# Beurteilung von Unterrichtsqualität in der Chemielehrkräftebildung – Die Vernetzung von Unterrichtsforschung und -praxis

Von der Naturwissenschaftlichen Fakultät der
Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation
von
Benjamin Heinitz, M. Ed.

2024

Referent: Prof. Dr. rer. nat. Andreas Nehring
Korreferentin: Apl. Prof. Dr. rer. nat. Friederike Korneck
Tag der Promotion: 04.06.2024

# I. Zusammenfassung

Die Erfassung von Unterrichtsqualität ist ein zentrales Anliegen in der Unterrichtsforschung und ebenso bedeutend für die Praxis der Lehrkräftebildung in der zweiten Phase. Dennoch ist Unterrichtsqualität nicht eindeutig definiert, sondern vielmehr ein Konglomerat unterschiedlicher Ansätze. Diese weisen teilweise Überschneidungen, aber auch Widersprüche auf. Die drei Basisdimensionen der Unterrichtsqualität sind eine häufig verwendete Grundlage in der Unterrichtsforschung des deutschsprachigen Raums. Jedoch wird auch bei diesen auf fehlende Aspekte oder abweichende Konzeptualisierungen hingewiesen. Vor allem Fachspezifika scheinen dabei eine besondere Rolle einzunehmen, da sie nur schwierig aus einer generischen Perspektive erfasst werden können. Neuere Entwicklungen in der Unterrichtsqualitätsforschung adressieren diesen Ergänzungsbedarf und bieten durch das *Syntheseframework* einen breiteren Ansatz zur Erfassung der Unterrichtsqualität. Die Passung für das Fach Chemie musste jedoch zunächst untersucht werden.

Als erster Beitrag der Dissertation wurde ein Review zu Videostudien der Naturwissenschaftsdidaktiken durchgeführt. Damit sollte herausgestellt werden, welche Qualitätsmerkmale aus Sicht der Naturwissenschaften relevant sind und inwiefern sie sich im *Syntheseframework* verorten lassen. Daraus wurde das Framework *naturwissenschaftsdidaktischer Perspektivierung* abgeleitet. Um die Nutzbarkeit des adaptierten Frameworks zu überprüfen, wurde im zweiten Beitrag ein Abgleich mit Ansätzen der Unterrichtsqualität aus den Fächern Biologie und Physik durchgeführt. Dabei konnte unter anderem die breite Anwendbarkeit des Frameworks herausgestellt werden, wodurch sich im Folgenden eine Möglichkeit bot, unterschiedliche Ansichten von naturwissenschaftlicher Unterrichtsqualität in einem gemeinsamen Rahmen abzubilden.

Es gibt nur wenige zentrale Vorgaben für die zweite Phase der Lehrkräftebildung, sodass eine einheitliche Entwicklung der Unterrichtsqualität bei Referendar*innen gehemmt ist. Die Unterrichtsqualitätsbeurteilung in der zweiten Phase sollte vergleichbar sein. Es wird jedoch an unterschiedlichen Stellen von einer Personenabhängigkeit der zweiten Phase berichtet. Im dritten Beitrag wurde deshalb eine Interviewstudie mit Fachleiter*innen und Referendar*innen im Fach Chemie durchgeführt, um deren Sichtweisen auf Unterrichtsqualität abzubilden. Dazu sollte die Unterrichtsqualität einer aufgezeichneten Chemiestunde beurteilt werden, wobei entsprechend zur Praxis kein Instrument zur Beurteilung vorgegeben wurde. Stattdessen wurden die freien Beurteilungen im Nachhinein im naturwissenschaftsdidaktischen Framework verortet, sodass eine direkte Gegenüberstellung der Beurteilungen ermöglicht wurde. Dabei hat sich gezeigt, dass die Beurteilungen in der Praxis stark voneinander abweichen können, was auch in der abschließenden Benotung der Stunde deutlich wurde. Dieselbe Stunde wurde personenabhängig zwischen „sehr gut“ und „ungenügend“ benotet. Die Anzahl der verwendeten Merkmale, die konkrete Auswahl und auch die Beurteilung derselben Merkmale unterschied sich sowohl innerhalb als auch zwischen den Gruppen. Die fehlende gemeinsame Grundlage der Unterrichtsqualität in der zweiten Phase der Lehrkräftebildung führt offensichtlich zu einer abweichenden Entwicklung. Im vierten Beitrag wurde deshalb ein Ansatz erprobt, die professionelle Unterrichtswahrnehmung von Referendar*innen vignettengestützt im bestehenden Rahmen der Fachseminare weiterzuentwickeln. Der Fokus lag auf der Dimension der kognitiven Aktivierung, da deren Beurteilung besonders große gruppenspezifische Unterschiede aufgewiesen hat, sie aber dennoch wichtig für die Lernzuwächse der Schüler*innen ist. Dabei konnte der individuelle Einfluss der Fachleiter*innen auf die Entwicklung der professionellen Unterrichtswahrnehmung deutlich herausgestellt werden. Die Ergebnisse sprechen dafür, die Ausbildung von angehenden Lehrkräften an gemeinsamen Grundlagen zu orientieren und den individuellen Einfluss der Fachleiter*innen zu reduzieren. Ein möglicher Ansatz dafür stellt die breite Anwendung von webbasierten Plattformen für Unterrichtsaufzeichnungen dar, wie sie im fünften Beitrag am Beispiel von VirtU-net Chemie vorgestellt wird. Dabei muss perspektivisch der Austausch zwischen den Fachseminaren gestärkt und die Unterrichtsforschung und -praxis stärker vernetzt werden, um einer abweichenden Entwicklung in der Lehrkräftebildung entgegenzuwirken.

Schlagwörter: Unterrichtsqualität, Chemieunterricht, Lehrkräftebildung, Referendariat, Professionelle Unterrichtswahrnehmung

# I. Abstract

Evaluating instructional quality is a central concern in research and equally important for pre-service teacher education. Nevertheless, instructional quality is not clearly defined, but rather a conglomerate of different approaches. Some of these show an overlap, but there are also contradictions. The three basic dimensions of instructional quality are frequently used in research in German-speaking countries, but the research also points out missing aspects and deviating conceptualizations. Subject-specific aspects are particularly difficult in this regard, since they are difficult to capture accurately from a generic perspective. Recent developments in research on instructional quality address the need to supplement and adapt the basic dimensions and offer a broader approach with a synthesized framework. However, its application to the subject of chemistry had to be investigated first.

The first part of this project was a review of video studies used in science education. The aim was to determine relevant criteria of instructional quality for science lessons. These criteria were compared with the synthesized generic framework and the adapted science education perspectives framework was derived. In order to ensure the usability of the science education perspectives framework, a comparison with instruments of instructional quality from the subjects of biology and physics was conducted in the second part of the project. This supported the broad applicability of the science education perspectives framework, which subsequently provided an opportunity to map different views of instructional quality within a common framework.

There are only a few central guidelines for the second phase of pre-service teacher education, which impacts the comparability of development regarding instructional quality. The evaluation of instructional quality should be comparable for pre-service teachers, but many reports imply a dependence on individual advisors. In the third part of the project, interviews with chemistry specific advisors and chemistry pre-service teachers were conducted. They were asked to evaluate the instructional quality of a recorded chemistry lesson, without the use of pre-set criteria. This setting was chosen for a higher ecological validity, since there are usually no pre-set criteria or standardized instrument in the second phase. Instead, the evaluations were subsequently coded using the science education perspectives framework for a direct comparison of evaluations. The results showed that evaluations can differ to a large extent, which was also evident in the final grading of the lesson. The participants graded the lesson between "very good" to "insufficient”. The differences in evaluations can be found both within and between the groups. The lack of a common basis for instructional quality in the second phase of teacher education obviously leads to a divergent development. Thus, the fourth part of the project was used to test an approach for the comparable develop of professional vision in the second phase. The pre-service teachers were trained in a video-based setting within their usual teacher education seminars. The evaluation of cognitive activation was focused because it showed particularly large group-specific differences in the first comparison. Nevertheless, it is important for the students' learning gains. The research highlighted the influence of subject specific advisors on the development of professional vision. The results suggest that the training of pre-service teachers should be based on a common framework of instructional quality and the individual influence of their advisors should be reduced. One possible approach to these difficulties would be the widespread use of web-based platforms with lesson recordings, as presented with VirtU-net Chemistry in the fifth part of the project. In the future, the exchange between teacher education seminars must be strengthened. Furthermore, research and practice must be more closely connected in order to counteract a divergent development in teacher education.

Keywords: Instructional quality, chemistry education, pre-service teacher education, professional vision

## II. Vorwort

Die vorliegende kumulative Dissertation entstand am Institut für Didaktik der Naturwissenschaften in der Abteilung Didaktik der Chemie an der Leibniz Universität Hannover. Dieser Beitrag stellt eine zusammenfassende Darstellung der Dissertation dar und rahmt die Einzelbeiträge vor dem Hintergrund des gemeinsamen Forschungsanliegens. Die vollständige Forschungsarbeit umfasst auch die im Folgenden aufgeführten Artikel, die jeweils Einzelbeiträge der gesamten Dissertation darstellen. Die bisher unveröffentlichten Artikel sind im aktuellen Bearbeitungsstand abgebildet und können sich vor der abschließenden Veröffentlichung ändern. Die Eigenanteile der jeweiligen Artikel werden im Anhang aufgeführt und orientieren sich an der *Contributor Roles Taxonomy* (Allen et al., 2019).

Veröffentlichte Beiträge

1. Heinitz, B., & Nehring, A. (2020). Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung. In Unterrichtswissenschaft (Vol. 48, Issue 3, pp. 319–360). Springer VS. https://doi.org/10.1007/s42010-020-00074-8
   *Forschungsbeitrag - Double blind peer review*
2. Heinitz, B., Szogs, M., Förtsch, C., Korneck, F., Neuhaus, B. J., & Nehring, A. (2022). Unterrichtsqualität in den Naturwissenschaften. Eine vergleichende Gegenüberstellung von Ansätzen zwischen Fachspezifik und Generik. Zeitschrift Für Didaktik Der Naturwissenschaften, 28(1), 10. https://doi.org/10.1007/s40573-022-00146-5
   *Forschungsbeitrag - Double blind peer review*
3. Heinitz, B., & Nehring, A. (2023). Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors. International Journal of Science Education. https://doi.org/10.1080/09500693.2023.2213382
   *Forschungsbeitrag - Double blind peer review*
4. Heinitz, B., & Nehring, A (2024). Virtuelle Unterrichtshospitationen im Chemieunterricht – Eine Vernetzung der ersten und zweiten Phase der Lehrkräftebildung. *Der mathematische und naturwissenschaftliche Unterricht : MNU*, 182-190.
   *Praxisbeitrag - Herausgeber*innenreview*

Bisher unveröffentlichte Beiträge

5. Heinitz, B., & Nehring, A (submitted). Perception and Interpretation of Cognitive Activation in Science Pre-Service Teacher Education – The Necessity of Establishing a Common Approach
   *Forschungsbeitrag*

## III. Danksagung

An dieser Stelle möchte ich mich bei allen Personen bedanken, die mich während meiner Promotion begleitet haben und mein Vorhaben auf die einen oder andere Weise unterstützt haben.

Zuerst möchte ich mich bei meinem Doktorvater Prof. Dr. Andreas Nehring für die Betreuung meiner Arbeit und die Unterstützung in den letzten Jahren bedanken. Die vielen Freiheiten und Möglichkeiten bei der Ausgestaltung meines Projekts, das reichhaltige Feedback, sowie die vielen dazugehörigen Diskussionen haben zu einer sehr individuellen Arbeit geführt, bei der es nie langweilig wurde.

Weiterhin danke ich Prof. Dr. Friederike Korneck für die Zusammenarbeit im Rahmen des zweiten Beitrags (und auch darüber hinaus), das viele konstruktive Feedback auf Tagungen und die Übernahme des Zweitgutachtens.

Ich danke außerdem meinen Kolleg*innen vom IDN, mit denen ich in den letzten Jahren nicht nur viel Zeit bei der Arbeit, sondern auch darüber hinaus verbracht habe. Ich kann an dieser Stelle nicht alle Personen nennen, möchte jedoch besonders Steffi, Catha, Marco, Marvin und Alex als direkter Arbeitsgruppe danken, aber selbstverständlich auch Malte, Dennis und Robert, mit denen ich genauso gern und oft meine Zeit verbracht habe.

Mein besonderer Dank gilt außerdem Steffi für das stets offene Ohr und die hilfreichen Ratschläge, falls die Schwierigkeiten und Probleme der Promotion mal wieder zu groß schienen. Damit wurde mir ein Bild einer offenen und kooperativen Zusammenarbeit in der Wissenschaft vermittelt, das ich ebenso fortführen möchte.

Ich danke auch den Referendar*innen und Fachleiter*innen für die Teilnahme an meinen Studien, sowie den Lehrkräften, die sich für die Aufzeichnung ihres Unterrichts bereit erklärt haben. Ohne diese Beteiligung wäre das Projekt in dieser Form nicht möglich gewesen.

Abschließend danke ich meiner Familie, meinen Freunden und meiner Freundin Iris, die mich alle stets im Rahmen ihrer Möglichkeiten unterstützt haben.

# Inhaltsverzeichnis

## 1. Ausgangslage der Dissertation

*„We're all mad here. I'm mad. You're mad." […] "And how do you know that you're mad?" "To begin with," said the Cat, "a dog's not mad. You grant that?" "I suppose so," said Alice. "Well, then," the Cat went on, "you see a dog growls when it's angry, and wags its tail when it's pleased. Now I growl when I'm pleased, and wag my tail when I'm angry. Therefore I'm mad."*

– Alice's Adventures in Wonderland, Lewis Carroll

Im Kern wird hier ein simpler Gedanke vermittelt: Kommunikation kann durch Bezüge zu unterschiedlichen Referenzsystemen stark beeinträchtigt werden. Auch wenn sich die *Cheshire Ca*t bei dieser Erklärung sicher nicht auf Unterrichtsqualität bezogen hat, ergibt sich doch eine Parallele zur Beurteilung derselben.

