt, werden fallbezogene, deskriptive Statistiken berechnet, um die Urteilsprozesse anhand der erstellten Ersthypothese, der Prozesse der Informationssuche und der Enddiagnose zu kategorisieren und insbesondere den (idealen) normativen Prozess und den konfirmatorisch verzerrten Prozess zu vergleichen.

Darüber hinaus wurden verallgemeinerte lineare gemischte Modelle zur Bestimmung der Effekte der Art der Ersthypothese und der Anzahl und Art der verwendeten Aufgaben (Informationssuche) auf die Akkuratheit der Enddiagnose erstellt. Dieses Vorgehen bietet die Möglichkeit, jeden individuellen Urteilsprozess der angehenden Lehrkräfte fallweise zu untersuchen und somit der Tatsache gerecht zu werden, dass personenspezifische Charakteristika und Fallcharakteristika Auswirkungen auf den Urteilsprozess haben (Brauer und Curtin 2017; Brown 2021). Der Einfluss des Personenmerkmals der Präferenz für Intuition oder Deliberation (PID) auf die Wahrnehmung der Situation, die Informationssuche und die Urteilsakkuratheit wurden korrelativ untersucht.

Alle teilnehmenden Personen, die das Erhebungstool vollständig bearbeitet hatten, wurden in die Auswertung einbezogen.

5 Ergebnisse

In Tab. 2 werden die Häufigkeiten der akkuraten Enddiagnosen gezeigt. Dabei haben die 79 teilnehmenden angehenden Lehrkräfte, die den Fragebogen komplett ausgefüllt haben, einen (LIK) bzw. zwei Fälle (KK, KT und NLS) pro zugrundeliegendem Fehlkonzept bearbeitet. Trotz der Möglichkeit auf spezifisches fachdidaktisches Wissen über die typischen Fehlkonzepte anhand der Übersicht über Fehlkonzepte zurückgreifen zu können, wurden häufig falsche Diagnosen erstellt, die Urteilsakkuratheit beträgt über alle Fälle 47,3 % (SD = 17,9).

Dabei fällt die Diagnose der Nullstrategie (NLS) besonders inakkurat aus. Beim Vergleich der durchschnittlichen Akkuratheit der Diagnose der Nullstrategie (8 %) mit der Diagnosegenauigkeit aller anderen Fälle zeigen sich hochsignifikante Unterschiede ($F(561) = 29{,}1$, $p \leq 0{,}001$), die als Bodeneffekte (Döring und Bortz 2016)

Tab. 2 Übersicht über die Urteilsakkuratheit bei der Enddiagnose von Fehlkonzepten im Bereich der Dezimalbrüche

	Alle Fälle	Vorliegendes Fehlkonzept: LIK (1 Fall)	Vorliegendes Fehlkonzept: KK (2 Fälle)	Vorliegendes Fehlkonzept: KT (2 Fälle)	Vorliegendes Fehlkonzept: NLS (2 Fälle)
n	553	79	158	158	158
Durchschnittlich richtige Diagnosen (SD)	47,3 % (17,9)	68,4 % (46,8)	80,4 % (39,8)	43,8 % (49,8)	8,0 % (26,8)
Durchschnittlich benötigte Diagnosezeit in Sekunden (SD)	1557 (328)	159 (10)	136 (53)	215 (109)	214 (45)

interpretiert werden. Dieses Fehlkonzept scheint aufgrund seiner Abhängigkeit von der Komma-trennt-Strategie und der häufig vorkommenden richtigen Lernendenantworten (Padberg und Wartha 2017) in diesem Kontext besonders schwierig zu diagnostizieren. Padberg und Wartha (2017) schlagen zur Diagnose dieser Strategie den Größenvergleich von drei Brüchen vor, welcher dann fehlerhafte Ergebnisse wie $0{,}009 < 0{,}03 < 0{,}029$ liefert. Da in der Studie ein vergleichbarer Rahmen für alle Fälle geschaffen wurde, wurde auch bei diesem Fehlkonzept nur der Größenvergleich von zwei Brüchen umgesetzt und in der den teilnehmenden Personen vorliegenden Wissensvermittlung auf Ausführungen dieser Besonderheit verzichtet. Aufgrund der schwierigen Diagnostizierbarkeit dieses Fehlkonzepts im Allgemeinen und beim Größenvergleich von nur zwei Brüchen im Besonderen werden die beiden Vignetten, die Fälle mit der Nullstrategie beinhalten, von der weiteren Analyse und der Diskussion ausgeschlossen. Da sich das diagnostische Potenzial der Aufgaben durch die dargestellte Diagnosesituation in den einzelnen Vignetten über die Beziehung der Zahlmerkmale der Dezimalbrüche ergibt, hat der Ausschluss von gesamten Vignetten keine Auswirkungen auf die theoretischen Grundannahmen oder das diagnostische Potenzial der verwendeten Aufgaben in anderen Fällen. Eine leichte Verfälschung der Ergebnisse mag entstanden sein, da die Nullstrategie in dem zur Verfügung stehenden Material beschrieben ist, und so anzunehmen ist, dass die teilnehmenden Personen die zu diagnostizierende Fälle auch auf das Vorliegen der Nullstrategie überprüft haben und daher ggf. eine größere Anzahl an Aufgaben angeklickt und verarbeitet haben. Die durchschnittliche Urteilsakkuratheit bei der Diagnose der übrigen fünf Fälle beträgt 63,3 % ($SD = 0{,}48$).

5.1 Explorative Untersuchung nach Urteilskategorien

Die in Abschn. 2.4 vorgestellten Prozesse bei der Diagnose von Fehlkonzepten wurden auf der Grundlage der Modellierung der Diagnose als soziales Hypothesentesten postuliert. Im Rahmen dieser mehrdeutigen Urteilssituation wird angenommen, dass eine kognitive Verzerrung die Informationsverarbeitung während der Urteilsfindung beeinflusst. Anhand der theoretischen Überlegungen und der Darstellung in Abb. 3 können sieben Kategorien der Urteilsbildung (Urteilskategorien) differenziert und empirisch abgebildet werden. Im Folgenden werden die kognitiven Prozesse der