Unterrichtsqualität ist ein Konzept, das mit vielen theoretischen Ansätzen verknüpft ist und dabei (oder auch gerade deshalb) an vielen Stellen eher abstrakt und schwer zugänglich ist. Die meisten Personen haben, speziell, wenn es um die Frage nach „gutem Unterricht" geht, ein implizites Verständnis, was diesen ausmacht (Praetorius et al., 2012). Dieses Verständnis wird früh durch die eigenen Erfahrungen geprägt, die jede Person beim Durchlaufen des Schulsystems macht. Somit entsteht bei allen eine individuelle Vorstellung, was einen „guten Unterricht" ausmacht und diese werden wiederum mit in die Lehrkräftebildung übertragen (Lortie, 1975). Der Einfluss individueller Vorstellungen geht dabei sogar so weit, dass Personen, die explizit zur Evaluation von Schulen verantwortlich sind, bei der Beurteilung einer Unterrichtsaufzeichnung ebenso individuelle Unterschiede aufweisen (Taut & Rakoczy, 2016). Weiterhin zeigt sich für Personen aus allen Gruppen der Lehrkräftebildung, dass das Verständnis von „gutem Unterricht" nicht immer mit einem effektiven Lernzuwachs auf Seiten der Schüler*innen einhergeht (Strong et al., 2011). Die Abweichung zwischen hoher Unterrichtsqualität und Lernzuwächsen lässt sich auch in der Unterrichtsforschung finden und kann in der Verwendung einer abweichenden Konzeptualisierung begründet sein (Christ et al., 2022).

Damit diese individuellen Ansätze systematisch zusammengeführt werden können, ist es zunächst einmal wichtig, die Frage zu klären, was alles für einen qualitativ hochwertigen Unterricht relevant sein kann. In der Forschung zur Unterrichtqualität gibt es viele Ansätze, die unterschiedliche Merkmale benennen (Helmke & Schrader 2008). Dabei weisen diese Ansätze teilweise Überschneidungen auf, nennen aber auch teils unterschiedliche Merkmale. Der besondere Anreiz von kurzen Auflistungen ist offensichtlich, geht aber auch immer mit dem Kritikpunkt der Unvollständigkeit einher. Besonders vor dem Hintergrund einer stetigen Entwicklung des Unterrichts (z.B. durch Digitalisierung) stellt sich auch stets die Frage nach einem Ergänzungsbedarf bestehender Ansätze. Einen weit verbreiteten Ansatz stellen die drei Basisdimensionen der Unterrichtsqualität dar (Klieme et al., 2001). Durch die konzeptuelle Rahmung von Dimensionen, die wiederum durch Merkmale weiter ausdifferenziert werden, bietet sich ein größeres Feld, in welchem Unterricht abgebildet werden kann. Viele Studien nehmen Bezug auf die Basisdimensionen und verorten ihre eigenen Ansätze in den bestehenden Dimensionen. Bei der Verortung unterschiedlicher Ansätze in den Basisdimensionen wird jedoch auch immer wieder ein Ergänzungsbedarf herausgestellt, der sowohl generische als auch fachspezifische Aspekte der Unterrichtsqualität umfasst (Praetorius et al., 2018, 2020a). Die Basisdimensionen sind breit definiert und entsprechend anwendbar. Dabei ist jedoch eine konkrete Operationalisierung notwendig, die dann wiederum die grundlegende Struktur infrage stellt. Aus dem Ergänzungsbedarf hat sich das Bestreben entwickelt, eine breitere Basis für die Abbildung der Unterrichtsqualität zu finden, die durch das *Syntheseframework* (Praetorius & Charalambous, 2018) angeboten wird. Offen ist hierbei jedoch die Frage geblieben, inwiefern dieser Ansatz, der ebenfalls auf Mathematikunterricht beruht, auch für andere Fächer anwendbar ist und ob damit eine umfassende Abbildung der Unterrichtsqualität möglich ist.

Die Pluralität der Forschungsansätze zur Unterrichtsqualität ist ebenfalls für die Praxis der Lehrkräftebildung relevant. Diese ist wenig standardisiert (Weber & Czerwenka, 2021), wodurch

Fachleiter*innen[1] häufig unterschiedliche theoretische Grundlagen verwenden und die Ausbildung entlang einer individuellen Erwartung an „guten Unterricht“ auslegen (Wiernik, 2020). Dies sorgt wiederum dafür, dass unterschiedliche Schwerpunkte in der Ausbildung zu finden sind und der Eindruck vermittelt wird, dass die zweite Phase besonders standortabhängig ist (Döbrich & Storch, 2012). Angehende Lehrkräfte werden somit im Verlauf der Ausbildung mit mehreren, möglicherweise widersprüchlichen Vorstellungen zur Unterrichtsqualität konfrontiert. So haben sie zunächst eine individuelle Vorstellung auf Basis ihrer Erfahrungen generiert, dieses adaptieren sie, wenn sie neue theoretische Inhalte im Studium vermittelt bekommen und im Referendariat lernen sie wiederum neue Sichtweisen kennen, die von diesen abweichen können.
Aus diesem Grund wäre es wichtig, die Ausbildung angehender Lehrkräfte an einer gemeinsamen Grundlage auszurichten und die erste und zweite Phase stärker miteinander zu vernetzen (Weber & Czerwenka, 2021). Diese Grundlage sollte genutzt werden, um Unterschiede gezielt zu thematisieren und ein gemeinsames Bild im Sinne einer *shared vision of teaching* zu entwickeln (Kang & van Es, 2019). Hierzu ist es wichtig, an den aktuellen Stand der Lehrkräftebildung anzuschließen, wozu dieser zunächst näher untersucht werden musste. Dadurch ergaben sich drei grundlegende Aufgaben für die Dissertation. Zunächst die Abbildung von Unterrichtsqualität für Chemieunterricht aus Sicht der Forschung. Dann die Erfassung von Unterrichtsqualitätsbeurteilungen in der zweiten Phase der Chemielehrkräftebildung und abschließend eine mögliche Zusammenführung beider Perspektiven für eine Verbesserung der Lehrkräftebildung.

## 2. Theoretischer Hintergrund

Durch den kumulativen Ansatz hat sich der theoretische Hintergrund im Verlauf der Zeit parallel zu den Einzelbeiträgen weiterentwickelt. So gab es zum einen neue Erkenntnisse im Bereich der Systematisierung von Unterrichtsqualität, besonders, wenn es um kollaborative Ansätze geht und genauso im Bereich der professionellen Unterrichtswahrnehmung. Die eigenen Arbeiten haben ebenso dazu geführt, dass sich neue theoretische Grundlagen für die folgenden Arbeiten ergeben haben. An dieser Stelle soll jedoch ein vollständiger Überblick über die relevante theoretische Grundlage erfolgen. Der Aufbau orientiert sich an der Reihenfolge der Einzelbeiträge, ausgehend von einer theoretischen Betrachtung der Unterrichtsqualität, hin zu einer Entwicklung der professionellen Unterrichtswahrnehmung in der Praxis der Lehrkräftebildung. Die Grundlagen der einzelnen Beiträge des Projekts werden in Abschnitt 6 „Vernetzung der Einzelbeiträge“ jeweils im Unterpunkt „Ausgangslage“ umrissen.

### *2.1. Zentrale Terminologien für diesen Beitrag*

Durch die Pluralität der Forschungsansätze zur Unterrichtsqualität ergibt sich eine breite Fülle an Terminologien, die in diesem Themenbereich verwendet werden. Häufig werden dabei jedoch eher abstrakte Konzepte beschrieben, was dazu führt, dass sich die Terminologien an einigen Stellen nicht eindeutig voneinander abgrenzen lassen. Teilweise werden dieselben Begriffe für unterschiedliche Konzepte benutzt und an anderer Stelle werden unterschiedliche Begriffe mit derselben Bedeutung verwendet. Auch im Rahmen der Einzelbeiträge der Dissertation hat sich durch den Bezug zu unterschiedlichen theoretischen Grundlagen und die stetige Weiterentwicklung teilweise eine Adaption der verwendeten Terminologie ergeben. Damit in diesem Beitrag ein eindeutiges Verständnis vorliegt, werden in Tabelle 1 die grundlegenden Begriffe definiert.

[1] Die Ausbilder*innen an den Fach-/Studienseminaren werden abhängig vom Standort unterschiedlich bezeichnet. Zur Vereinfachung wird in diesem Beitrag durchgängig der Begriff Fachleiter*innen verwendet.

*Tabelle 1: Zentrale Terminologien des Manuskripts. Erweitert und adaptiert nach Heinitz et al., (2022).*

| Begriff | Definition |
|---|---|
| Systematisierung/Framework (der Unterrichtsqualität) | Eine Auflistung von Qualitätsmerkmalen, die auf Basis einer gemeinsamen theoretischen Grundlage auf mehrere Abstraktionsebenen verteilt ist. |
| Instrument | Eine Möglichkeit zur Messung von Unterrichtsqualität auf Grundlage unterschiedlicher Merkmale, Kriterien und Indikatoren. Instrumente können zweckbezogen aus einem Ansatz abgeleitet werden. |
| Dimension | Die höchste Ebene einer Systematisierung |
| Merkmal | Facette zur Erfassung der Unterrichtsqualität, die sich einer Dimension unterordnen lässt. |
| Kriterium | Ausdifferenzierung eines Merkmals[2]. Kriterien weisen einen Bezug zum beobachtbaren, unterrichtlichen Handeln von Lehrenden und Lernenden auf. |
| Indikator/Item | Fragen oder Aufforderungen eines Instruments, die ein Urteil in Bezug auf die Ausprägung eines Kriteriums oder eines Merkmals einfordern. |
| Hierarchie (der Systematisierung/ des Frameworks) | Der zugrundeliegende Aufbau einer Systematisierung von Unterrichtsqualität. Dimensionen, Merkmale, Kriterien etc. werden auf verschiedenen Ebenen abgebildet. Eine Ebene weist jeweils eine vergleichbare Abstraktion auf. Prinzipiell können zwischen der höchsten Ebene (Dimension) und den direkt beobachtbaren Indikatoren/Items beliebig viele Ebenen liegen, sofern sie sich theoretisch begründen lassen. |
| Abstraktionsgrad | Bildet die Generalisierung innerhalb einer Hierarchie ab. Hohe Ebenen sind abstrakt beschrieben, erfassen inhaltlich mehrere Aspekte der Unterrichtsqualität und benötigen Schritte der Operationalisierung zur Beurteilung von Unterricht. Niedrigere Ebenen werden typischerweise konkreter im Hinblick auf die Beurteilung von unterrichtlichen Aktivitäten beschrieben. |
| Beurteilung | Umfasst mehrere Schritte und stellt somit einen übergreifenden Begriff dar. Hierzu gehört die Auswahl von Merkmalen oder Kriterien, die Bewertung derselben und auch die Zusammenführung aller Bewertungen zu einem abschließenden Fazit bzw. einer Note. |
| Bewertung | Ist Teil einer Beurteilung, bezieht sich aber nur auf ein einzelnes Merkmal oder Kriterium. Je nach Skala kann sie konkrete Punkte umfassen oder auch nur eine Tendenz im Sinne von „gut“ oder „schlecht“. |

### *2.2. Unterrichtqualität aus Sicht der Unterrichtsforschung*

Der Begriff Unterrichtsqualität folgt keiner einheitlichen Definition. Vielmehr ist es ein abstraktes (mehrdimensionales) Konstrukt, das abhängig von der Untersuchung unterschiedlich ausgelegt werden kann oder auch (fach-)spezifisch fokussiert wird (Brunner, 2018). Unterrichtsqualität ist abhängig vom

[2] Die Abgrenzung zwischen Merkmalen und Kriterien ist in der Unterrichtsforschung nicht immer trennscharf und kann in unterschiedlichen Studien voneinander abweichen. In diesem Beitrag wird nach Möglichkeit der Begriff Merkmal verwendet, um unterschiedliche theoretische Grundlagen auf einer abstrakteren Ebene abzubilden. Der Begriff „Kriterium“ wird explizit verwendet, wenn eine Ebene mit niedrigerer Abstraktion beschrieben wird.

jeweiligen Ziel des Unterrichts und kann entsprechend durch unterschiedliche Merkmale beschrieben werden. Sowohl in der Forschung als auch in der Praxis kann jedoch eine möglichst einheitliche und objektive Beurteilung der Unterrichtsqualität als wünschenswert angenommen werden. Im Zusammenhang der Unterrichtsqualität stellt sich auch die Frage, was einen „guten Unterricht" ausmacht. Ein möglichst hoher Lernzuwachs der Schüler*innen scheint dabei naheliegend, allerdings wird dies nicht als einziger Faktor für einen „guten Unterricht" herangezogen. So ist auch eine Unterscheidung in „good teaching", das sich an normativen Standards orientiert und „effective teaching", das sich an am Erreichen der Lernziele orientiert, möglich. In Kombination ergeben diese beiden Aspekte dann das „quality teaching" (Berliner, 2005). Dass die Aufgabe von Unterricht mehr als Lernzuwächse umfasst, spiegelt sich auch in den Curricula und Bildungsstandards wider (z.B. Bildungsbeitrag des Faches Chemie, Niedersächsisches Kultusministerium, 2022). Die Qualität des Unterrichts muss folglich vor seinem jeweiligen Ziel beurteilt werden und auch den Kontext berücksichtigen. Dadurch haben sich im Verlauf der Zeit diverse Ansätze und Instrumente zur Beurteilung der Unterrichtsqualität entwickelt, die auf unterschiedliche Aspekte des Unterrichts fokussieren und diesen mehr oder weniger umfänglich abbilden (Helmke & Schrader 2008). Diese sind teilweise aus theoretischen Überlegungen begründet und teilweise aus empirischen Daten abgeleitet. Eine grundlegende Unterscheidung, die bei der Beschreibung der Merkmale getroffen werden kann, ist die in Sicht- und Tiefenstrukturen. Sichtstrukturen umfassen die übergeordnete Strukturierung, die direkt ersichtlich ist, wohingegen Tiefenstrukturen die Qualität der Interaktion zwischen Lehrenden, Lernenden und dem Unterrichtsgegenstand umfassen. Besonders für die Erreichung der Lernziele sind in vielen Fällen die Tiefenstrukturen von großer Bedeutung (Kunter & Ewald, 2016).

### *2.2.1. Drei Basisdimensionen der Unterrichtsqualität*

Je nach Verständnis der Unterrichtsqualität kann die Herangehensweise für deren Beurteilung abweichen. In der Regel erfolgt eine Beurteilung der Unterrichtsqualität jedoch über zuvor festgelegte Merkmale oder Kriterien. Einer der am weitesten verbreiteten Ansätze in der deutschsprachigen Forschung zur Systematisierung von relevanten Qualitätsmerkmalen sind die drei Basisdimensionen (Klieme et al., 2001). Die Basisdimensionen ließen sich empirisch aus der TIMS-Videostudie ableiten und umfassen die „Unterrichts- und Klassenführung", die „Schülerorientierung" und die „kognitive Aktivierung". Die Dimensionen sind nicht unabhängig voneinander, haben jedoch jeweils eine spezifische Bedeutung für den Unterricht. Die Unterrichts- und Klassenführung stellt die notwendigen Rahmenbedingungen her, die Schülerorientierung zielt auf die Motivation und das Interesse der Schüler*innen ab und die kognitive Aktivierung bestimmt das Ausmaß des Lernzuwachses.
Die Basisdimensionen finden eine breite Anwendung, was sich auch in einem internationalen Literatur-Review gezeigt hat (Praetorius et al., 2018). Dabei konnten 21 quantitative empirische Studien (z.B. PISA, Klieme & Rakoczy, 2003; COACTIV, Kunter & Voss, 2013) mit insgesamt 39 dazugehörigen Artikeln herausgestellt werden, die einen direkten Bezug zu den drei Basisdimensionen aufweisen. Auch wenn die Basisdimensionen zunächst aus dem Mathematikunterricht abgeleitet wurden, waren die Studien in unterschiedlichen Fächern verortet und umfassten auch den naturwissenschaftlichen Unterricht. Auch die Klassenstufe und die Perspektive der beurteilenden Personen variierten zwischen den Studien, was die breite Anwendbarkeit der Basisdimensionen weiter unterstreicht. Dennoch hat sich im Rahmen des Reviews auch gezeigt, dass die angenommenen theoretischen Zusammenhänge zwischen den Basisdimensionen und ihrem Effekt auf den Unterricht nicht immer eintreten. Somit konnte zwar die breite Anwendbarkeit gezeigt, aber auch der Bedarf nach einer Anpassung der Basisdimension herausgestellt werden. Dieser begründet sich nicht nur über fehlende Aspekte, sondern auch eine fachspezifische Ausdifferenzierung der Dimensionen.

### *2.2.2. Anpassung und Erweiterung der drei Basisdimensionen*

Die Beurteilung von Unterrichtsqualität kann durch viele Faktoren beeinflusst werden. Dennoch, oder auch gerade deshalb, besteht in vielen Fällen das Bestreben, eine gemeinsame generische Grundlage zu verwenden. Dazu haben sich die drei Basisdimensionen in vielen Fällen etabliert, wobei jedoch ein

grundsätzlicher Anpassungs- und Ergänzungsbedarf bezüglich der Vollständigkeit, Operationalisierung und Fachspezifität herausgestellt werden kann (Praetorius et al., 2018, 2020a).
Eine erweiterte Grundlage wird durch das *Syntheseframework* (Praetorius & Charalambous, 2018) geboten. Dieses konnte durch eine Zusammenführung von 12 empirisch genutzten Beobachtungsinstrumenten der Mathematik herausgearbeitet werden und lässt sich konzeptuell in sieben Dimensionen der Unterrichtsqualität unterteilen. Dabei werden die Basisdimensionen teilweise aufgeteilt und teilweise durch weitere Aspekte ergänzt. Das *Syntheseframework* umfasst dabei die Dimensionen I. Auswahl und Thematisierung von Inhalten und Fachmethoden, II. Kognitive Aktivierung, III. Unterstützung des Übens, IV. Formatives Assessment, V. Unterstützung des Lernens aller Schüler*innen, VI. Sozio-emotionale Unterstützung und VII. Klassenführung[3]. Grundsätzlich ist das *Syntheseframework* so angelegt, dass es zwar generisch formuliert ist, aber ebenso fachspezifische und auch hybride Merkmale der Unterrichtsqualität erfasst. Um die fächerübergreifende Anwendbarkeit des *Syntheseframeworks* zu überprüfen, wurde es zur Verortung von Ansätzen aus den Fachbereichen Geschichte, Sport und Naturwissenschaften genutzt (Praetorius et al., 2020b). Dabei ließe sich an vielen Stellen bereits eine direkte Anwendbarkeit des *Syntheseframeworks* herausstellen, wobei es an anderen Stellen ebenso einen fachspezifischen Ergänzungs- oder Differenzierungsbedarf aufwies.
Häufig kann der theoretische Anpassungs- oder Ergänzungsbedarf direkt aus dem Verwendungszweck heraus abgeleitet werden, sodass unterschiedliche Grundlagen zusammengeführt werden. Es lässt sich jedoch auch aus empirischen Studien eine Aufteilung oder auch Ergänzung der Basisdimensionen ableiten (Jentsch et al., 2021).

### *2.2.3. Herausforderungen bei der Beurteilung von Unterrichtsqualität*

Den Tiefenstrukturen des Unterrichts kommt eine besondere Bedeutung für die Lernzuwächse der Schüler*innen zu (Kunter & Ewald, 2016). Jedoch sind es gerade hoch inferente Merkmale, die bei einer Beurteilung der Unterrichtsqualität zu hohen Abweichungen zwischen unterschiedlichen Personen führen können. Der hohe Interpretationsrahmen sorgt dafür, dass die Beurteilung eines Merkmals an unterschiedliche Handlungen geknüpft sein kann. Dabei können unterschiedliche theoretische Grundlagen und Erfahrungen eine Erwartungshaltung für eine gute oder auch schlechte Ausprägung eines Merkmals erzeugen. Diese „impliziten Theorien" über guten Unterricht können sehr individuell ausgeprägt sein und ggf. nicht einmal der beurteilenden Person selbst vollständig bewusst vorliegen (Praetorius et al., 2012). Durch individuelle Vorstellungen über guten Unterricht werden auch durch Expert*innen Indikatoren herangezogen, die nicht notwendigerweise in einem festen Framework verortet sind oder auf spezifischen Theorien beruhen, sodass die Beurteilung entsprechend für andere Personen nicht zugänglich ist (Taut & Rakoczy, 2016). Auch stimmt die Vorstellung eines „guten Unterrichts" bei vielen Personen in der Lehrkräftebildung nicht immer mit einem hohen Lernzuwachs der Schüler*innen überein (Strong et al., 2011). Besonders für die Ausbildung angehender Lehrkräfte können hieraus klar ersichtlich Schwierigkeiten erwachsen.
Auch bei der Nutzung eines standardisierten Instruments und damit Indikatoren, die in einem festen Framework verortet sind, können Schwierigkeiten auftreten. Ein reliables Instrument führt nicht dazu, dass die Ergebnisse entsprechend immer reliabel sind, da auch die Beurteilenden selbst, sowie das Ratertraining und das zugrundeliegende Beurteilungssystem einen Einfluss auf die Anwendung des Instruments haben (Hill et al., 2012). Da diese Rahmenbedingungen bereits die Erstellung und Testung eines Instruments beeinflussen können, sprechen sich Hill et al. (2012) auch dafür aus, die Gütekriterien eines Instruments entsprechend vorsichtig zu interpretieren. Dies gilt vor allem dann, wenn sie in unmittelbarer Nähe bestimmter Grenzwerte liegen. Die Anwendung eines Instruments zur Beurteilung der Unterrichtsqualität ist in der Regel an ein Ratertraining gebunden. Hierbei hat sich jedoch gezeigt, dass sich diese nicht notwendigerweise positiv auf die Reliabilität der Beurteilenden auswirken (Praetorius et al., 2012). Entsprechend müsste bereits das Ratertraining an eigene Gütekriterien gebunden werden. Praetorius et al. (2012) stellen außerdem heraus, dass eine hohe Reliabilität nicht notwendigerweise mit einer hohen Validität zusammenhängt, sodass diese

[3] Für den direkten Vergleich mit den drei Basisdimensionen wurde hier direkt die deutsche Übersetzung des *Syntheseframeworks* gewählt, auch wenn diese erst später veröffentlicht wurde.

Gütekriterien grundsätzlich beide in Betracht gezogen werden sollten. Grundsätzlich könnte eine hohe Reliabilität auch durch die Messung eines abweichenden Konstrukts erreicht werden.
Die Beurteilung von Unterrichtsqualität durch externe Beobachter*innen hängt mit einigen Vor-, aber auch Nachteilen zusammen (Praetorius, 2013). Dabei wird vor allem herausgestellt, dass die Beurteilenden selbst nicht Teil des Unterrichtsgeschehens sind und in der Regel durch Trainings speziell geschult werden. Weiterhin sind die externen Beobachter*innen häufig an vielen Unterrichtsbeobachtungen beteiligt, sodass sie eine entsprechend hohe Vergleichsmenge haben. Als Nachteil kann dabei vor allem die beschränkte Zugänglichkeit zu relevanten Informationen gezählt werden, die in Abschnitt 2.2.4 weiter ausgeführt wird.
Insgesamt führen die Schwierigkeiten bei der Beurteilung dazu, dass bestimmte Unterrichtsqualitätsmerkmale nur mit sehr viel Aufwand beurteilt werden können. So führt die erschwerte Zugänglichkeit dazu, dass bis zu zehn Rater*innen für eine reliable Beurteilung der Lernendenunterstützung benötigt werden (Praetorius et al., 2012). Auch der Umfang der Beobachtungen kann variieren. Eine einzelne Stunde kann möglicherweise einen zu beschränkten Einblick bieten, sodass für eine reliable Beurteilung von komplexeren Dimensionen, wie z.B. der kognitiven Aktivierung, bis zu neun Stunden notwendig sein können (Praetorius et al., 2014). Dies kann neben der Zugänglichkeit und den Interpretationsmöglichkeiten auch durch die zeitliche Stabilität bedingt sein. Vor allem Merkmale der kognitiven Aktivierung weisen eine geringe zeitliche Stabilität auf (z.B. Gabriel-Busse & Lipowsky, 2021). Es gibt auch Hinweise darauf, dass bereits ein kurzer Unterrichtsausschnitt ausreichen kann, um eine reliable und valide Beurteilung der Unterrichtsqualität zu erzielen (Begrich et al., 2017). Dabei ist jedoch eine Differenzierung zwischen spezifischen Merkmalen der Unterrichtsqualität weniger genau möglich. Darüber hinaus hat sich auch gezeigt, dass zeitlich weniger stabile Merkmale z.B. im Bereich der kognitiven Aktivierung nur unzuverlässig in einem kurzen Unterrichtsausschnitt beurteilt werden können (Begrich et al., 2020, 2021). Ein kurzer Einblick in eine Unterrichtsstunde kann jedoch ausreichen, um eine intuitive Beurteilung hervorzurufen, die auch über einen längeren Beobachtungszeitraum hinweg stabil bleibt und bei hohem domänenspezifischem Wissen eher zutreffend ist (Begrich et al., 2020).
Die bisherigen Überlegungen setzten voraus, dass dasselbe Instrument zur Erfassung der Unterrichtsqualität verwendet wird. Eine Beurteilung derselben Unterrichtsstunde mit unterschiedlichen Instrumenten kann jedoch auch zu einer stark unterschiedlichen Einschätzung der Unterrichtsqualität führen (Brunner, 2018). In der Regel werden deshalb Studien mit unterschiedlichen theoretischen Grundlagen auch nicht direkt miteinander verglichen. Selbst bei Verwendung einer gemeinsamen theoretischen Grundlage, kann es jedoch vorkommen, dass diese unterschiedlich ausgelegt wird, was ebenfalls zu Abweichungen in der Beurteilung führen kann (Christ et al., 2022).

### *2.2.4. Perspektivenunterschiede bei der Beurteilung von Unterrichtsqualität*

Die Beurteilung von Unterrichtsqualität erfolgt in der Forschung häufig aus der Perspektive von externen Beobachter*innen (Clausen, 2002). Grundsätzlich ist jedoch auch eine Beurteilung der Unterrichtsqualität aus Perspektive der Schüler*innen und der Lehrkräfte möglich. Viele Studien zeigen jedoch, dass diese drei Perspektiven in ihrer Beurteilung häufig voneinander abweichen (z.B. Clausen, 2002; Camburn & Barnes, 2004; Kunter and Baumert, 2006; Desimone et al., 2010; Fauth et al., 2014). Es lassen sich zwar immer wieder Gemeinsamkeiten in der Beurteilung finden, diese sind jedoch auf einzelne Merkmale und Gruppen begrenzt. Eine vollständige Übereinstimmung zwischen allen drei Perspektiven ist eher selten und kommt in erster Linie bei direkt beobachtbaren Merkmalen der Oberflächenstruktur vor. Die Unterschiede zwischen den Gruppen sind durch eine abweichende Zugänglichkeit zu spezifischen Merkmalen, sowie die eigene Betroffenheit bedingt und können durch die „reference perspective matrix“ (Fauth et al., 2020) eingängig visualisiert werden. Dasselbe Merkmal kann unterschiedlich bewertet werden, wenn unterschiedliche Referenzpunkte einbezogen werden.
Abhängig von der Stellung der Person im System kann auch die Grundhaltung beeinflusst sein. Vor allem bei der Beurteilung des eigenen Unterrichts kann es bei Lehrkräften zu einem *protective bias* kommen, der zu einer positiveren Beurteilung der eigenen Leistung führt (Vazire, 2010). Ähnliche positivere Ausprägungen in der Beurteilung können auch angenommen werden, wenn persönliche

Beziehungen zwischen der beurteilenden und der beurteilten Person bestehen. Eine persönliche Beziehung kann jedoch auch zu einer valideren Beurteilung führen, wenn Persönlichkeitsmerkmale aus externer Perspektive abgefragt werden (Vazire, 2010). Eine Beurteilung aus externer Perspektive ist grundsätzlich durch die begrenzte Verfügbarkeit von Kontextinformationen gekennzeichnet. So zeigt sich auch bei erfahrenen Lehrkräften, dass diese ihre Expertise in der Beurteilung des eigenen Unterrichts nicht unmittelbar auf den Unterricht fremder Personen übertragen können (Berliner, 1989). Die Frage nach der Expertise in der Unterrichtsbeurteilung ist damit nicht eindeutig zu klären und wird in Abschnitt 2.6.1 weiter vertieft.