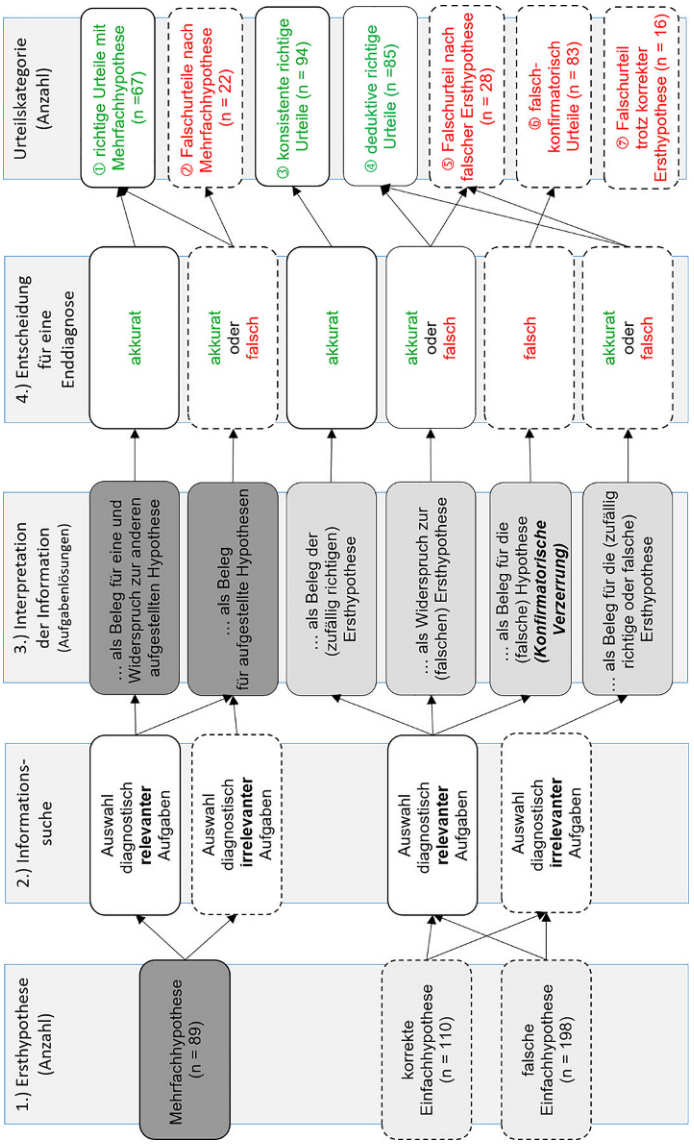

Abb. 7 Urteilskategorien bei der Diagnose von Fehlkonzepten im Bereich der Dezimalbrüche, wobei der nicht beobachtbare Prozess der Interpretation der Information deduziert wurde (Bei Prozess ② wurde in sieben Fällen eine falsche Mehrfachhypothese erstellt, d. h. das zu diagnostizierende Fehlkonzept wurde nicht genannt und der Urteilsprozess führte dann zu einem Falschurteil. Die *durchgezogenen Ränder* der *Küsten* stehen für die normative Informationsverarbeitung, die *gestrichelten Ränder* stehen für erwarbare, aber nicht optimale Prozesse. Darüber hinaus stehen die *dicken Ränder* der *Kästchen* für beobachtbares Verhalten, während *dünne Ränder* interne Prozesse darstellen)

Tab. 3 Deskriptive Übersicht über die Informationssuche und die Urteilssicherheit zur Differenzierung der sieben Urteilskategorien

	N	Durchschnittliche Anzahl an gestellten Aufgaben (SD)	Anteil an diagnostischen Aufgaben (SD)	Urteilssicherheit bei Ersthypothese (SD)	Urteilssicherheit bei Enddiagnose (SD)
① richtige Urteile mit Mehrfachhypothese	67	2,52 (2,01)	0,73 (0,28)	83,05 (25,99)	67,84 (31,53)
② Falschurteile nach Mehrfachhypothese	22	2,59 (2,30)	0,54 (0,38)	62,73 (41,53)	57,18 (36,19)
③ konsistente richtige Urteile	94	2,15 (1,75)	0,69 (0,35)	66,37 (30,84)	58,03 (21,40)
④ deduktive richtige Urteile	85	3,47 (2,59)	0,66 (0,28)	75,07 (22,40)	54,14 (17,31)
⑤ Falschurteile nach falscher Ersthypothese	28	3,82 (3,60)	0,65 (0,32)	55,34 (31,75)	57,71 (22,24)
⑥ falsch-konfirmatorische Urteile	83	2,48 (2,06)	0,47 (0,37)	62,67 (32,10)	55,71 (25,52)
⑦ Falschurteile trotz korrekter Ersthypothese	16	4,56 (4,24)	0,70 (0,24)	53,74 (34,05)	56,81 (23,79)

einzelnen Kategorien erklärt, die gefundenen Anzahlen dargestellt (Abb. 7) und die Informationsverarbeitung deskriptiv unterschieden (Tab. 3). Anhand dieser Urteilskategorien wird der Einfluss der Urteilsverzerrung auf die Prozesse der Informationsverarbeitung abgebildet.

Urteile, die nach dem Erstellen von Mehrfachhypothesen zustande kommen, werden somit in ① *richtige Urteile mit Mehrfachhypothese* bzw. in ② *Falschurteile nach Mehrfachhypothese* unterteilt.

Eine Einzelhypothese, die sich zufällig mit dem vorliegenden Lernendenfehlkonzept deckt, kann zu akkuraten Diagnosen führen und wird als ③ *konsistentes richtiges Urteil* bezeichnet. Der Urteilsprozess, welcher trotz korrekter Ersthypothese zu einer falschen Diagnose führt, wird im Folgenden als ⑦ *Falschurteil trotz korrekter Ersthypothese* kategorisiert.

Deduktive richtige Urteile (④) kommen trotz der Formulierung einer falschen Ersthypothese zustande. Falschurteile trotz oder ohne Verwerfen einer falschen Einzelhypothese werden als ⑤ *Falschurteile nach falscher Ersthypothese* bezeichnet. Das Beibehalten einer falschen Einzelhypothese wird als ⑥ *falsch-konfirmatorisches Urteil* kategorisiert. Diese Unterteilung erlaubt nun den Einfluss der Urteilsverzerrung, in diesem Fall von konfirmatorischer Informationsverarbeitung, zu untersuchen.

Tab. 3 spiegelt den explorativen Zugang zur Untersuchung der Urteilskategorien wider und zeigt einen deskriptiven Überblick über die durchschnittliche Anzahl an verarbeiteten Informationen (hier: gestellte Aufgaben), den Anteil an diagnostischen Aufgaben an der Informationssuche und die Urteilssicherheit bei der Enddiagnose.

Die Anzahl der weiteren Aufgaben wird erfasst als die durchschnittliche Anzahl an Aufgaben, die pro Vignette zusätzlich ausgewählt wurden. Der Anteil an diagnostischen Aufgaben ist der prozentuelle Anteil der diagnostisch relevanten Aufgaben in Bezug zu allen ausgewählten Aufgaben.

Bei konsistenten ③ und konfirmatorischen Urteilsprozessen ⑥ werden deskriptiv die wenigsten Aufgaben verarbeitet, während bei nicht-konfirmatorischen Urteilen ohne Formulierung von Mehrfachhypothesen (unbestimmt Falschurteile ⑤ und ⑦ und deduktiv richtige Urteile ④) die meisten Aufgaben während des Diagnoseprozesses ausgewählt werden.

Bezüglich der Verarbeitung von diagnostisch relevanten Aufgaben während des Diagnoseprozesses zeigt sich, dass bei den richtigen Urteilen unter Verwendung einer Mehrfachhypothese ① der größte Anteil an diagnostischen Aufgaben verwendet wurde, während bei den falsch-konfirmatorischen Urteilen ⑥ der kleinste Anteil an diagnostischen Aufgaben zugrunde liegt.

Die teilnehmenden Personen, die eine richtige Diagnose unter Verwendung einer Mehrfachhypothese getroffen haben, geben die höchste Urteilssicherheit sowohl bei der Formulierung der Ersthypothese als auch bei der Enddiagnose an, alle anderen Urteilskategorien (②–⑦) zeichnen sich durch eine geringere Sicherheit aus.

Basierend auf den Modellannahmen können aus den oben genannten Kategorien zwei für die übergreifende Forschungsfrage interessante Urteilsprozesse fokussiert werden: Einerseits der normative Urteilsprozess (d. h. richtige Urteile auf Basis von Mehrfachhypothesen ①) und andererseits der konfirmatorisch fehlgeleitete Urteilprozess (d. h. falsche Urteile durch Beibehalten einer falschen eindeutigen Ersthypothese ⑥). Hier zeigen sich nur geringe Unterschiede in der Anzahl der gestellten Aufgaben (2,52 vs. 2,48; $t(148) = 0{,}12$; $p = 0{,}91$), wohl aber im Anteil diagnostischer Aufgaben und der Urteilssicherheit: *Richtige Urteile mit Mehrfachhypothese* kommen unter Verwendung eines größeren Anteils von diagnostischen Aufgaben zustande (0,73 vs. 0,47; $t(148) = 4{,}76$; $p \leq 0{,}001$) und die urteilenden Personen sind sich ihrer Diagnose sicherer (67,84 vs. 55,71; $t(148) = 2{,}60$; $p = 0{,}001$) als bei falsch-konfirmatorischen Urteilen.

5.2 Prädiktoren für eine hohe Urteilsakkuratheit

Zur Beantwortung der ersten Forschungsfrage sollen die Prädiktoren für eine hohe Urteilsakkuratheit bei der Diagnose von Fehlkonzepten im Bereich der Dezimalbrüche ermittelt werden. Auf der Grundlage der theoretischen Modellierung des Diagnoseprozesses als Hypothesentesten wird die Formulierung der Ersthypothese, die Anzahl an gestellten Aufgaben und der Anteil an diagnostischen Aufgaben (d. h. Aufgaben mit relevanten Informationen) untersucht. Tab. 4 stellt jeweils die durchschnittlichen Werte der Prädiktoren für Urteilsakkuratheit nach akkuraten bzw. falschen Diagnosen dar.

Laut den Hypothesen 1 und 2 (s. Kap. 3) sollte die Wahrnehmung der Mehrdeutigkeit der Situation, (H1) – erhoben durch die Formulierung der Ersthypothese – die Anzahl und die Art gestellten Aufgaben (H2) bei der Diagnose von Fehlkonzepten im Bereich der Dezimalbrüche durch Lehramtsstudierende prädiktiv für eine hohe Urteilsakkuratheit sein. Zur Überprüfung dieser Annahmen werden verallgemeinerte

Tab. 4 Darstellung der durchschnittlichen Werte der angenommenen Prädiktoren für Urteilsakkuratheit nach akkuraten bzw. falschen Diagnosen

	Akkurate Diagnosen ($n = 250$)	Falsche Diagnosen ($n = 145$)
Durchschnittliche Anzahl an Mehrfachhypothesen (SD)	0,26 (0,44)	0,15 (0,36)
Durchschnittliche Anzahl an gestellten Aufgaben (SD)	2,73 (2,28)	2,94 (2,73)
Durchschnittlicher Anteil an diagnostischen Aufgaben (SD)	0,68 (0,31)	0,54 (0,36)

lineare gemischte Modelle mit der abhängigen Variablen Urteilsakkuratheit gerechnet, um die Effekte der Prädiktoren zu schätzen (Tab. 5). In dieser Auswertung werden jeweils fünf Falldiagnosen von jeder der teilnehmenden Personen untersucht, so dass der Einfluss der Itemschwierigkeit (Fälle) und der einzelnen Personen berücksichtigt wird. Die in Tab. 5 angegebenen R^2-Werte beziehen sich einmal auf die festen Effekte (*marginal* R^2), während der zweite Werte (*conditional* R^2) sowohl feste wie auch zufällige Effekte berücksichtigt.

Bei der sukzessiven Berechnung der Effekte einzelner Prädiktoren (Mehrfachhypothese, Anzahl an gestellten Aufgaben bzw. Anteil an diagnostischen Aufgaben) konnten außer den in Tab. 5 dargestellten keine weiteren signifikante Ergebnisse zur Vorhersage der Urteilsakkuratheit gefunden werden. Auf die einzelne Darstellung wird daher zugunsten des Modells mit der größten Varianzaufklärung verzichtet. Der Anteil der verwendeten diagnostischen Aufgaben beim Urteilsprozess hat einen signifikanten Effekt auf die Wahrscheinlichkeit, ein vorhandenes Fehlkonzept korrekt zu diagnostizieren. Darüber hinaus zeigt das Odds Ratio von 1,71, dass die Formulierung von Mehrfachhypothesen die Wahrscheinlichkeit für akkurate Diagnosen des vorliegenden Fehlkonzeptes deskriptiv, aber nicht signifikant, erhöht. Die Anzahl an gestellten Aufgaben zeigt keine Effekte. Beim Vergleich der beiden Modelle liefert das Prädiktorenmodell niedrigere AIC-Werte als das Nullmodell, das Nullmodell jedoch niedrigere BIC-Werte als das Prädiktorenmodell. Diese Werte liefern daher keine eindeutigen Argumente für eine Modellpräferenz. Auch hinsichtlich der Varianzaufklärung der Urteilsakkuratheit ist kein Modell eindeutig zu bevorzugen: Die Aufklärung über die festen Effekte der Prädiktoren (*marginal* R^2) ist in beiden Modellen sehr niedrig. Mit einer Varianzaufklärung von 4 % (im Vergleich zu 0 %) wäre das Prädiktorenmodell dennoch zu bevorzugen ($\chi^2(3) = 10{,}707$, $p = 0{,}013$). Bei der Varianzaufklärung über die festen und zufälligen Effekte, also unter Berücksichtigung der individuellen Unterschiede der teilnehmenden Personen (*conditional* R^2), zeigt das Nullmodell hingegen deskriptiv leicht höhere Werte. Der Unterschied in der Varianzaufklärung ist jedoch nicht signifikant ($\chi^2(3) = 0{,}183$ $p = 0{,}757$). Es konnte darüber hinaus kein signifikanter Zusammenhang zwischen den erstellten Mehrfachhypothesen und dem Anteil diagnostischer Aufgaben gefunden werden ($r = -0{,}082$, $N = 394$, $p = 0{,}104$).

Tab. 5 Verallgemeinerte lineare gemischte Modelle zur Darstellung der Effekte der Prädiktoren auf die Urteilsakkuratheit sowie die Varianz innerhalb (σ^2) und zwischen (τ_{00}) den Prädiktoren, der Intraclass-Correlation (ICC), der Varianzaufklärung (R^2) und des Akaike-Information-Criterion (AIC) bzw. des Bayesian-Information-Criterion (BIC) zur Modellselektion

	Urteilsakkuratheit Nullmodell			Urteilsakkuratheit Prädiktorenmodell		
Prädiktoren	Odds Ratios	CI	p	Odds Ratios	CI	p
(Intercept)	2,00	0,88–4,55	0,097	0,91	0,35–2,37	0,851
Mehrfachhypothese	–			1,71	0,87–3,38	0,120
Anzahl an gestellten Aufgaben				1,00	0,91–1,11	0,954
Anteil an diagnostischen Aufgaben				2,90	1,38–6,09	*0,005*
Random Effects						
σ^2	3,29	–		3,29	–	
$\tau 00$	Fälle: 0,76 Studierende: 0,67		–	Fälle: 0,62 Studierende: 0,55		–
ICC	0,30	–		0,26	–	
N	Fälle: 5 Studierende: 79		–	Fälle: 5 Studierende: 79		–
Observations	395	–		395	–	
Marginal R^2/Conditional R^2	0,000/0,303			0,041/0,293		
AIC	479,95			475,24		
BIC	491,89			499,12		

5.3 Personenmerkmal Präferenz für Intuition oder Deliberation

In einem letzten Auswertungsschritt wird untersucht, ob das über den PID-Fragebogen erhobene Personenmerkmal der Präferenz für Intuition oder Deliberation die Varianz beim Erkennen der Mehrdeutigkeit der Diagnosesituation, bei der (verzerrten) Informationsverarbeitung oder bei der Urteilsakkuratheit aufklären kann. Für das Personenmerkmal PID konnten nach der im Auswertungsleitfaden vorgeschlagenen Methode des Medians splits und der anschließenden Kategorisierung 15 Personen mit einer Präferenz für Intuition (Typ-I), 39 Personen mit einer Präferenz für Deliberation (Typ-D) und 25 Personen, die als situationsabhängige Urteilende (Typ-S) zu charakterisieren sind, gefunden werden (Betsch 2004).