#### *2.2.5. Fachspezifik in der Beurteilung von Unterrichtsqualität*

Neben den Unterschieden, die durch die Perspektive der Beurteilenden bedingt sind, können auch Unterschiede durch das betrachtete Fach entstehen. Ein Fachunterricht enthält zwar Elemente, die aus einer generischen Perspektive heraus beurteilt werden können, es gibt jedoch auch häufig fachspezifische Merkmale (Praetorius et al., 2020b). Die Frage danach, wie Fachspezifik verstanden wird, kann auch den Vergleich unterschiedlicher Ansätze zur Beurteilung der Unterrichtsqualität erschweren (Heinitz et al., 2022). Das zugrundeliegende Verständnis von Fachspezifik beeinflusst nicht direkt die Beurteilung der Unterrichtsqualität, kann aber einen Einfluss auf die Auswahl oder Operationalisierung eines Instruments haben. Grundlegend lässt sich das Professionswissen einer Lehrkraft neben dem fächerübergreifenden pädagogischen Wissen auch in ein fachliches und fachdidaktisches Wissen unterteilen (Shulman, 1986, 1987). Weist ein Instrument fachspezifische Anteile auf, kann entsprechend fachliches oder fachdidaktisches Wissen für die Beurteilung der beobachteten Unterrichtshandlung notwendig sein (Neuhaus, 2021). Ist dieser Anteil im Instrument nicht direkt ersichtlich, kann die Beurteilung durch ein fehlendes fachliches oder fachdidaktisches Wissen unbemerkt beeinflusst werden. Somit hat das Verständnis der Fachspezifik einen indirekten Einfluss auf die Beurteilung der Unterrichtsqualität. Grundsätzlich lassen sich in der Forschung Hinweise darauf finden, dass fachspezifische Merkmale einen großen Einfluss auf die Lernzuwächse der Schüler*innen haben (Seidel & Shavelson, 2007). Dadurch wird die Bedeutung von fachlichem und fachdidaktischem Wissen hervorgehoben und gezeigt, dass ein rein generisches Instrument in den meisten Fällen nicht für eine umfassende Beurteilung der Unterrichtsqualität geeignet ist. Somit lässt sich auch für relativ breit akzeptierte und verwendete generische Ansätze zur Erfassung der Unterrichtsqualität, wie z.B. die drei Basisdimensionen, ein fachspezifischer Ergänzungsbedarf herausstellen. Brunner (2018) hebt dies vor allem dadurch hervor, dass eine Unterrichtsstunde mit fachlichen Fehlern durchaus positiv in einer generischen Beurteilung abschneiden könnte, wenn die fachliche Korrektheit nicht explizit erfasst wird. Entsprechend werden generische Ansätze zwar in vielen Fällen verwendet, jedoch fachspezifisch angepasst und erweitert (z.B. Korneck et al., 2017; Neuhaus, 2021; Heinitz & Nehring, 2020).

### *2.3. Kollaborative Ansätze zur Zusammenführung der Unterrichtsqualität*

Durch die Vielzahl von Ansätzen zur Beurteilung der Unterrichtsqualität wird die Zusammenführung und Analyse unterschiedlicher Studien erschwert (Schlesinger & Jentsch, 2016, Christ et al., 2022). Die Pluralität der Forschungsansätze kann dabei durchaus zielführend sein, um den Diskurs zwischen unterschiedlichen Standpunkten zu fördern. Für eine effektive und kumulative Weiterentwicklung des Forschungsfelds scheint die Stärkung kollaborativer Ansätze jedoch notwendig zu sein (Charalambous & Praetorius, 2022). Zu diesem Zweck können Ansätze, wie das *Syntheseframework* (Praetorius & Charalambous, 2018) hilfreich sein, um unterschiedliche Konzeptualisierungen in einem breiten Rahmen zusammenzuführen. Dabei ergeben sich jedoch auch Schwierigkeiten, die bereits im Vorfeld antizipiert werden sollten. So müssen gemeinsame Ziele festgelegt, Differenzen in verwendeten Terminologien und Strukturen berücksichtigt, methodische Herausforderungen explizit adressiert, Transparenz geschaffen und auch Fragen der Finanzierung gemeinsamer Vorhaben beantwortet werden (Charalambous et al., 2021).

Die grundlegende Idee des *Syntheseframeworks* wurde deshalb weiterentwickelt, woraus das *MAIN-Teach-Modell* abgeleitet werden konnte (Charalambous & Praetorius, 2020). Das *MAIN-Teach-Modell*

bietet eine Betrachtung des *Syntheseframeworks* in mehreren Ebenen. Dabei werden eine Kernebene, eine begünstigende Ebene und eine zugrundeliegende Ebene benannt, die insgesamt durch acht Dimensionen der Unterrichtsqualität ausdifferenziert werden (Praetorius et al., 2023). Die Abweichung in der Anzahl der Dimensionen zum *Syntheseframework* ergibt sich durch die zugrundeliegende Ebene der Adaptivität, die zuvor nur untergeordnet enthalten war. Aus dem *MAIN-Teach-Modell* konnte weiterführend ein Beobachtungsinstrument zur Beurteilung der Unterrichtsqualität abgeleitet werden (Wemmer-Rogh et al., 2023).

### *2.4. Lernzuwächse als Ergebnis hoher Unterrichtsqualität*

Lernzuwächse von Schüler*innen sind ein relevanter Indikator für Unterrichtsqualität und in vielen Fällen ein Ergebnis von "effective teaching" (Berliner, 2005). Dabei lassen sich in unterschiedlichen Studien Zusammenhänge zwischen Merkmalen der Unterrichtsqualität und Lernzuwächsen von Schüler*innen zeigen (Seidel & Shavelson, 2007; Kersting et al., 2012; Kyriakides et al., 2013; Labudde et al., 2014; Kang, 2020). Diese sind jedoch nicht immer für alle Merkmale gleichermaßen zu finden. Zum einen kann es natürlich darauf zurückzuführen sein, dass die entsprechenden Merkmale eher Teil des „good teaching“ (Berliner, 2005) sind und damit keinen unmittelbaren Einfluss auf die Lernzuwächse haben, sondern eher die Rahmenbedingungen beeinflussen. Zum anderen kann es aber auch daran liegen, dass die Messungen der Unterrichtsqualität nicht immer valide sind, besonders wenn es sich um hoch inferente Merkmale handelt (Praetorius et al., 2012, Schlesinger & Jentsch, 2016). Dabei stellen gerade hoch inferente Merkmale häufig einen besseren Prädiktor für die Leistung der Schüler*innen dar (Begrich et al., 2017). Auf Seite der Lernenden ist es vor allem das bereits bestehende Vorwissen, das einen großen Einfluss auf das Lernen hat (Simonsmeier et al., 2021). Somit sind Lernzuwächse auch nicht unmittelbar erwartbar, wenn die untersuchten Merkmale unabhängig vom Vorwissen positiv ausgeprägt sein können. Dadurch würde sich eine ähnliche Diskrepanz zwischen der Beurteilung durch ein Instrument und einer theoretisch begründeten Voraussetzung ergeben, wie es bei Brunner (2018) für die fachliche Korrektheit beschrieben wird.

### *2.5. Professionelle Unterrichtswahrnehmung*

Die Grundidee einer professionellen Wahrnehmung wurde durch Goodwin (1994) geprägt. Dabei ging es zunächst nicht um eine spezifische Profession, sondern um eine übergreifende Gemeinsamkeit von Fachkräften. Die Wahrnehmung ist bei diesen so spezifisch angepasst, dass sie besonders relevante Aspekte sofort erkennen und ihre Handlungen entsprechend ausrichten. Die professionelle Wahrnehmung ist zwar ein individueller Prozess und an das Wissen und die Erfahrungen einer Person gekoppelt, es lassen sich jedoch vergleichbare Wahrnehmungsmuster innerhalb einer Profession finden.

Die Übertragung in die Lehrkräftebildung im Sinne einer professionellen Unterrichtswahrnehmung ist stark durch Sherin (2001) geprägt. Dabei geht es zunächst darum, Lehrkräfte auf spezifische Aspekte hinzuweisen, damit diese erkannt und gezielt in den Blick genommen werden (*selective attention*). In einem nächsten Schritt sollten sie dann auf Basis ihres Wissens flexibel und angemessen auf die jeweiligen Unterrichtssituationen reagieren können (*knowledge based reasoning*). Somit ergibt sich die Möglichkeit, die Unterrichtsqualität im laufenden Unterrichtsprozess zu beeinflussen.

Die professionelle Unterrichtswahrnehmung kann als situationsspezifische Fähigkeit beschrieben werden (Blömeke et al., 2015). Der Prozess ist dabei in drei Schritte unterteilt, die im Modell nach Blömeke et al. als Wahrnehmung (*perception*), Interpretation (interpretation) und Entscheidungsfindung (decision making) bezeichnet werden. Diese sind durch verschiedene Voraussetzungen auf Seiten der Lehrkraft beeinflusst, die als Dispositionen zusammengefasst werden können. Die professionelle Unterrichtswahrnehmung mündet in einer beobachtbaren Handlung (*performance*), die auch als vierter Schritt in der Kette betrachtet werden kann.

#### *2.5.1. Entwicklung von professioneller Unterrichtswahrnehmung*

Ein gezieltes Training der professionellen Unterrichtswahrnehmung kann mit Hilfe von Unterrichtsaufzeichnungen erfolgen und wurde von den ursprünglichen Überlegungen heraus zu

sogenannten *video clubs* weiterentwickelt (z.B. Sherin, 2007; Sherin & van Es, 2009). Im Gegensatz zu vollständigen Unterrichtsaufzeichnungen werden Videovignetten (also kurze Ausschnitte aus dem Unterrichtsvideo) häufig eingesetzt, um die Komplexität des Unterrichts auf spezifische Aspekte zu reduzieren (Beck et al., 2002). Dadurch können Videovignetten so gewählt werden, dass sie eine direkte Verknüpfung zwischen dem theoretischen Wissen und spezifischen Unterrichtssituationen ermöglichen (Sherin & van Es, 2009).
Theoretisches Wissen hat einen besonderen Stellenwert, da es einen größeren Effekt auf die professionelle Wahrnehmung hat, als praktische Erfahrungen (Stürmer et al., 2014).
Dadurch bietet sich eine Förderung der professionellen Wahrnehmung bereits in der stärker theoretisch ausgelegten ersten Phase der Lehrkräftebildung an. Es ist daher auch nicht überraschend, dass der Einsatz von Unterrichtsvideos in der universitären Lehrkräftebildung weit verbreitet ist (Blomberg et al., 2013). Mittlerweile gibt es darüber hinaus ein breites Angebot verschiedener Onlineplattformen, die Videovignetten für die Lehrkräftebildung zur Verfügung stellen (Junker et al., 2022).
Das Training der professionellen Unterrichtswahrnehmung kann dafür genutzt werden, um einen gemeinsamen Blick auf das Unterrichten zu entwickeln, was Kang und van Es (2019) als eines der wertvollen Ziele der Lehrkräftebildung beschreiben. Auch in empirischen Untersuchungen konnte bereits ein mediierender Effekt der professionellen Wahrnehmung zwischen dem Wissen einer Lehrkraft und den Lernzuwächsen der Schüler*innen aufgezeigt werden (Blömeke et al., 2022). Ein erfolgreiches Training der professionellen Wahrnehmung setzt allerdings auch voraus, dass notwendiges Wissen auf Seite der angehenden Lehrkräfte vorhanden ist, sodass dies für eine erfolgreiche Entwicklung berücksichtigt werden muss (Martin et al., 2023).

### *2.5.2. Fachspezifik der professionellen Unterrichtswahrnehmung*

Das Training professioneller Unterrichtswahrnehmung kann mit unterschiedlichen Merkmalen der Unterrichtsqualität verknüpft sein. Dabei werden teils generische Merkmale fokussiert (z.B. Stürmer et al., 2013; Hellermann et al., 2015; Kramer et al., 2017), allerdings auch teils fachspezifische Merkmale (z.B. Steffensky et al., 2015; Sunder et al., 2016). Wie auch bei der Beurteilung von Unterrichtsqualität können grundsätzlich fachspezifische Anteile in der professionellen Unterrichtswahrnehmung festgestellt werden (Blomberg et al., 2011; Steffensky et al., 2015). So lässt sich auch eine Korrelation zwischen dem fachdidaktischen Wissen und der professionellen Unterrichtswahrnehmung finden (Meschede et al., 2017). Entsprechend sollten diese auch besonders bei den Voraussetzungen für eine erfolgreiche Entwicklung der professionellen Wahrnehmung berücksichtigt werden. Auch bei generischen Dimensionen der Unterrichtsqualität, wie z.B. der Klassenführung, lassen sich Abweichungen in der professionellen Unterrichtswahrnehmung zwischen unterschiedlichen Fächern finden (Stahnke & Friesen, 2023). Dabei zeigten die Lehrkräfte der Biologie und Mathematik zwar vergleichbare Muster in der Wahrnehmung, wichen jedoch in der Ableitung von Handlungsalternativen voneinander ab.