Ein über zwei dummy-codierte Variablen berechnetes verallgemeinertes lineares gemischtes Modell mit den drei Gruppen der PID-Kategorisierung als Prädiktor zeigt weder für die Wahrnehmung der Mehrdeutigkeit der Situation ($CI = 0{,}35–22{,}32$, $p = 0{,}328$), noch für die Informationssuche (Anzahl an gestellten Aufgaben: $CI = –0{,}50–0{,}14$, $p = 0{,}275$, Anteil diagnostische Aufgaben $CI = –0{,}02–0{,}08$, $p = 0{,}250$), noch für die Urteilsakkuratheit ($CI = –0{,}04–0{,}10$, $p = 0{,}405$) signifikante Ergebnisse. Anhand einer einfaktoriellen Varianzanalyse (ANOVA) wurde getestet, ob sich die über die fünf Fälle gemittelten abhängigen Variablen der Wahrnehmung der Situation, der Informationssuche und der Urteilsakkuratheit zwischen den drei Gruppen unterscheiden. Dabei zeigt sich, dass Personen, die eine Präferenz für Deliberation

Tab. 6 Darstellung der durchschnittlichen Werte der Wahrnehmung der diagnostischen Situation, der Informationssuche und der Urteilsakkuratheit gemittelt über die 5 Fälle nach den Typen der Präferenz für Intuition und Deliberation (Betsch 2004)

	Typ-I-Personen ($n = 15$)	Typ-D-Personen ($n = 39$)	Typ-S-Personen ($n = 25$)
Durchschnittliche Anzahl formulierter Mehrfachhypothesen (SD)	0,16 (0,31)	0,31 (0,36)	0,11 (0,32)
Durchschnittliche Anzahl an gestellten Aufgaben (SD)	3,09 (1,29)	2,70 (1,03)	2,80 (2,51)
Durchschnittlicher Anteil an diagnostischen Aufgaben (SD)	0,62 (0,16)	0,65 (0,18)	0,59 (0,36)
Durchschnittliche Urteilsakkuratheit (SD)	0,63 77.152	0,67 77.164	0,58 77.176

zeigen, signifikant mehr Mehrfachhypothesen aufstellen als Personen, die den anderen Gruppen zugeordnet sind ($F(2, 392) = 10{,}37$, $p \leq 0{,}001$). Diese Gruppe verarbeitet deskriptiv durchschnittlich die geringste Anzahl an gestellten Aufgaben ($F(2, 392) = 0{,}69$, $p = 0{,}503$), wählt den höchsten Anteil an diagnostischen Aufgaben aus ($F(2, 392) = 1{,}68$, $p = 0{,}189$) und erzielt die höchste Urteilsakkuratheit ($F(2, 392) = 1{,}13$, $p = 0{,}325$), wobei die Unterschiede zwischen den Gruppen nicht statistisch signifikant sind (Tab. 6).

6 Diskussion

Eine akkurate Diagnose von Fehlkonzepten durch die Lehrkraft ist für den Aufbau belastbarer Vorstellungen bei Lernenden von entscheidender Bedeutung (Bradshaw und Templin 2014). Das Forschungsinteresse der hier vorliegenden Studie fokussiert die Entstehung solcher Urteile über die Untersuchung von Indikatoren der zugrundeliegenden kognitiven Prozesse bei der Situationswahrnehmung und der Informationsverarbeitung (Loibl et al. 2020; Rieu et al. 2022). Für eine solche mehrdeutige Urteilssituation wird angenommen, dass Verzerrungen stattfinden, welche sich negativ auf die Urteilsakkuratheit auswirken (Oswald und Grosjean 2004).

Die vorliegende Studie definiert diagnostisches Urteilen als Prozess der Informationsverarbeitung, wobei diese von den Dispositionen der urteilenden Person und den Informationen der diagnostischen Situation beeinflusst werden. Dafür wurde eine schultypische Diagnosesituation gewählt, in der mehrdeutige Inferenzen vom manifesten Verhalten von Lernenden (deren Lösung) auf die latenten Fehlkonzepte gemacht werden. Durch die Modellierung der Diagnose von Fehlkonzepten bei Dezimalbrüchen als soziales Hypothesentesten kann ein normativer Prozess und eine davon abweichende kognitive Verzerrung beschrieben werden. Die empirische Untersuchung der erstellten Ersthypothese, der verarbeiteten Informationen im Rahmen der Informationssuche und der Enddiagnose im experimentellen Rahmen erlaubt Rückschlüsse, ob die kognitiven Prozesse einer konfirmatorischen Verzerrung unterliegen.

6.1 Beschreibung der Urteilsprozesse

Das übergeordnete Ziel der Studie ist die Untersuchung des Einflusses von Urteilsverzerrungen auf die Urteilsakkuratheit bei der Diagnose von Fehlkonzepten. Dazu wurde die Verzerrung bei der Wahrnehmung der Situation anhand der Art der formulierten Ersthypothese und bei der Informationssuche, während der Urteilsprozesse anhand der Anzahl und der Art der verarbeiteten Informationen angenommen und unterschieden.

Die theoretischen Annahmen bezüglich der kognitiven Verzerrung bei der Wahrnehmung der komplexen Urteilssituation und der Informationsverarbeitung, während der anschließenden Urteilsprozesse konnte durch verallgemeinerte lineare gemischte Modelle und varianzanalysierende Verfahren empirisch untermauert werden.

Dabei zeigt sich, dass bei den konfirmatorischen Urteilsprozessen ⑥ die wenigsten Informationen verarbeitet werden. Angehende Lehrkräfte, die trotz richtiger Ersthypothese (unter Verkennung der Mehrdeutigkeit der abgebildeten Urteilssituation) zu einer falschen Diagnose kommen, verarbeiten die meisten Aufgaben während des Diagnoseprozesses (vgl. Kategorie *Falschurteile trotz korrekter Ersthypothese* ⑦). Dieses Ergebnis deutet darauf hin, dass bei diesem Urteilsprozess eine informationsintegrierende Strategie (Fiske 2012; Böhmer et al. 2015) angewendet wird, die aufgrund der Mehrdeutigkeit der Urteilssituation aber dennoch nicht immer zur richtigen Diagnose führt.

Als Beleg für die Modellannahmen des Auftretens der konfirmatorischen Verzerrung bei der Diagnose von Fehlkonzepten im Bereich der Dezimalbrüche, zeigen die erhobenen Daten, dass bei den *falsch-konfirmatorischen Urteilen* ⑥ der geringste Anteil an diagnostisch relevanten Aufgaben verwendet wurde. Die Modellannahme, dass die verzerrte Informationssuche konfirmatorisch im Sinne der erstellen Einzelhypothese durchgeführt wird, dass die vorliegenden Informationen als Belege für diese Ersthypothese interpretiert werden und dass mehrdeutige Hinweise ignoriert werden, lässt sich anhand der Datenauswertung unterstreichen (Dreger 2012; Schulz-Hardt und Köhnken 2000). So ist zu erklären, dass trotz der Verarbeitung von 53 % relevanter Informationen die Ersthypothese nicht revidiert und das Fehlkonzept nicht akkurat diagnostiziert wird.

Das Ergebnis, dass die teilnehmenden Personen, die eine richtige Diagnose unter Verwendung einer Mehrfachhypothese getroffen haben, die höchste Urteilssicherheit sowohl zu Beginn als auch am Ende des Urteilsprozesses angeben (Hausmann und Läge 2008), könnte einen Hinweis darauf geben, dass diese Personen die Mehrdeutigkeit der Urteilssituation erkannt, daher die vorhandenen Informationen bewusst und strategisch verarbeitet und somit größeres Vertrauen in ihre Diagnose haben.

Die Kategorisierung und die Gegenüberstellung des normativen Prozesses ①, der zu akkuraten Diagnosen nach Formulierung einer Mehrfachhypothese führt, mit dem konfirmatorischen Prozess, der zu falschen Diagnosen führt ⑥, unterstreicht die Unterschiede der beiden Vorgehensweisen bei der Wahrnehmung der Situation, bei der Art der verarbeiteten Informationen und der Urteilssicherheit und belegt auch hier die Annahmen des kognitiven Modells.

6.2 Prädiktoren für die Urteilsakkuratheit

Zur Überprüfung der aufgestellten Hypothesen wurden die Effekte von Prädiktoren auf die Urteilsakkuratheit bei Einschätzung von Fehlkonzepten im Bereich der Dezimalbrüche untersucht. Bisherige Studien konnten bereits die wissensbasierte Informationsverarbeitung in diagnostischen Situationen zeigen (Ostermann et al. 2018; Rieu et al. 2022). Das fachdidaktische Wissen über die Fehlkonzepte, welches anhand der Übersicht über Fehlkonzepte vorlag, wurde in dieser Studie daher als notwendige Bedingung konzipiert und lag allen Teilnehmenden gleichermaßen vor. Auch wenn ein ideales Vorgehen die Wahrnehmung der mehrdeutigen Urteilssituation und somit die Formulierung von Mehrfachhypothesen voraussetzt, so konnte statistisch kein signifikanter Einfluss auf die Urteilsakkuratheit festgestellt werden. Dieses Ergebnis kann darin begründet liegen, dass die Prozesse der Informationssuche und Interpretation den angenommenen Effekt der Formulierung von Mehrfachhypothesen bei der Ersthypothese auf die Akkuratheit der Enddiagnose mediieren und aufgrund der Komplexität dieses Urteilsprozesses die Art der Erstdiagnose nicht deterministisch auf die weiteren Prozessschritte wirkt. Eine weitere Interpretation könnte sein, dass einige teilnehmende Personen die Mehrdeutigkeit der Urteilssituation zwar wahrnehmen, aber möglicherweise dennoch nur eine Hypothese formulieren und im Anschluss den Urteilsprozess entsprechend ihrer Wahrnehmung hypothesenprüfend gestalten. Solch ein Phänomen könnte möglicherweise in einer weiteren Studie durch Triangulierung mit qualitativen Daten aufgedeckt werden.

Als Prädiktor für akkurate Diagnosen konnte hingegen ein hoher Anteil von diagnostischen Aufgaben gefunden werden. Eine gezielte Informationssuche und die bewusste Integration der relevanten Informationen führt wahrscheinlicher zu einer akkuraten Enddiagnose. Dieses Ergebnis entspricht den formulierten Hypothesen und unterstreicht die Bedeutung von informationsverarbeitenden Strategien bei Urteilsprozessen.

Bei keiner der erhobenen soziodemographischen Variablen (Alter, Geschlecht und Praxiserfahrung) konnte ein Einfluss auf das Urteilsverhalten nachgewiesen werden. Die Präferenz für intuitive oder deliberate Entscheidungen als Personenmerkmal scheint sich bei der Diagnose von Fehlkonzepten mit kleinen Effekten auf die Wahrnehmung der Mehrdeutigkeit der Situation und die Informationssuche auszuwirken. Das Ergebnis unterstreicht im Ansatz die personenabhängige Diagnosesensitivität als Disposition (Kron et al. 2021). Die Varianz bei der Informationssuche und bei der Urteilsakkuratheit kann durch dieses Personenmerkmal allerdings nicht aufgeklärt werden.

6.3 Limitationen

Die vorliegende Studie stellt einen ersten Schritt zum besseren Verständnis von Urteilsverzerrungen in mehrdeutigen diagnostischen Situationen dar. Weitere Untersuchungen sollten die im Folgenden genannten Grenzen des beschriebenen Vorgehens und der Ergebnisse fokussieren.

Auch wenn die dargestellte diagnostische Situation sehr realitätsbezogen ist und Lehrkräften im Alltag begegnet, so kann dennoch die Frage gestellt werden, inwie-

fern auch Studierenden ohne Praxiserfahrung diese Authentizität wahrnehmen und deren wissensbasierte Urteilsprozesse vergleichbar mit erfahrenen Lehrkräften sind. Dieser Frage wird im Rahmen von Folgestudien nachgegangen. Für Lehrkräfte in der Unterrichtssituation ist die Überprüfung, inwiefern der auftretende Fehler tatsächlich auch systematisch auftritt, der erste Schritt bei der Feststellung von Fehlkonzepten. Im Rahmen der Studie wurde den teilnehmenden Personen mitgeteilt, dass es sich bei den Fällen um ein konsistent vorkommendes Fehlkonzept der Lernenden handelt. Daher kann die Einteilung in relevante und irrelevante Aufgaben anhand ihres diagnostischen Potenzials in diesem Rahmen getroffen werden. Im schulischen Kontext hingegen könnte die Lösung einer Aufgabe desselben Typs wie die vorliegende fehlerhaft gelöste Aufgabe eine sinnvolle, diagnostische Information zur Absicherung eines vorliegenden, systematischen Fehlers sein. Auch die Einschränkung auf vier häufig vorkommende Fehlkonzepte limitiert die Komplexität der Urteilssituation; diese Einschränkung beruht auf der Zielsetzung der Studie und ermöglicht die kontrollierte, systematische Untersuchung der angenommenen Prozesse.

Die Einteilung der Aufgaben nach ihrem diagnostischen Potenzial ermöglicht situationsunabhängige Aussagen über die Relevanz der Informationen für den diagnostischen Prozess. Unabhängig von der Ersthypothese sind somit Aufgaben, die trotz vorliegendem Fehlkonzept richtig gelöst werden und Aufgaben, die keine weiteren Informationen als die bereits dargestellte Aufgabe liefern, irrelevant für die Diagnose des Fehlkonzeptes. Alle anderen Aufgaben, die weitere Informationen zur Abtrennung von zwei möglichen Fehlkonzepten liefern, werden als relevant für den Diagnoseprozess interpretiert. Diese Einteilung könnte noch verfeinert werden: Die irrelevanten Aufgaben könnten getrennt werden in Aufgaben, die von allen Lernenden trotz Fehlkonzept richtig gelöst werden und solchen, die keine weiteren Diagnoseinformationen enthalten, da sie strukturgleich zur bereits dargestellten Aufgabe sind. Bei den relevanten Aufgaben könnte eine Unterscheidung nach der Trennung zwischen den verschiedenen Fehlkonzepten vorgenommen werden. Für das hier verwendete Design scheint die vorgenommene, dichotome Einteilung ausreichend präzise.

Zur Reduzierung möglicher Einflüsse von unterschiedlichem fachdidaktischem Vorwissen wurde eine Kohorte angehender Lehrkräfte zu Beginn ihres Grundschullehramtsstudiums gewählt. Diese Entscheidung wurde bewusst getroffen, da Dezimalbrüche zwar bereits in der Grundschule behandelt werden (z. B. beim Thema Größen und Messen), allerdings kein Schwerpunktthema für die Ausbildung von Primarstufenlehrkräften ist. So wurde versucht sicherzustellen, dass der diagnostischen Situation einerseits eine hohe Praxisrelevanz zugeschrieben wird, andererseits die Praktikums- und Praxiserfahrung bei den Studierenden sehr gering ausgeprägt ist und im Rahmen des Studiums noch keine fachdidaktischen Veranstaltungen besucht wurden. Allen teilnehmenden Personen wurde das fachdidaktische Wissen über die typischen und operationalisierten Fehlkonzepte im Bereich der Dezimalbrüche während der gesamten Studie schriftlich zur Verfügung gestellt (Übersicht über Fehlkonzepte). Bei der untersuchten Kohorte kann davon ausgegangen werden, dass nur die vorgegebenen, häufig auftretenden Fehlkonzepte in den Blick genommen werden, da keine Kenntnisse über weitere Fehlkonzepte zu erwarten sind. Dieses Vorgehen wurde gewählt, um die notwendige Voraussetzung für die

ablaufenden Urteilsprozesse der angehenden Lehrkräfte sicherzustellen (Rieu et al. 2022). In weiteren Studien sollte zusätzlich untersucht werden, wie sich die Bereitstellung des fachdidaktischen Wissens auf die Formulierung von Ersthypothesen und die Enddiagnosen auswirkt und wie berufserfahrene Lehrkräfte (auch ohne diese Hilfestellung) diagnostizieren. Aufgrund der kognitiven Komplexität der Urteilssituation und der zeitintensiven Auseinandersetzung mit den Fallvignetten wurde auf eine zusätzliche Erhebung von weiterer Disposition (v. a. fachdidaktisches Wissen) der angehenden Lehrkräfte verzichtet. Um mögliche persönliche Unterschiede auszugleichen, wurde eine möglichst homogene Studierendenkohorte zu Beginn des Studiums ausgewählt. Hier könnten weitere Studien ansetzen und den Einfluss von fachdidaktischem Wissen nicht nur auf die Urteilsakkuratheit, sondern auch auf die Urteilsprozesse untersuchen.