## *2.6. Ausbildung angehender Lehrkräfte*

Die Ausbildung angehender Lehrkräfte ist in Deutschland nur in einem begrenzten Maß standardisiert und wird in vielen Punkten durch die Bundesländer individuell ausgestaltet. Die Beschlüsse der Kultusministerkonferenz dienen dabei zwar als Orientierung für eine einheitliche Ausgestaltung, sind aber keine bindenden Vorgaben.
Die folgende Übersicht orientiert sich an der Beschreibung von Kunz & Uhl (2021). Eine Gemeinsamkeit und Besonderheit des deutschen Systems ist die Unterteilung in zwei Phasen. Die erste Phase ist dabei stärker theoretisch orientiert und an den Universitäten verortet. Zwar gibt es auch praktische Anteile, aber zunächst sollte in dieser Phase fachliches, fachdidaktisches und pädagogisches Wissen erworben und die Grundlage für einen reflektierten Umgang mit Unterrichtssituationen gelegt werden. In der Regel spezialisieren sich die Studierenden auf zwei bis drei Fächer und eine spezifische Schulform. Die theoretischen Inhalte werden durch Praktika an Schulen bereits teilweise mit konkreten Unterrichtshandlungen verknüpft. Die universitäre Ausbildung wurde an den meisten Standorten auf

das Bachelor-Master-System umgestellt und endet somit in der Regel nach fünf Jahren mit dem Erreichen des Master of Education. Die zweite Phase ist stärker praxisorientiert und begleitet die angehenden Lehrkräfte bei ihrer Arbeit in der Schule. Diese Phase wird auch als Referendariat bezeichnet und ist weit weniger standardisiert als die erste Phase. Grundsätzlich sollen dabei die theoretischen Grundlagen mit konkreten Unterrichtssituationen zusammengeführt werden. Das Referendariat ist jedoch stärker von den jeweiligen Bundesländern abhängig und hat eine Dauer zwischen 16 und 24 Monaten. Teilweise wird es durch sogenannte Praxissemester stärker mit der ersten Phase verknüpft und damit zeitlich entlastet. In der zweiten Phase nehmen die Referendar*innen an regelmäßigen Studienseminaren teil, die fachspezifische und allgemeine pädagogische Inhalte behandeln. Zusätzlich gibt es Unterrichtsbesuche durch die jeweiligen Seminarleitungen, wobei die Anzahl erneut abhängig vom Bundesland ist. Einige Bundesländer legen dabei eine Anzahl pro Fach fest (z.B. zwei Besuche pro Fach und Ausbildungsabschnitt in Berlin), andere geben eine Gesamtzahl für das Referendariat an (z.B. 12 im Verlauf der Ausbildung in Niedersachsen). Abschließend werden nach demselben Schema Prüfungsstunden gezeigt, die für die Benotung der zweiten Phase genutzt werden. Die angehenden Lehrkräfte sind sowohl innerhalb der ersten Phase als auch beim Übergang in die zweite Phase nicht an feste Standorte gebunden. Es ist dadurch nicht ungewöhnlich, dass Referendar*innen von unterschiedlichen Universitäten in einem Fachseminar gemeinsam ausgebildet werden. Die Varianz innerhalb eines Fachseminars wird noch weiter gesteigert, da die Referendar*innen sich teilweise an unterschiedlichen Zeitpunkten in ihrer Ausbildung befinden und unterschiedliche Fächerkombinationen unterrichten. Darüber hinaus gibt es auch die Möglichkeit des Quereinstiegs, wobei kein vorangegangenes Lehramtsstudium notwendig ist, um in die zweite Phase einzusteigen. Die Fachleiter*innen stehen dabei wiederum vor der Herausforderung, die unterschiedlichen Voraussetzungen der Referendar*innen zusammenzubringen. Dabei gibt die eher offene Rahmung zwar einige Freiheiten, sorgt aber auch dafür, dass die Auslegung nach individuellen Maßstäben erfolgen muss.

Der Bedarf nach einer Ausbildung im Bereich der professionellen Wahrnehmung wird auch in der zweiten Phase deutlich. So stellt die Verknüpfung zwischen dem theoretischen Wissen und der Unterrichtspraxis gerade für angehende Lehrkräfte häufig eine Schwierigkeit dar (Korthagen & Kessels, 1999). Außerdem sind angehende Lehrkräfte häufig noch nicht dazu in der Lage, die relevanten Situationen einer Unterrichtsstunde zu fokussieren (Star & Strickland, 2008). Beides soll jedoch im Rahmen der Fachseminare vermittelt werden.

Die offene Rahmung in der Ausbildung angehender Lehrkräfte sorgt auch dafür, dass Unterschiede in der Lehrkräftebildung festgestellt werden können (z.B. Wiernik, 2020) und durch die Referendar*innen eine Standortabhängigkeit der Ausbildung wahrgenommen wird (Döbrich & Storch, 2012). Grundsätzlich können individuelle Schwerpunktsetzungen zielführend sein, um die unterschiedlichen Voraussetzungen der angehenden Lehrkräfte zu berücksichtigen. Allerdings sollte der Anspruch einer Vergleichbarkeit bestehen, wenn die Personen am Ende formal dieselbe Ausbildung erhalten haben und variabel in ihrer Standortwahl sind. Die Offenheit der Rahmung ist auch auf internationaler Ebene zu finden (z.B. Boyd et al., 2009; Tatto, 2021) und kann zu einer Varianz in der Ausprägung der Unterrichtsqualität führen (Blömeke et al., 2016a).

### *2.6.1. Expertise in der Lehrkräftebildung*

Bei der Ausbildung angehender Lehrkräfte stellt sich auch die Frage, inwiefern eine Expertise im Lehrberuf erreicht werden kann und was diese ausmacht. Berliner (1989) stellt unterschiedliche Abstufungen in der Expertise von Lehrkräften heraus. Dabei unterscheiden sich die Stufen zwischen Noviz*innen und Expert*innen zunächst in ihrer Lehrerfahrung, die vereinfacht an der Dauer der Berufsausübung gemessen werden kann. Die Unterschiede zwischen den beiden Extremgruppen äußern sich nach Berliner auch in ihren Unterrichtsbeobachtungen. Noviz*innen fokussieren eher auf Oberflächenstrukturen, wohingegen Expert*innen auch Tiefenstrukturen beurteilen. Dabei ist die Aufmerksamkeit von Noviz*innen wenig fokussiert, sodass sie nicht nur relevante Merkmale beurteilen. Bei Expert*innen ist es wiederum so, dass sie zwar fokussieren und sich nicht von irrelevanten Merkmalen ablenken lassen, jedoch häufig nicht alles explizit äußern können. Vielmehr

gehen sie in ihrem eigenen Unterricht in eine Art „flow“-Erlebnis über, das lediglich durch störende Merkmale unterbrochen wird, sodass positive Merkmale nicht weiter bemerkt werden. Berliner führt dabei jedoch auch aus, dass Expert*innen ihre Expertise nicht unbedingt auf unbekannte Kontexte übertragen können. Damit ist fragwürdig, inwiefern eine erfahrene Lehrkraft überhaupt zu einer professionellen Unterrichtswahrnehmung aus externer Perspektive befähigt ist. Wird die Expertise wiederum in Zusammenhang mit einer professionellen Wahrnehmung gesehen, wie es bei Goodwin (1994) der Fall ist, ergibt sich eine abweichende Definition von Expert*innen. So gibt es zum einen erfahrene Lehrkräfte, die „guten Unterricht“ machen, aber diesen nicht explizieren können und zum anderen gibt es Expert*innen der Unterrichtsbeobachtung. Diese Unterscheidung wird auch weiter durch die Erkenntnisse von Stürmer et al. (2014) unterstützt, nach denen vor allem theoretisches Wissen gegenüber praktischer Erfahrung einen großen Einfluss auf die Zunahme der professionellen Wahrnehmung hat. Es ist vor allem die Fähigkeit, relevante Situationen zu interpretieren und nicht die direkte Wahrnehmung, die vom Wissen beeinflusst wird, wie sich am Beispiel des pädagogischen Wissens zeigen lässt (König et al., 2014; Wolff et al., 2016).

Es zeigt sich jedoch auch, dass Expertise im Bereich der Unterrichtsbeobachtung nicht unbedingt zu reliablen Beurteilungen führt (Praetorius et al., 2012). Expertise hatte bei Praetorius et al. einen Einfluss darauf, welche Merkmale für die Beurteilung genutzt wurden. Expert*innen nutzten insgesamt mehr Merkmale für ihre Beurteilung, die als essenziell für die Unterrichtsqualität eingestuft wurden. Allerdings ließ sich kein Unterschied in der Reliabilität der Beurteilung im Vergleich zu den Noviz*innen finden.

Die Frage nach Expertise ist in der Forschung zur Lehrkräftebildung nicht an einheitliche Voraussetzungen gebunden. Selbst bei Verwendung derselben theoretischen Grundlage werden teilweise unterschiedliche Kriterien zur Identifikation von Expertise verwendet (Palmer et al., 2005). Auch bei der Beschränkung auf einzelne Kriterien, wie z.B. die Lehrerfahrung, haben andere Kriterien einen Einfluss, sodass am Ende individuelle Unterschiede erkennbar sind. Somit ist die Gruppe der Expert*innen in vielen Fällen heterogen und kann als Vergleichsgruppe nicht immer zu eindeutigen Ergebnissen führen (Gabriel-Busse et al., 2020). Statt einer übergreifenden Expertise scheint es daher zielführender, verschiedene Qualitätskriterien für die Beurteilung von Lehrkräften zu nutzen. Somit lässt sich eine übergreifende Qualität der Lehrkraft aus dem Bildungshintergrund der Lehrkraft, der Erfahrung im Lehrberuf, der Teilnahme an Fortbildungen, persönlichen Charakteristika und Selbstwirksamkeit erstellen, die auch einen direkten Zusammenhang zur Unterrichtsqualität aufweisen (Blömeke et al., 2016a).

### *2.6.2. Kommunikation in der Lehrkräftebildung*

Für ein Training der professionellen Unterrichtswahrnehmung ist eine erfolgreiche Kommunikation notwendig, wofür ein *common ground* zwischen den beteiligten Personen gefunden werden muss (Clark & Schaefer, 1989). Ein größeres Problem in der Lehrkräftebildung ist dabei die teilweise Schwierigkeit von erfahrenen Lehrkräften, ihre professionelle Unterrichtswahrnehmung zu explizieren (Berliner, 1989). Dies wird weiter durch die hohe Fülle an Fachbegriffen und die damit einhergehende Abstraktion erschwert. Die Verwendung von Fachbegriffen soll grundsätzlich eine Kommunikation vereinfachen. So ist es in Fachbereichen, wie z.B. der Medizin üblich, Diagnosen über einen Fachbegriff zu bündeln, wobei Expert*innen des Felds diesen erschließen können (Schmidt & Rikers, 2007). Ein Fachbegriff ist an ein bestimmtes Fachwissen gebunden und setzt dieses für ein eindeutiges Verständnis voraus. Die hohe Vielfalt theoretischer Grundlagen und die teilweise abweichende Operationalisierung führt jedoch besonders im Bereich der Unterrichtsqualität dazu, dass Fachbegriffe nicht immer mit eindeutigen Bedeutungen verknüpft werden können (Christ et al., 2022). Weiterhin ist die Expertise nicht genau festgelegt (Abschnitt 2.6.1), sodass diese nicht mit einem klar definierten Fachwissen einhergeht, wie es in der Medizin der Fall ist. Besonders in Fachseminaren kommen eine Vielzahl unterschiedlicher Personen mit abweichenden Hintergründen zusammen. Dies kann zum einen daran liegen, dass sie zuvor an unterschiedlichen Standorten studiert haben, über den Quereinstieg in das Seminar kamen oder einfach an unterschiedlichen Zeitpunkten ihrer Ausbildung sind. Eine Verkapselung durch die Nutzung von Fachbegriffen kann deshalb dazu führen, dass

unterschiedliche Verständnisse bei den Referendar*innen eines Fachseminars ausgelöst werden. Somit kann die Verwendung von Fachsprache die Kommunikation erschweren, statt sie wie gewollt zu erleichtern. Der Lehrberuf ist weniger stark standardisiert als andere Professionen. Zusätzlich sind Forschung und Praxis in vielen Fällen weniger stark vernetzt, was die eindeutige Übertragung theoretischer Konstrukte in die Praxis weiter erschwert. Die Pluralität in der Forschung ist in vielen Fällen gut begründet, kann jedoch direkte negative Auswirkungen auf die Praxis haben. Für ein erfolgreiches Training der professionellen Unterrichtswahrnehmung muss sichergestellt werden, dass auf einer gemeinsamen Grundlage aufgebaut wird (Martin et al., 2023).

## 3. Übergreifende Forschungsfragen

Die Einzelbeiträge der Dissertation verfolgten separate Forschungsfragen. Dennoch lassen sich einige übergreifende Fragen herausstellen, die eine Rahmung bieten und auch für die weiterführende Diskussion relevant bleiben.

1. Was zeichnet Unterrichtsqualität aus Perspektive des Chemieunterrichts aus?

Die Betrachtung von Unterrichtsqualität aus Perspektive des Chemieunterrichts ist zentral für die theoretische Aufbereitung der folgenden Projektteile. Dabei wurde im ersten Beitrag spezifisch die Frage verfolgt, inwiefern das *Syntheseframework* (Praetorius & Charalambous, 2018) auch für den naturwissenschaftlichen Unterricht anwendbar ist. Im Folgenden musste dann überprüft werden, ob das erarbeitete *Framework der naturwissenschaftsdidaktischen Perspektivierungen* (Heinitz & Nehring, 2020) die Möglichkeit bietet, auch andere Ansätze aus den Naturwissenschaften abzubilden (Heinitz et al., 2022). Die Ansätze konnten zwar sinnvoll zusammengeführt werden, dennoch unterschieden sie sich abhängig von ihrem Verwendungszweck. Somit konnten zwar die ersten Forschungsfragen beantwortet werden, aber dennoch blieb ein zentraler Aspekt offen. Lässt sich ein Kern der naturwissenschaftlichen Unterrichtsqualität herausstellen, der als gemeinsame Grundlage für jede Unterrichtsstunde vorhanden sein muss, damit diese zielführend gestaltet ist?
Die theoretische Aufbereitung der naturwissenschaftsspezifischen Unterrichtsqualität erfolgte in den ersten beiden Beiträgen aus Perspektive des aktuellen Forschungsstandes. Es stellte sich weiterführend die Frage, inwiefern die Perspektive der Unterrichtsforschung auch in der zweiten Phase der Lehrkräftebildung wiederzufinden ist. Die Fachseminare stellen eine zentrale Lerngelegenheit in der zweiten Phase dar und verknüpfen die theoretische Grundlage mit konkreten Unterrichtssituationen. Deshalb wurde die Unterrichtsqualitätsbeurteilung von Fachleiter*innen und Referendar*innen in einem offenen Setting untersucht.

2. Inwiefern bestehen Unterschiede in der Beurteilung von Unterrichtsqualität in der zweiten Phase der Lehrkräftebildung?