Bei der Erhebung der Informationssuche wurden sowohl die Anzahl aller gestellten Aufgaben als auch der Anteil der diagnostischen Aufgaben berücksichtigt um ein möglichst komplettes Abbild der Sensitivität – also der Fähigkeit der beurteilenden Personen, relevante Informationen auszusuchen – zu erhalten (s. Kron et al. 2021). Gerade bei der Betrachtung des Anteils relevanter diagnostischer Aufgaben bei den konfirmatorisch-falschen Urteilsprozessen (ca. 53 % aller Aufgaben) liegt die Vermutung nah, dass diese Personen alle Informationen ohne Evidenzregel konfirmatorisch als Beleg für ihre Einzelhypothese verwendet haben. Diese Annahme müsste allerdings durch eine Triangulation mit weiteren Forschungsmethoden, wie beispielsweise einem *think-aloud* Ansatz, überprüft werden.

6.4 Ausblick

Die Ergebnisse der vorliegenden Studie erlauben erste Erkenntnisse zu Urteilsprozessen von angehenden Lehrkräften in der komplexen Diagnose von Fehlkonzepten im Bereich der Dezimalbrüche. Die vorgenommene Kategorisierung anhand der Wahrnehmung der Situation (Art der Ersthypothese) und der weiteren Informationsverarbeitung lässt eine erste Unterscheidung zwischen normativ-akkuraten und konfirmatorisch-verzerrten Urteilsprozessen zu. Der Befund, dass eine konfirmatorisch-verzerrte Informationssuche nachgewiesen werden konnte, klärt die Varianz der Urteilsakkuratheit der angehenden Lehrkräfte in dieser Urteilssituation teilweise auf und liefert Anhaltspunkte auch für andere Domänen (Südkamp et al. 2012). Das beschriebene normative Vorgehen, welches ausgehend von alternativen Hypothesen diagnostisch relevante Aufgaben aussucht, um abschließend zu einer akkuraten Diagnose zu gelangen, sollte als Urteilsstrategie in mehrdeutigen Situationen in die Lehrkräfteausbildung einfließen, um eine höhere Diagnosegenauigkeit im Bereich von Fehlkonzepten zu erzielen und damit die Adaptivität von Unterricht zu steigern. Ein vergleichbarer explorativer Ansatz mit berufserfahrenen Lehrkräften könnte darüber hinaus wichtige Informationen liefern, inwiefern sich die Informationsverarbeitung und die Urteilsbildung dieser Personengruppe von den hier berichteten Ergebnissen unterscheidet.

Darüber hinaus bieten die Erkenntnisse den Ausgangspunkt für weitere Untersuchungen. So wurden nur zwei der sieben unterschiedenen Urteilskategorien näher beleuchtet, ein weiteres Augenmerk sollte beispielsweise auf die Kategorien der *de-*

duktiv richtigen Urteile bzw. der *Falschurteile trotz korrekter Ersthypothese* gerichtet werden. Es konnte gezeigt werden, dass bei diesen Prozessen viele Informationen verarbeitet werden, die zu unterschiedlichen Urteilen führen. In weiteren Untersuchungen sollten zusätzlich noch Aussagen über die Reihenfolge der Verarbeitung gemacht werden (z. B. anhand von Logfiles), um den stattfindenden Urteilsprozess noch präziser nachverfolgen zu können. Da außerdem gezeigt wurde, dass Personenmerkmale nur eine bedingte Rolle bei der Vermeidung von Urteilsverzerrungen spielen, können Effekte von Interventionen auf die Reduzierung der Verzerrung bei der Wahrnehmung der mehrdeutigen Situation und bei der Informationssuche erwartet werden.

Abschließend soll der forschungsmethodische Impuls dieser Studie unterstrichen werden, da über die Erstellung von Modellannahmen und der Umsetzung von geeigneten Erhebungsmöglichkeiten zur Untersuchung der Informationsverarbeitung ein empirischer Beleg für die Genese von diagnostischen Urteilen im Speziellen und von menschlichem Denken im Allgemeinen aufgezeigt wurde (Loibl et al. 2020; Leuders et al. 2022). Konkret wurden die Erkenntnisse zum *confirmation bias* auf die Diagnoseprozesse von angehenden Lehrkräften übertragen. Die Ergebnisse stellen einen möglichen Erklärungsansatz für die mäßige Akkuratheit und hohe Varianz von Lehrkräftediagnosen dar.

7 Anhang

7.1 Diagnosevignetten

Für die möglichen Fehlkonzepte wurden folgende Abkürzungen verwendet: Komma-trennt (KT), Kein-Komma (KK), Länger-ist-kleiner (LIK) und Nullstrategie (NLS) (Abb. 8).

In dieser Vignette (Abb. 9) liegt das Fehlkonzept „Komma-trennt“ vor. Dieses Fehlkonzept kann aufgrund der fehlerhaft gelösten Aufgabe 4,08 > 4,7 zunächst nicht eindeutig zugeordnet werden, da bei dieser Aufgabe auch das Fehlkonzept „Kein-Komma“ zu einer fehlerhaften Lösung führt. Die diagnostizierenden Lehrkräfte müssen weitere Aufgaben suchen, die bei den beiden Fehlkonzepten unterschiedlich gelöst würden. Aufgaben mit diesem diagnostischen Potenzial liegen in Runde 1 an der 3. und 4. Stelle von links, in Runde 2 an der 1. und 4. Stelle von links, in Runde 3 an der 2. und 4. Stelle und in Runde 4 an der 2. und 3. Stelle voSn links vor. Die teilnehmenden Personen mussten das vorliegende Fehlkonzept spätestens nach Runde 4 diagnostizieren.

Diese Logik wurde in allen weiteren Vignetten vergleichbar umgesetzt.

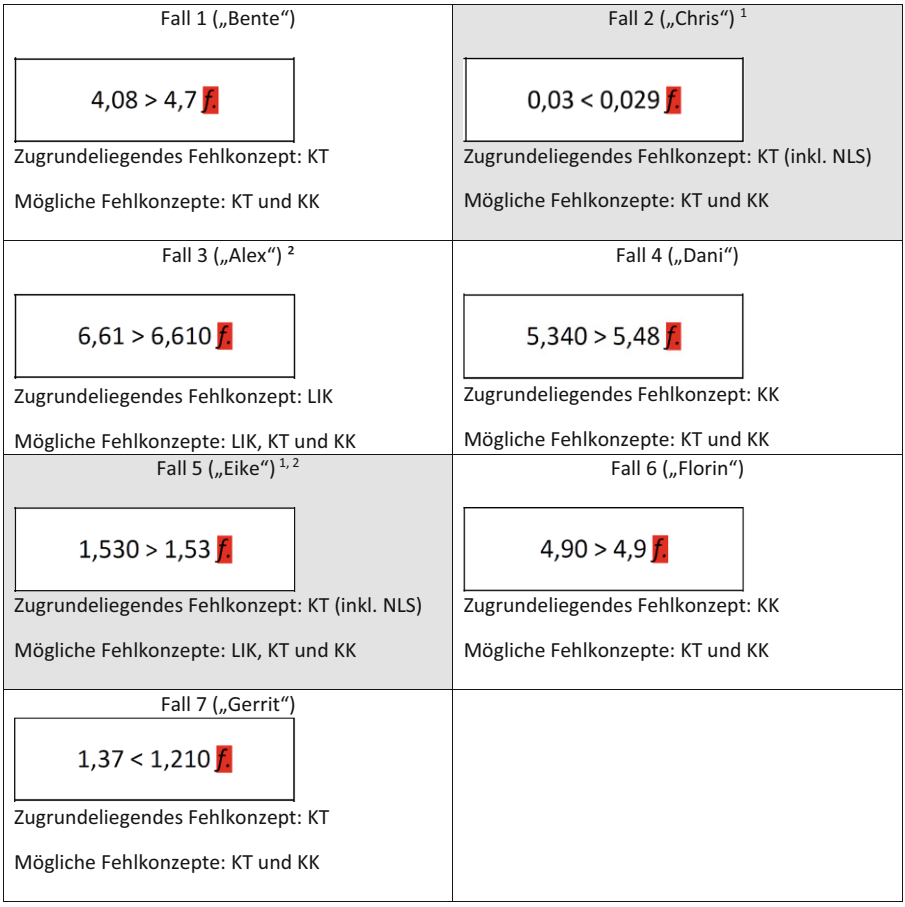

Abb. 8 Mögliche Fehlkonzepte. ([1] Die *Fälle 2* und *5* wurden aufgrund von Bodeneffekten bei der Urteilsakkuratheit von der Berechnung ausgeschlossen und sind hier nur der Vollständigkeit abgebildet; [2] Die Mehrdeutigkeit der *Fälle 3* und *5* beruht auf der Tatsache, dass Lernende mit KK- und KT-Strategie bzw. mit LIK-Strategie diese Aufgabe ebenfalls fehlerhaft lösen, das Ungleichheitszeichen allerdings vertauschen)