Der dritte Beitrag stellt eine Abbildung der Praxis dar und vergleicht die beiden Gruppen in ihrer Rolle als externe Beobachtende einer Unterrichtsstunde. Die gruppenspezifischen Unterschiede, die dabei herausgestellt werden konnten, haben dann im Folgenden einen stärkeren Fokus auf die Dimension der kognitiven Aktivierung und damit auf weitgehend fachspezifische Merkmale der Unterrichtsqualität gelegt. Dabei ging es um eine vertiefende Analyse der bestehenden Unterschiede zwischen den beiden Gruppen, aber auch weiterführend um das Anliegen einer einheitlicheren Ausbildung angehender Lehrkräfte. Die zuvor gefundenen Unterschiede boten Grund zu der Annahme, dass die Lehrkräftebildung nicht einheitlich verläuft und teilweise stark durch individuelle Schwerpunkte beeinflusst wird. Der vierte Beitrag behandelte somit die dritte übergreifende Frage.

3. Inwiefern kann die professionelle Unterrichtswahrnehmung im bestehenden System der Lehrkräftebildung einheitlicher entwickelt werden?

Diese Frage wurde zum Teil im Rahmen des vierten Beitrages beantwortet und erste Schwierigkeiten und Möglichkeiten herausgestellt. Im fünften Beitrag wurde die Frage weiterverfolgt und die bisherigen Erkenntnisse zu einem Output für die Praxis zusammengeführt. Dieser stellt natürlich nur einen ersten Ansatz dar, die gesammelten Erkenntnisse in die Praxis zu übertragen, zeigt aber ebenso viele Möglichkeiten für die Weiterarbeit auf. Die übergreifenden Fragen sind sehr weit gefasst und nicht in einem einzelnen Projekt zu beantworten. Dennoch sollen sie den Zusammenhang der Dissertation beleuchten und genutzt werden, um den Leitgedanken zu verdeutlichen.

## 4. Zusammenfassende Darstellung der Methoden

Die Dissertation greift in den einzelnen Beiträgen auf unterschiedliche methodische Ansätze zurück. Dennoch gibt es einige Gemeinsamkeiten, die über die Beiträge hinweg gleich blieben und auf den besonderen Charakter der benötigten Daten zurückzuführen sind. Da viele Abschnitte des Projekts sehr offene Fragen verfolgten, zu denen bisher nur wenige Forschungsgrundlagen bestehen, zeichnet sich das gesamte Projekt durch einen sehr explorativen Charakter aus. Dafür sollte eine möglichst hohe ökologische Validität erreicht werden, die, wie von Holleman et al. (2020) gefordert, für diesen Fall spezifiziert werden soll. Für die Beurteilung der Unterrichtsqualität durch die Fachleiter*innen und Referendar*innen wurde ein Unterrichtsvideo gewählt, um eine vergleichbare Basis zu finden. Es wurde jedoch nicht in die Unterrichtsbeobachtung eingegriffen, um die Situation einer Beurteilung im Referendariat möglichst realistisch nachzustellen. Dafür durfte das Video nicht angehalten werden und es wurden keine Instrumente zur Beobachtung vorgegeben. Sollten individuelle Proband*innen normalerweise feste Beobachtungsinstrumente verwenden, durften sie diese nutzen. Da im Normalfall keine standardisierten Instrumente verwendet werden, wurde im Projekt davon abgesehen, da offene und geschlossene Herangehensweisen unterschiedliche Konstrukte erfassen (Müller & Gold, 2022). Darüber hinaus kann die Übertragung von Instrumenten auf ein anderes Setting zu Verzerrungen führen, gerade wenn die Beurteilenden keine Erfahrung im Umgang mit diesen haben (Schlesinger & Jentsch, 2016). Dadurch sind jedoch auch sehr individuelle Daten erhoben worden, wodurch eine qualitative Auswertung naheliegend war. Dabei hat besonders die qualitative Inhaltsanalyse (Mayring, 2014) den wichtigsten Zugang geboten. Als deduktive Grundlage wurde das zu Beginn entwickelte Framework verwendet, welches wiederum aus einem systematischen Literaturreview (Moher et al., 2009) abgeleitet wurde. Auf dieser breiten Basis ließen sich die meisten Aussagen verorten, wobei das Kodiermanual nach der ersten Studie zusätzlich induktiv durch eine weitere Ebene und Unterteilung der bestehenden Merkmale ergänzt wurde. Das Kodiermanual ist dabei so konzipiert, dass es zur Kategorisierung von Aussagen über Unterrichtsqualität verwendet werden kann. Die Aussagen werden dafür zunächst in „idea units“ (Jacobs & Morita, 2002) unterteilt, wobei die Abschnitte so gewählt werden, dass sie jeweils einer Kategorie zugeordnet werden können. Wichtig war es an dieser Stelle auf Doppelkodierungen zurückzugreifen, da das Manual sehr umfassend ist und die Daten einen hohen Interpretationsspielraum bieten. So war es jederzeit möglich, dass eine zweite Person die „idea units“ anders unterteilen oder auch getrennte Abschnitte wieder zusammenführen konnte. Im vierten Beitrag musste aufgrund der vorangegangenen Forschung davon ausgegangen werden, dass sich die Fachseminare grundlegend voneinander unterscheiden. Um sie dennoch miteinander zu vergleichen, wurden sie als separate Fälle einer mutiple case study (Yin, 2018) betrachtet.

Für die Ableitung von allgemeingültigeren Aussagen, Hypothesen und Implikationen hat es sich angeboten, die durchgeführten Analysen weiter zu quantifizieren. Die genaue Umsetzung ist in den Einzelbeiträgen abgebildet. Auch wenn die Stichproben insgesamt eher klein waren, was vor allem auf die spezifische Untersuchungsgruppe zurückzuführen war, ließen sich damit einige zusammenfassende Einblicke in die zweite Phase der Lehrkräftebildung generieren.

## 5. Fachliche Schwerpunkte der Unterrichtsbeispiele

Im Rahmen des Projekts wurden drei Vignetten aus dem Chemieunterricht verwendet. Grundsätzlich muss bei den offenen Erhebungssettings berücksichtigt werden, dass die Vignetten die Schwerpunkte

bei der Beurteilung beeinflusst haben können. Für einen besseren Einblick sind die Vignetten deshalb in Tabelle 2 als Übersicht dargestellt. Die Vignetten wurden mit unterschiedlichen Klassen gefilmt.
Für die Untersuchung der Entwicklung professioneller Unterrichtswahrnehmung im Fachseminar wurden zwei Vignetten verwendet. Diese wurden so gewählt, dass sie gemeinsame Merkmale der kognitiven Aktivierung adressieren, aber unterschiedliche Fachinhalte behandelten.

*Tabelle 2: Übersicht über die Unterrichtshandlungen und damit verknüpfte fachliche und fachdidaktische Inhalte der verwendeten Unterrichtsvignetten.*

| Vignette zum Vergleich der Fachleiter*innen und Referendar*innen<br>Thema: Umkehrbarkeit der chemischen Reaktion<br>Klassenstufe: 7<br>Dauer: 27 Minuten (Stundendauer 45 Minuten)<br>Fokus: vollständige Stunde mit verkürzter Darstellung der Arbeitsphasen und Phasenübergänge | |
|---|---|
| Unterrichtshandlung | Zentrale fachliche und fachdidaktische Inhalte |
| Unterrichtseinstieg und Kontextualisierung mit einem Comic. Schüler*innen formulieren mit Hilfe der Lehrkraft eine Problemstellung, die durch ein Experiment bearbeitet werden soll. | Silber wurde oxidiert und liegt als (vermeintlich wertloses) Silberoxid vor. Die chemische Reaktion von Silber und Sauerstoff zu Silberoxid soll umgekehrt werden. Die Erstellung von Silberoxid wird an dieser Stelle nicht thematisiert. |
| Die Schüler*innen erhalten von der Lehrkraft eine Versuchsbeschreibung und sollen diese in Gruppen durchführen. Dabei soll Silberoxid in einem Reagenzglas über einem Bunsenbrenner erhitzt und anschließend eine Glimmspanprobe durchgeführt werden. Die Schüler*innen sollen ihre Beobachtungen sammeln. | Die Thermolyse von Silberoxid setzt bei 130 °C ein. Während der Reaktion wird gasförmiger Sauerstoff freigesetzt, der durch eine Glimmspanprobe nachgewiesen werden kann. Nach dem Erhitzen verbleibt im Reagenzglas Silber als Feststoff.<br>$2Ag_2O_{(s)} \xrightarrow{\Delta T} 4Ag_{(s)} + O_{2(g)}$ |
| Sammlung der Beobachtungen an der Tafel und Deutung des Versuchs im Plenum. Die Lehrkraft moderiert die Aussagen der Schüler*innen. Die Beobachtungen werden einzeln aufgegriffen und den Deutungen gegenübergestellt. Abschließend wird die Fragestellung vom Beginn der Stunde aufgegriffen. | Schüler*innen durchlaufen exemplarisch einen naturwissenschaftlichen Arbeitsprozess. Die einzelnen Schritte sind dabei: Problemstellung, Hypothese, Experiment, Beobachtungen, Deutung und Hypothesenprüfung. |
| Vignette zur Untersuchung der Beurteilung kognitiver Aktivierung im Fachseminar (Lernvignette)<br>Thema: Einführung der chemischen Reaktion<br>Klassenstufe: 7<br>Dauer: 6 Minuten (Stundendauer 90 Minuten)<br>Fokus: Auswertung eines durchgeführten Experiments | |
| Unterrichtshandlung | Zentrale fachliche und fachdidaktische Inhalte |
| Vor dem Start der Vignette wurde ein Experiment durchgeführt, bei dem die Schüler*innen Schwefel und Kupfer in einem Reagenzglas über einem Bunsenbrenner erhitzt haben. Die Schüler*innen sollten die Eigenschaften der Edukte und Produkte notieren und miteinander vergleichen. | Zu Beginn des Experiments liegen Schwefel und Kupfer als Feststoffe im Reagenzglas vor. Beim Erhitzen verdampft der Schwefel und reagiert mit dem Kupfer zu Kupfersulfid.<br>$Cu_{(s)} + S_{(s)} \xrightarrow{\Delta T} CuS_{(s)}$<br>Kupfer liegt vor dem Experiment als rötliches, biegsames Blech vor. Schwefel liegt als gelbliches Pulver vor. Das entstandene Kupfersulfid ist schwarz-bläulich und brüchig. |
| Besprechung des Experiments im Plenum | Die Schüler*innen durchlaufen exemplarisch einen naturwissenschaftlichen Arbeitsprozess. |

| | |
|---|---|
| | Beobachtungen vor und nach dem Experiment werden gegenübergestellt und daraus eine Deutung zum Ergebnis des Experiments angestellt. |
| Vergleich mit einem Modellexperiment der vorangegangenen Stunde. | Beim Modellexperiment wurde ein Gemisch aus Erbsen und Hirse hergestellt und die Volumenkontraktion thematisiert. Die Schüler*innen sollten diesen Vergleich als Anlass nehmen, die chemische Reaktion als Entstehung neuer Stoffe mit neuen Eigenschaften zu deuten. Durch die thematische Bindung des vorangegangenen Experiments zum Teilchenmodell deuten sie jedoch die Reaktion von Kupfer und Schwefel mit dem Teilchenmodell. Dabei werden Hybridvorstellungen zwischen dem Teilchenmodell und dem Atommodell nach Dalton deutlich. |
| Vignette zur Untersuchung der Beurteilung kognitiver Aktivierung im Fachseminar (Transfervignette)<br>Thema: Modelle als Teil naturwissenschaftlicher Denk- und Arbeitsweisen<br>Klassenstufe: 8<br>Dauer: 3 Minuten (Stundendauer 45 Minuten)<br>Fokus: Stundeneinstieg | |
| Unterrichtshandlung | Zentrale fachliche und fachdidaktische Inhalte |
| Einstieg mit einem stummen Impuls unter Verwendung einer schematischen Darstellung. | Die Darstellung basiert auf dem einfachen Teilchenmodell und stellt die Aggregatzustände in drei separaten Abbildungen dar. |
| Interpretation der Abbildung durch die Schüler*innen mit Rückfragen durch die Lehrkraft. | Die Rückfragen sollen den Modellcharakter der Abbildung herausstellen und die Schüler*innen dazu bringen, diese im Plenum zu explizieren. |
| Beginn der ersten Einzelarbeitsphase mit einer zentralen Aufgabenstellung für alle Schüler*innen. | Die Schüler*innen sollen ihr gewohntes Vorgehen aus einer Metaperspektive betrachten und damit Modelle als Teil naturwissenschaftlicher Denk- und Arbeitsweisen herausstellen. |

## 6. Vernetzung der Einzelbeiträge

Die Dissertation ist kumulativ ausgelegt, sodass die jeweiligen Einzelbeiträge aufeinander aufbauen und sich die Erkenntnisse eines Projektteils auf die Planung und Durchführung des nächsten Teils ausgewirkt haben. Insgesamt sind es somit fünf Einzelbeiträge, die untereinander vernetzt sind, sich aber gleichzeitig entlang der übergreifenden Forschungsfragen in drei Abschnitte unterteilen lassen. Die Beiträge eins und zwei bilden die theoretische Grundlage der Unterrichtsqualität im naturwissenschaftlichen Unterricht ab. Die Beiträge drei und vier werfen einen vertiefenden Blick auf den aktuellen Stand der Praxis, wobei Beitrag vier zusammen mit Beitrag fünf auch Implikationen für eine mögliche Entwicklung der professionellen Unterrichtswahrnehmung in der zweiten Phase behandelt.

Im Folgenden werden jeweils die Ausgangslagen der jeweiligen Beiträge beschrieben und die zentralen Inhalte in einer kurzen Zusammenfassung dargestellt. Daraus ergeben sich wiederum weiterführende Punkte, die teilweise direkt in Folgebeiträgen aufgegriffen wurden und diese somit vernetzen. Einige dieser Punkte konnten bislang noch nicht abschließend bearbeitet werden und bilden somit wichtige

Diskussionspunkte in der abschließenden Betrachtung der Dissertation. Alle Beiträge sind als Anhang beigefügt.