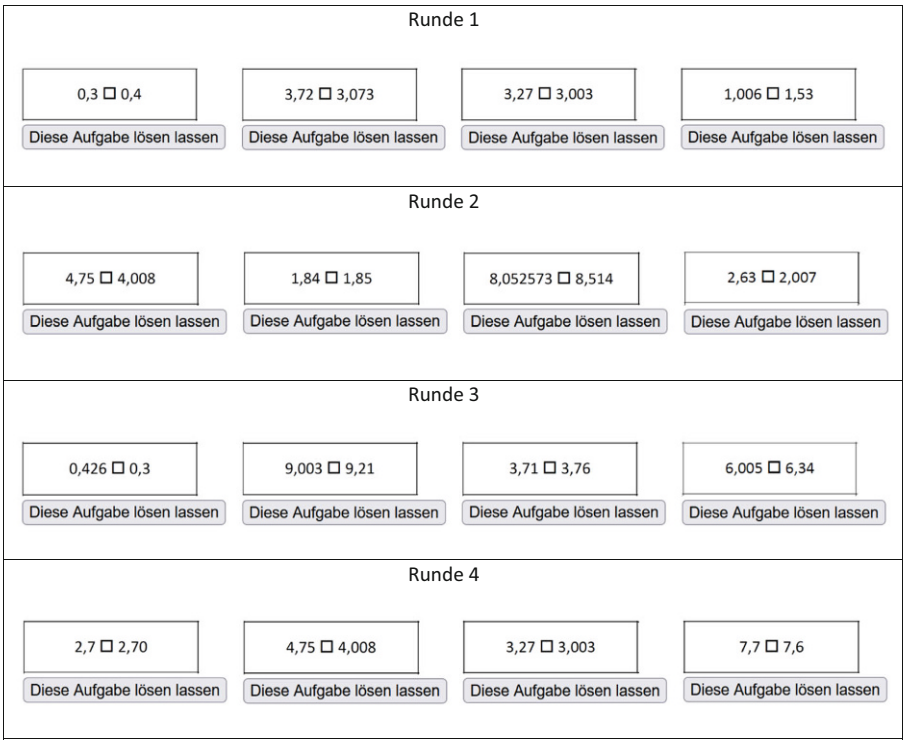

Abb. 9 Gezeigte Aufgaben in 4er-Blöcken für Fall 1

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

Ay, Y. (2017). A review of research on the misconceptions in mathematics education. *Education Research Highlights in Mathematics, Science and Technology*, *2017*, 21–31.

Bashir, T., Rasheed, S., Raftar, S., Fatima, S., & Maqsood, S. (2013). Impact of behavioral biases on investor decision making: Male vs female. *Journal of Business and Management*, *10*(3), 60–68.

Baur, A. (2018). Fehler, Fehlkonzepte und spezifische Vorgehensweisen von Schülerinnen und Schülern beim Experimentieren. *Zeitschrift für Didaktik der Naturwissenschaften*, *24*(1), 115–129.
Beck, E., Baer, M., Guldimann, T., Bischoff, S., Brühwiler, C., Müller, P., Niedermann, R., et al. (Hrsg.). (2008). *Adaptive Lehrkompetenz: Analyse und Struktur, Veränderbarkeit und Wirkung handlungssteuernden Lehrerwissens*. Pädagogische Psychologie und Entwicklungspsychologie, Bd. 63. Münster: Waxmann.
Becker, S., Spinath, B., Ditzen, B., & Dörfler, T. (2020). Der Einfluss von Stress auf Prozesse beim diagnostischen Urteilen – eine Eye Tracking-Studie mit mathematischen Textaufgaben. *Unterrichtswissenschaft*, *48*(4), 531–550.
Betsch, C. (2004). Präferenz für Intuition und Deliberation (PID). *Zeitschrift Für Differentielle Und Diagnostische Psychologie*, *25*(4), 179–197.
Binder, K., Krauss, S., Hilbert, S., Brunner, M., Anders, Y., & Kunter, M. (2018). Diagnostic skills of mathematics teachers in the COACTIV study. In *Diagnostic competence of mathematics teachers*. Springer.
Böhmer, I., Hörstermann, T., Gräsel, C., Krolak-Schwerdt, S., & Glock, S. (2015). Eine Analyse der Informationssuche bei der Erstellung der Übergangsempfehlung: Welcher Urteilsregel folgen Lehrkräfte? *Journal for educational research online*, *7*(2), 59–81.
Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*(3), 403–425.
Brauer, M., & Curtin, J.J. (2017). Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411.
Brown, V.A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1).
Brunner, K., Obersteiner, A., & Leuders, T. (2021). How prospective teachers detect potential difficulties in mathematical tasks—an eye tracking study. *RISTAL: Research in Subject Matter Teaching and Learning*, *4*, 108–125.
Confrey, J., & Kazak, S. (2006). A thirty-year reflection on constructivism in mathematics education in PME. In *Handbook of research on the psychology of mathematics education*. Brill.
Corno, L.Y.N. (2008). On teaching adaptively. *Educational psychologist*, *43*(3), 161–173.
Dreger, B. (2012). *Diagnose: Confirmation bias. Wie die anfängliche Überzeugtheit von einer klinisch-psychologischen Prüfhypothese, die Berufserfahrung und das Fachwissen die Validität klinischer Diagnosen beeinflussen*. Doctoral dissertation
Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). https://doi.org/10.1007/978-3-642-41089-5.
Dünnebier, K., Gräsel, C., & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung: Eine experimentelle Studie zu Ankereffekten. *Zeitschrift für Pädagogische Psychologie*, *23*(34), 187–195.
Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of personality and social psychology*, *71*(2), 390.
Fiedler, K. (1997). Die Verarbeitung sozialer Informationen für Urteilsbildung und Entscheidungen. In *Sozialpsychologie*. Springer.
Fiske, S.T. (2012). The continuum model and the stereotype content model. *Handbook of theories of social psychology*, *1*, 267–288.
Fujii, T. (2020). Misconceptions and alternative conceptions in mathematics education. In S. Lerman (Hrsg.), *Encyclopedia of mathematics education* (S. 453–455). Springer.
Gatlin, K.P., Cooley, L.G., & Elam, A.G. (2019). Confirmation bias: Does it vary by culture or education level. *International Journal of Business Marketing and Management*, *4*(2), 40–43.
Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of economic literature*, *55*(1), 96–135.
Hausmann, D., & Läge, D. (2008). Sequential evidence accumulation in decision making: The individual desired level of confidence can explain the extent of information acquisition. *Judgment and Decision Making*, *3*(3), 229–243. https://doi.org/10.1017/S1930297500002436.
Heitzman, 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.

Herppich, S., Praetorius, A. K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., & Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher education*, *76*, 181–193.

Hinson, J. M., Jameson, T. L., & Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 298–306.

Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der Pädagogischen Diagnostik* (6. Aufl.). Beltz.

Klieme, E. (2020). Guter Unterricht – auch und besonders unter Einschränkungen der Pandemie? In *„Langsam vermisse ich die Schule …"* (S. 117–135). Waxmann.

Krauss, S., Neubrand, M., Blum, W., et al. (2008). Die Untersuchung des professionellen Wissens deutscher Mathematik-Lehrerinnen und -Lehrer im Rahmen der COACTIV-Studie. *Journal für Mathematik-Didaktik*, *29*, 233–258.

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*.

Kuo, B. C., Chen, C. H., & de la Torre, J. (2018). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement*, *42*(3), 179–191.

Leuders, T., & Loibl, K. (2021). Beyond subject specificity—Student and teacher thinking as sources of specificity in teacher diagnostic judgments. *RISTAL*, *4*, 60–70.

Leuders, T., Dörfler, T., Leuders, J., & Philipp, K. (2018). Diagnostic competence of mathematics teachers: unpacking a complex construct. In *Diagnostic competence of mathematics teachers* (S. 3–31). Springer.

Leuders, T., Loibl, K., Sommerhoff, D., et al. (2022). Toward an overarching framework for systematizing research perspectives on diagnostic thinking and practice. *Journal für Mathematik-Didaktik*, *43*, 13–38.

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic Judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education*, *91*(3).

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern. *Zeitschrift für Pädagogische Psychologie*, *23*(3).

McKown, C., & Weinstein, R. S. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development*, *74*(2), 498–515.

Morris, A. K., Hiebert, J., & Spitzer, S. M. (2009). Mathematical knowledge for teaching in planning and evaluating instruction: What can preservice teachers learn? *Journal for research in mathematics education*, *40*(5), 491–529.

Mosandl, C., & Sprenger, L. (2014). Von den natürlichen Zahlen zu den Dezimalzahlen – nicht immer ein einfacher Weg! *Praxis der Mathematik in der Schule*, *56*, 16–21.

Moser Opitz, E., & Nührenbörger, M. (2015). Diagnostik und Leistungsbeurteilung. In *Handbuch der Mathematikdidaktik* (S. 491–512). Springer Spektrum.

Nesher, P. (1987). Towards an instructional theory: The role of student's misconceptions. *For the learning of mathematics*, *7*(3), 33–40.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175–220.

Nitsch, R. (2015). *Diagnose von Lernschwierigkeiten im Bereich funktionaler Zusammenhänge*. Wiesbaden: Springer.

Ostermann, A., Leuders, T., & Nückles, M. (2018). Improving the judgment of task difficulties: Prospective teachers' diagnostic competence in the area of functions and graphs. *Journal of Mathematics Teacher Education*, *21*(6), 579–605.

Oswald, M. E., & Grosjean, S. (2004). Confirmation Bias. In R. Pohl (Hrsg.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (1. Aufl. S. 79–96). Psychology Press.

Oudman, S., Van de Pol, J., Bakker, A., Moerbeek, M., & Van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education*, *76*, 214–226.

Padberg, F., & Wartha, S. (2017). *Didaktik der Bruchrechnung*. Springer.

Philipp, K. (2018). Diagnostic Competences of mathematics teachers with a view to processes and knowledge resources. In T. Leuders, K. Philipp & J. Leuders (Hrsg.), *Diagnostic competence of mathematics teachers* (S. 109–127). Springer.

van de Pol, J., van Gog, T., & Thiede, K. (2021). The relationship between teachers' cue-utilization and their monitoring accuracy of students' text comprehension. *Teaching and Teacher Education*, *107*.
Prediger, S. (2008). The relevance of didactical categories for analysing obstacles in conceptual change—Revisiting the case of multiplication of fractions. *Learning and Instruction*, *18*(1), 3–17.
Prediger, S., & Wittmann, G. (2009). Aus Fehlern lernen – (wie) ist das möglich. *Praxis der Mathematik in der Schule*, *51*(3), 1–8.
Prediger, S., & Zindel, C. (2017). Deepening prospective mathematics teachers' diagnostic judgments: interplay of videos, focus questions and didactic categories. *European Journal of Science and Mathematics Education*, *5*(3), 222–242.
Rieu, A., Loibl, K., Leuders, T., & Herppich, S. (2020). Diagnostisches Urteilen als informationsverarbeitender Prozess – Wie nutzen Lehrkräfte ihr Wissen bei der Identifizierung und Gewichtung von Anforderungen in Aufgaben? *Unterrichtswissenschaft*, *48*(4), 503–529.
Rieu, A., Leuders, T., & Loibl, K. (2022). Teachers' diagnostic judgments on tasks as information processing—The role of pedagogical content knowledge for task diagnosis. *Teaching and Teacher Education*, *111*.
Rubie-Davies, C., Hattie, J., & Hamilton, R. (2006). Expecting the best for students: Teacher expectations and academic outcomes. *British Journal of Educational Psychology*, *76*(Pt 3), 429–444.
Schons, C., Obersteiner, A., Reinhold, F., Fischer, F., & Reiss, K. (2022). Developing a simulation to foster prospective mathematics teachers' diagnostic competencies: the effects of scaffolding. *Journal für Mathematik-Didaktik*, , 1–24.
Schreiter, S., Vogel, M., Rehm, M., & Dörfler, T. (2021). Teachers' diagnostic judgment regarding the difficulty of fraction tasks: A reconstruction of perceived and processed task characteristics. *RISTAL*, *4*, 126–145.
Schulz-Hardt, S., & Köhnken, G. (2000). Wie ein Verdacht sich selbst bestätigen kann: Konfirmatorisches Hypothesentesten als Ursache von Falschbeschuldigungen wegen sexuellen Kindesmißbrauchs. *Praxis der Rechtspsychologie*, *10*(Sonderheft 1), 60–88.
Smith, J.P., diSessa, A.A., & Roschelle, J. (1994). Misconceptions reconceived: a constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, *3*(2), 115–163.
Stacey, K. (2005). Travelling the road to expertise: A longitudinal study of learning. In H.L. Chick & J.L. Vincent (Hrsg.), *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education* (S. 19–36). PME.
Steinle, V., & Stacey, K. (1998). The incidence of misconceptions of decimal notation amongst students in Grades 5 to 10. In C. Kanes, M. Goos & E. Warren (Hrsg.), *Teaching mathematics in new times* (S. 548–555).
Stewart, J.E. (2008). Locus of control and self-attribution as mediators of hazardous attitudes among aviators: A review and suggested applications. *International Journal of Applied Aviation Studies*, *8*(2), 263–279.
Südkamp, A., & Praetorius, A.-K. (2017). *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen*. Waxmann.
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.
Tobisch, A., & Dresel, M. (2017). Negatively or positively biased? Dependencies of teachers' judgments and expectations based on students' ethnic and social backgrounds. *Social Psychology of Education*, *20*, 731–752.
Tomlinson, C.A., Brighton, C., Hertberg, H., Callahan, C.M., Moon, T.R., Brimijoin, K., et al. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: a review of literature. *Journal for the Education of the Gifted*, *27*(2-3), 119–145.
Trope, Y., & Liberman, A. (1996). Social hypothesis testing: cognitive and motivational mechanisms. In E.T. Higgins & A.W. Kruglanski (Hrsg.), *Social psychology: handbook of basic principles* (S. 239–270). Guilford.
Tschan, F., Semmer, N.K., Gurtner, A., Bizzari, L., Spychiger, M., Breuer, M., & Marsch, S.U. (2009). Explicit reasoning, confirmation bias, and illusory transactive memory: A simulation study of group medical decision making. *Small Group Research*, *40*(3), 271–300.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.
Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*.

Van Ophuysen, S. (2006). Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahnempfehlung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *38*(4), 154–161.

Westhoff, K., & Kluck, M.-L. (2014). *Psychologische Gutachten schreiben und beurteilen*. Springer.

Witteman, C., Van den Bercken, J., Claes, L., & Godoy, A. (2009). Assessing rational and intuitive thinking styles. *European Journal of Psychological Assessment*, *25*(1), 39.