## *6.1. Beitrag 1: Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung*

### *6.1.1. Ausgangslage*

Die Ausgangslage des ersten Beitrags wurde durch das *Syntheseframework* (Praetorius & Charalambous, 2018) geprägt. Das *Syntheseframework* hat den Anspruch, einen breiteren Ansatz zur Systematisierung von Unterrichtsqualität zu bieten und die drei Basisdimensionen um bisher fehlende Aspekte zu ergänzen. Dabei ist es aus generischer Perspektive heraus formuliert, beinhaltet jedoch ebenso fachspezifische und hybride Merkmale. Bisher stand eine direkte Anwendung im naturwissenschaftlichen Fachbereich aus, weshalb eine Gegenüberstellung des *Syntheseframeworks* mit der Unterrichtsqualität aus Sicht der Naturwissenschaften angestrebt wurde. Dazu musste zunächst die Frage geklärt werden, welche Qualitätsmerkmale für den naturwissenschaftlichen Unterricht relevant sind. Hierzu wurde die breite empirische Basis der Naturwissenschaftsdidaktiken genutzt und in einem Literaturreview Merkmale herausgearbeitet, die in vorangegangenen Videostudien zur Unterrichtsqualitätsbeurteilung genutzt worden sind. Ausgangspunkt für die Recherche bildete ein bestehendes Review zur methodischen und inhaltlichen Ausrichtung quantitativer Videostudien im mathematisch-naturwissenschaftlichen Fachbereich (Dorfner et al., 2017). Die fachliche Rahmung wurde dabei auf Studien aus dem naturwissenschaftlichen Bereich beschränkt (Biologie, Chemie, Physik und Sachunterricht).

### *6.1.2. Zusammenfassung der Ergebnisse*

Das durchgeführte Literaturreview baute auf dem Suchraster des Reviews von Dorfner et al. (2017) auf. Für eine Abbildung des aktuellen Standes wurden zusätzlich die Jahre 2016 – 2019 ergänzt. Dabei unterschieden sich die Studien in ihren theoretischen Grundlagen und auch den Zielen, die sie mit der Erhebung verfolgten. Somit mussten die bestehenden Instrumente zunächst zusammengeführt werden, um die Frage nach einer naturwissenschaftsspezifischen Unterrichtsqualität zu beantworten. Alle Kriterien aus den gefundenen Studien wurden gesammelt und dem *Syntheseframework* gegenübergestellt.

Insgesamt ließen sich 72 % der Kriterien mit einer „vollständigen Übereinstimmung“ im *Syntheseframework* verorten. In einigen Fällen mussten jedoch fachspezifische Formulierungen stärker abstrahiert werden. Darüber hinaus gab es einige Kriterien, die nicht direkt verortet werden konnten und 21 % der Kriterien wiesen eine „teilweise Übereinstimmung“ auf. Dies war zum einen darin begründet, dass die Kriterien mit einer breiteren Grundlage aufgebaut waren und sich über mehrere Abschnitte des *Syntheseframeworks* erstreckt haben. Zum anderen passten Kriterien nur anteilig in das *Syntheseframework*, wenn Aspekte des Kriteriums unberücksichtigt blieben. Bei 7 % der Kriterien gab es eine „starke Abweichung“, da sie vollständig im *Syntheseframework* fehlten und somit eine Erweiterung der bestehenden Systematik notwendig machten. Aus der Analyse der Gemeinsamkeiten und Unterschiede ließ sich insgesamt eine naturwissenschaftsspezifische Adaption des generischen *Syntheseframeworks* ableiten. Dazu wurden die verorteten Kriterien zu Clustern zusammengeführt und fehlende Aspekte im *Syntheseframework* ergänzt. Dadurch entstand das *Framework der naturwissenschaftsdidaktischen Perspektivierungen*. Das adaptierte Framework nutzt dieselben sieben generischen Dimensionen und ordnet darunter 50 Merkmale der Unterrichtsqualität an, die aus der Clusterung der Kriterien hervorgegangen sind.

### *6.1.3. Vernetzung der Ergebnisse*

Das *Framework der naturwissenschaftsdidaktischen Perspektivierungen* stellt eine Adaption des *Syntheseframeworks* dar. Dabei wurden an vielen Stellen fachspezifische Merkmale ergänzt, was bedeutsam für eine potenzielle Unterrichtsqualitätsbeurteilung im Chemieunterricht ist. Gerade fachspezifische Merkmale haben einen großen Einfluss auf die Lernzuwächse der Schüler*innen (Seidel & Shavelson, 2007) und können essenziell für eine vollständige Beurteilung einer Unterrichtsstunde

sein (Brunner, 2018). Dies macht jedoch auch die Notwendigkeit von fachlichem und fachdidaktischem Wissen bei der Anwendung des adaptierten Frameworks deutlich (Neuhaus, 2021).
Grundlegend bleibt es jedoch über den direkten Bezug und die Verortung der einzelnen Kriterien vergleichbar zur generischen Basis. Genau über diese Basis lässt sich das naturwissenschaftsdidaktische Framework wiederum auch mit anderen Fächern vergleichen (Praetorius et al., 2020b). Es wurde somit nicht nur ein neuer Ansatz zur Beurteilung der Unterrichtsqualität geschaffen, sondern explizit auf einer bestehenden Systematik aufgebaut. Damit wurde die Bestrebung verfolgt, gemeinsame Grundlagen zu nutzen und kollaborativ an einer stetigen Zusammenführung von Verständnissen der Unterrichtsqualität zu arbeiten (z.B. Charalambous et al., 2021). Die Zusammenführung unterschiedlicher Ansätze der Unterrichtsqualität und die Erarbeitung einer gemeinsamen Basis wird deshalb auch als ein zentraler Punkt in der weiterführenden Diskussion aufgegriffen.
Die Kriterien, die zur Operationalisierung des adaptierten Frameworks verwendet wurden, sind aus bereits durchgeführten empirischen Studien der Naturwissenschaftsdidaktiken entnommen und transparent im Framework verortet. Dadurch sind die erstellten Cluster der *naturwissenschaftsdidaktischen Perspektivierungen* weniger abstrakt und leichter nachvollziehbar. Es bietet sich dadurch weniger Interpretationsspielraum, als die Cluster allein durch ihre Terminologie zulassen würden. Hierbei muss jedoch auch berücksichtigt werden, dass die Verortung der Kriterien nur einseitig erfolgte. Für einen vollständigen Abgleich und eine eindeutige Verortung müssten auch die Autor*innen der jeweiligen Instrumente eine Verortung ihrer Kriterien im *Syntheseframework* vornehmen. Dies ist natürlich durch die Fülle der Studien, die im Rahmen des Reviews gefunden wurden, nicht realistisch, wäre aber nur konsequent, wenn die Ergebnisse des zweiten Beitrags berücksichtigt werden.

## *6.2. Beitrag 2: Unterrichtsqualität in den Naturwissenschaften. Eine vergleichende Gegenüberstellung von Ansätzen zwischen Fachspezifik und Generik*

### *6.2.1. Ausgangslage*

Das *Framework der naturwissenschaftsdidaktischen Perspektivierungen* ist durch eine Zusammenführung von Instrumenten mit unterschiedlichen theoretischen Grundlagen entstanden und führt diese vor dem Hintergrund des *Syntheseframeworks* zusammen. Auf dieser gemeinsamen theoretischen Grundlage war eine Gegenüberstellung mit Ansätzen anderer Fachbereiche (Sport und Geschichte) möglich (Praetorius et al., 2020b).
Die Instrumente aus den Naturwissenschaftsdidaktiken wurden im ersten Beitrag auf einer neuen theoretischen Grundlage zusammengeführt. Es ist jedoch offen, inwiefern sich andere Instrumente der Naturwissenschaftsdidaktiken mit dem adaptierten Framework vergleichen lassen oder es ggf. erweitern. Zum einen könnte das adaptierte Framework durch die breite theoretische Grundlage in der Lage sein, auch weitere Ansätze abzubilden. Es könnte jedoch auch möglich sein, dass die zugrundeliegenden Theorien zu stark voneinander abweichen und es nicht möglich ist, die Ansätze gemeinsam zu betrachten.
Wenn die Grundlagen es entsprechend zulassen, könnten unterschiedliche Ansätze künftig im selben Framework verortet und bei Bedarf modular zusammengesetzt werden. Damit wären sie untereinander vergleichbar und es wäre dennoch möglich, ein Instrument an einen spezifischen Untersuchungszweck anzupassen. Aus der Theorie heraus wird schnell ersichtlich, dass es viele unterschiedliche Ansätze gibt, die sich wiederum auf unterschiedliche generische Grundlagen berufen. Durch das adaptierte Framework wird eine Systematisierung vorgeschlagen, die Merkmale aus allen Fächern der Naturwissenschaftsdidaktiken enthält. Offen war dabei, ob es alle Konzepte so umfänglich abbilden kann, wie sie aus den jeweiligen Theorien heraus konzipiert sind.

### *6.2.2. Zusammenfassung der Ergebnisse*

Im Rahmen des Beitrags wurde eine Gegenüberstellung von Ansätzen zur Erfassung der Unterrichtsqualität aus den Fächern Biologie und Physik mit dem *Framework der naturwissenschaftsdidaktischen Perspektivierungen* durchgeführt. Diese Gegenüberstellung folgte demselben Vorgehen, wie es beim *Syntheseframework* (Praetorius & Charalambous, 2018) angewandt

wurde. Dabei war das Ziel in diesem Fall nicht die Erstellung eines gemeinsamen Ansatzes, sondern ein systematisches Herausstellen von Gemeinsamkeiten und Unterschieden. Das Vorgehen orientiert sich deshalb an denselben Fragen wie das *Syntheseframework* und einem schrittweisen Vergleich von theoretischen Grundlagen, Verwendungszweck und Operationalisierung. Die konkreten Fragen, die für alle Ansätze beantwortet werden mussten, wurden entsprechend angepasst.
Zusätzlich erfolgte eine gegenseitige Verortung der Ansätze in den jeweiligen Systematisierungen der anderen. Jede Teilgruppe der Fachbereiche bekam dabei die Aufgabe, ihren eigenen Ansatz aufzuteilen und der Systematik der anderen unterzuordnen. Damit wurde insgesamt kein Ansatz den anderen vollständig übergeordnet, sondern ein Vergleich aus unterschiedlichen Perspektiven erstellt. Dieses umfangreiche Verfahren wurde genutzt, um sicherzustellen, dass die Gegenüberstellung der Operationalisierungen auch von allen Autor*innen gleich verstanden wurde. Dabei ließen sich fünf generalisierte Aussagen über die Vergleichbarkeit von Ansätzen zur Beurteilung der Unterrichtsqualität in den Naturwissenschaftsdidaktiken ableiten, die auch für künftige Studien relevant sein können.
Im Rahmen der theoretischen Gegenüberstellung wurde außerdem die Frage nach einer Definition von Fachspezifität aufgeworfen. Dabei stellte sich heraus, dass die drei Ansätze jeweils eine leicht abweichende Definition von Fachspezifik nutzen, was einen Einfluss auf die gegenseitige Verortung hatte. Insgesamt konnten fünf mögliche Ansätze zur Definition von Fachspezifik abgeleitet werden. Diese beeinflussen nicht nur eine mögliche Zusammenführung unterschiedlicher Instrumente, sondern auch die Kommunikation im Bereich der Unterrichtsqualität. Häufig wird von fachspezifischen Ansätzen, Instrumenten oder sogar Kriterien gesprochen. Was genau damit gemeint ist, ist jedoch abhängig vom zugrundeliegenden theoretischen Verständnis. Dabei erstreckten sich diese abweichenden Definitionen nicht nur auf Ansätze aus den Naturwissenschaftsdidaktiken, sondern auch auf generische Ansätze.
Zusammenfassend ließ sich zeigen, dass die Ansätze mit einigen Ausnahmen sehr gut in den jeweiligen Systematisierungen der anderen Ansätze verortet werden konnten. Das *Framework der naturwissenschaftsdidaktischen Perspektivierungen* bot dabei eine Rahmung, in der die anderen Ansätze vollständig verortet werden konnten. An einigen Stellen haben Abweichungen in der theoretischen Grundlage und dem Verwendungszweck jedoch dazu geführt, dass die Operationalisierung innerhalb des Frameworks aufgeteilt werden mussten und somit z.B. durch ein einzelnes Merkmal mehrere Dimensionen der Unterrichtsqualität im *Framework der naturwissenschaftsdidaktischen Perspektivierungen* adressiert wurden. Dies unterstreicht noch einmal die Möglichkeit einer modularen Zusammensetzung von Instrumenten aus bestehenden, umfangreichen Frameworks.

### *6.2.3. Vernetzung der Ergebnisse*

Grundsätzlich hat sich gezeigt, dass das *Framework der naturwissenschaftsdidaktischen Perspektivierungen* so umfassend konzipiert ist, dass auch andere Ansätze der Naturwissenschaftsdidaktiken darin verortet werden können. Unterschiede in der theoretischen Grundlage und den Verwendungszwecken führten teilweise dazu, dass die Operationalisierung abweicht. Dennoch musste das Framework nicht durch weitere Merkmale ergänzt werden. Das bedeutet nicht, dass das Framework entsprechend für jede Beurteilung der Unterrichtsqualität in den Naturwissenschaften geeignet ist. Vielmehr zeigt es, dass darüber unterschiedliche Ansätze zusammengeführt werden können. Das systematische Vorgehen und die generalisierten Unterschiede und Gemeinsamkeiten sollten besonders bei der Entwicklung neuer Ansätze berücksichtigt werden. Wenn diese direkt mit Bezug zu einem übergreifenden Framework erstellt werden, bleiben sie auch weiterführend vergleichbar mit anderen Ansätzen. Damit könnten die Ergebnisse einer potenziell sehr spezifischen Studie vor einem breiteren Hintergrund betrachtet werden. Das Ziel ist dabei nicht, einen einzigen Ansatz zur Beurteilung der Unterrichtsqualität zu entwickeln, sondern eine gemeinsame Weiterentwicklung anzustreben.
Auch wenn es zunächst nicht im Fokus der Untersuchung stand, hat die Herausarbeitung von unterschiedlichen Verständnissen der Fachspezifik weitreichende Auswirkungen. Wenn in der Forschung z.B. vom Einfluss fachspezifischer Merkmale auf den Lernerfolg von Schüler*innen

gesprochen wird (z.B. Seidel & Shavelson, 2007), löst dies je nach vorliegendem Verständnis unterschiedliche Konnotationen aus. Gerade wenn ein komplexes Ergebnis einer Meta-Studie auf eine zentrale Aussage heruntergebrochen wird, sollte diese zumindest von allen gleich verstanden werden. Die unterschiedlichen Definitionen von Fachspezifik verdeutlichen noch einmal das grundlegende Problem der Kommunikation, was mit der Verwendung abstrakter Terminologien einhergeht und auch in der weiterführenden Diskussion (Abschnitt 7.2.3) breiter behandelt wird. Dies stellt ein Problem dar, das auch über den Bereich der Unterrichtsqualität hinausgeht.

## *6.3. Beitrag 3: Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors*

### *6.3.1. Ausgangslage*

Die theoretische Betrachtung der Unterrichtsqualität hat viele Punkte aufgeworfen, die auch für die Praxis der Lehrkräftebildung relevant sind. Für die Beurteilung von Unterrichtsqualität gibt es in der Forschung eine Pluralität unterschiedlicher Ansätze (Helmke & Schrader, 2008). Diese werden wiederum teilweise in der Lehrkräftebildung genutzt, allerdings gibt es dabei keine übergreifenden oder systematischen Vorgaben. Auf Länderebene werden zwar Vorgaben definiert, allerdings sind diese eher offen gehalten, sodass Raum für individuelle Interpretationen besteht (Kunz & Uhl, 2021). So entsteht bei angehenden Lehrkräften häufig der Eindruck einer unterschiedlichen Behandlung in Abhängigkeit von den jeweils Ausbildenden (Döbrich & Storch, 2012). Dabei spielen die offenen Richtlinien für die Ausbildenden genauso eine Rolle wie die Schwierigkeit, bisher gelernte theoretische Ansätze in die Praxis zu übertragen, welche häufig bei angehenden Lehrkräften ein Problem darstellt (Korthagen & Kessels, 1999). Die beiden Phasen der Lehrkräftebildung sollten grundsätzlich vernetzt sein, sodass das theoretische Wissen der ersten Phase direkt mit den Praxisbeispielen der zweiten Phase verknüpft werden kann. Diese Vernetzung findet aber häufig nicht in ausreichendem Maße statt (Weber & Czerwenka, 2021).

Dadurch hat sich für die folgende Untersuchung die Frage ergeben, inwiefern sich die Beurteilungen von Unterrichtsqualität durch Fachleiter*innen unterscheiden und weiterhin, inwiefern es Unterschiede zur Gruppe der Referendar*innen gibt. Aus den theoretischen Ansätzen zur Expertise in der Lehrkräftebildung ist ein Unterschied in der Beurteilung zwischen Expert*innen und Noviz*innen (gemessen an der Unterrichtserfahrung) zu erwarten (Berliner, 1989). Bei der Untersuchung von Berliner (1989) hat sich jedoch auch gezeigt, dass die Expertise bei der Beurteilung einer fremden Klasse aus einer externen Perspektive weniger stark zum Tragen kommt. Diese einfache Unterteilung wird weiter erschwert, wenn Expertise als Zusammenspiel unterschiedlicher Kriterien betrachtet wird (Blömeke et al., 2016a). Gerade da es keine festen Vorgaben für Fachleiter*innen gibt (Weber & Czerwenka, 2021), konnte zwar ein gewisser Gruppenunterschied erwartet werden, allerdings ist die Ausprägung aufgrund der heterogenen Gruppe nicht eindeutig absehbar gewesen. Weiterhin besteht eine Besonderheit des Systems gerade darin, dass die Beurteilenden häufig keine standardisierten Instrumente für die Beurteilung einer Unterrichtsstunde verwenden. Ein solcher offener Ansatz kann zu anderen Ergebnissen führen als ein geschlossener Ansatz (Müller & Gold, 2022). Diese Besonderheit wurde deshalb für die Studie berücksichtigt und ein explorativer, offener Ansatz für eine möglichst hohe ökologische Validität gewählt.

### *6.3.2. Zusammenfassung der Ergebnisse*

Der Gruppenvergleich zwischen Fachleiter*innen und Referendar*innen basierte auf der freien Beurteilung einer aufgezeichneten Chemieunterrichtsstunde (vgl. Tab. 2). In dieser wurde die Umkehrbarkeit der chemischen Reaktion thematisiert und in einem Experiment Silberoxid erhitzt. In der Studie wurde untersucht, welche Merkmale durch die jeweiligen Gruppen verwendet werden wurden und inwiefern sich die Bewertungen derselben Merkmale unterscheiden. Abschließend sollte die Unterrichtsstunde benotet werden.

Zur Auswertung wurde ein Kodiermanual aus dem *Framework der naturwissenschaftsdidaktischen Perspektivierung* entwickelt, welches eine möglichst umfangreiche Kategorisierung von Aussagen zur Unterrichtsqualität ermöglichen sollte. Es hat sich herausgestellt, dass die beiden Gruppen

grundsätzlich voneinander unterschieden werden können und Gruppenspezifika aufweisen. So haben Fachleiter*innen im Durchschnitt mehr unterschiedliche Merkmale der Unterrichtsqualität verwendet, um die Unterrichtsstunde zu beurteilen. Dabei sind sie in der Bewertung der Merkmale tendenziell negativer eingestellt als die Referendar*innen. Dies spiegelt sich auch in der abschließenden Benotung wider. Bei der Auswahl der Merkmale zeigen die Fachleiter*innen einen stärkeren Fokus auf die kognitive Aktivierung als die Referendar*innen. Diese nutzen tendenziell eher Merkmale der Oberflächenstruktur und beziehen sich auf die Klassenführung. Dabei verwenden sie im Durchschnitt weniger Merkmale pro Person. Werden die Beurteilungen der Fachleiter*innen zusammen betrachtet, lassen sich Referenzen zu 47 von 50 Merkmalen des Frameworks finden. Dieses Ergebnis verdeutlicht, dass die Unterrichtsstunde das Potenzial bietet, fast jeden Aspekt der Unterrichtsqualität zu thematisieren. Hierbei muss jedoch berücksichtigt werden, dass teilweise lediglich das Potenzial für bestimmte Merkmale genannt wurde, ohne dass diese direkt beobachtbar waren. Insgesamt weist die Benotung der Unterrichtsstunde eine sehr breite Streuung auf, welche das gesamte Notenspektrum von „sehr gut“ bis „ungenügend“ umfasst. Eine Streuung ließ sich in beiden Gruppen separat beobachten und ist nicht allein auf die Gruppenunterschiede zurückzuführen. Besonders in der Gruppe der Fachleiter*innen hätte die Benotung dazu geführt, dass die Referendarin in der Unterrichtsaufzeichnung in 7 von 17 Fällen durchgefallen wäre, obwohl alle dieselbe Aufzeichnung gesehen haben.
Der Beitrag enthält außerdem die englische Übersetzung des *Frameworks der* naturwissenschaftsdidaktischen Perspektivierungen: *Sience Education Perspectives-Framework*, wodurch sich im Folgenden das Akronym *SEP-Framework* ableitet.

### *6.3.3. Vernetzung der Ergebnisse*

Das Kodiermanual zum *SEP-Framework* wurde mit dem Ziel entwickelt, Aussagen zur Unterrichtsqualität in den 50 Merkmalen des Frameworks zu verorten. Bei der Auswertung konnten jedoch einige Merkmale herausgestellt werden, die einen verhältnismäßig breiten Interpretationsrahmen haben. Aus der Adaption des *Syntheseframeworks* ging dies zunächst nicht hervor. In der Auswertung führte dies jedoch dazu, dass z.B. Aussagen zur „Passung der Lerninhalte an die Lernvoraussetzungen“ der Schüler*innen zusammen mit Aussagen zur „Einbindung der Fachinhalte durch Gesprächsführung“ geclustert wurden. Auf der theoretischen Ebene passt beides zur Auswahl und Einbindung von Fachinhalten, für eine präzisere Analyse der Beurteilung könnte dies aber zu ungenau sein. Dies gab Anlass zu einer induktiven Erweiterung um sogenannte Subperspektivierungen, die im Rahmen einer Masterarbeit erstellt wurden (Meyer, 2021). Das daraus resultierende Kodiermanual erlaubt eine präzisere Verortung von Aussagen der Unterrichtsqualität und wurde im Folgenden für alle weiteren Auswertungen genutzt. Dies schloss ebenfalls die Interviews der Referendar*innen ein. Diese induktive Erweiterung hat zwei Punkte verdeutlicht. Zum einen ist eine weitere induktive Ausdifferenzierung des *SEP-Frameworks* möglich, die über die bisherigen Subperspektivierungen hinausgeht. Damit könnten Indikatoren abgeleitet werden, die direkt zu einer Beurteilung der Unterrichtsqualität genutzt werden könnten. Dabei hätte das gesamte Framework zwar einen zu großen Rahmen, aber es könnte abhängig von der Untersuchung innerhalb desselben Frameworks eine Auswahl an Merkmalen für eine Beurteilung gewählt werden. Zum anderen könnte der große Umfang des Kodiermanuals aber auch einen Anlass bieten, um über eine stärker automatisierte Kodierung von Aussagen zur Unterrichtsqualität nachzudenken (Abschnitt 7.4). Die Ergebnisse haben auch gezeigt, dass die Bewertung der Merkmale eine starke Korrelation zur Benotung aufgezeigt hat. Das zeigt, dass die Proband*innen trotz individueller Unterschiede in ihrer Beurteilung intraindividuell konsistent sind. Es zeigt aber auch, dass die gebildeten mittleren Bewertungen auf Basis des Kodiermanuals prädiktiv für die Benotung der Stunde sind. Diese Punkte werden deshalb in der weiterführenden Diskussion ebenfalls aufgegriffen.
Die Interviews haben gezeigt, dass eine sehr unterschiedliche Beurteilung der Unterrichtsqualität vorliegt. Dieses Ergebnis wirft viele Fragen über den aktuellen Stand der Lehrkräftebildung auf und wird deshalb ebenfalls in der weiterführenden Diskussion aufgegriffen. Hierbei schließen sich nicht nur

Fragen über eine mögliche Standardisierung der Lehrkräftebildung an, sondern auch über die Expertise bei der Beurteilung von Unterrichtsqualität.
Es scheint sinnvoll, die Lehrkräftebildung an gemeinsamen theoretischen Grundlagen zu orientieren. Inwiefern eine theoretische Vorgabe in der Praxis von allen direkt umgesetzt werden kann, ist jedoch offen. Gerade wenn die Vorgabe nicht direkt mit den bestehenden theoretischen Grundlagen vereinbar ist, kann es zu Problemen führen. Die anschließende Studie zur Entwicklung der professionellen Unterrichtswahrnehmung in den Fachseminaren greift unter anderem diese Problematik auf.

## *6.4. Beitrag 4: Perception and Interpretation of Cognitive Activation in Science Pre-Service Teacher Education*

### *6.4.1. Ausgangslage*

Bei der Analyse der freien Beurteilung im dritten Beitrag ist vor allem die Dimension der kognitiven Aktivierung durch große Gruppenunterschiede aufgefallen. Grundsätzlich war festzuhalten, dass die Fachleiter*innen häufiger Bezüge zu dieser Dimension herstellen. Damit lag die Annahme nahe, dass ein gewisser Grad von Unterrichtserfahrung zur Beurteilung von Merkmalen dieser Dimension notwendig ist, was sich auch in anderen Studien bestätigt (z.B. Meschede et al., 2017). Kognitive Aktivierung ist durch den hohen Interpretationsspielraum nur schwer zugänglich und reliable zu beurteilen (Praetorius et al., 2012). Dennoch weisen Studien darauf hin, dass ein Zusammenhang zwischen einer hohen kognitiven Aktivierung der Schüler*innen und deren Lernerträgen steht (Praetorius et al., 2018). Kognitive Aktivierung wird weiterhin als Teil der Tiefenstruktur von Unterricht aufgeführt und als grundlegend relevant für die Lernerträge der Schüler*innen herausgestellt (Kunter & Ewald, 2016). Hierbei muss jedoch beachtet werden, dass kognitive Aktivierung häufig unterschiedlich operationalisiert wird (Christ et al., 2022). Dies kann die Zusammenhänge zu den Lernerträgen der Schüler*innen sowohl positiv als auch negativ beeinflussen.
Aus dem vorangegangenen Beitrag ließ sich das Ziel ableiten, die Beurteilung von Unterrichtsqualität einheitlicher in die Lehrkräftebildung zu integrieren. Der Grundgedanke war hierbei, die professionelle Unterrichtswahrnehmung der Referendar*innen im Sinne eines *knowledge-based reasoning* (Sherin, 2001) zu trainieren, sodass sie ebenfalls in der Lage sind, bestimmte Merkmale der kognitiven Aktivierung zu erkennen. Videovignetten haben sich in der ersten Phase bereits als wichtiges Werkzeug zum Training der professionellen Unterrichtswahrnehmung etabliert (Blomberg et al., 2013). Offen ist hierbei allerdings, inwiefern dies auch in den bestehenden Strukturen der Fachseminare in der zweiten Phase möglich ist. Bisher gibt es nur wenige Untersuchungen zum Einsatz von Videovignetten in der zweiten Phase der Lehrkräftebildung.
Das Ziel der Studie war die Untersuchung von Veränderungen in der professionellen Unterrichtswahrnehmung von Referendar*innen, wenn diese an einem vignettenbasierten Fachseminar zur kognitiven Aktivierung teilnahmen. Dies sollte erneut mit einer möglichst hohen ökologischen Validität geschehen, um das bestehende System möglichst realistisch abzubilden. Damit sollten auch Schritte für die Entwicklung der professionellen Unterrichtswahrnehmung im Referendariat möglichst nah am bestehenden System abgeleitet werden. Die Vignetten wurden so ausgewählt, dass die kognitive Aktivierung mit den gleichen Merkmalen des *SEP-Frameworks* beurteilt werden kann. Inhaltlich behandeln sie jedoch unterschiedliche Inhalte, sodass ein Transfer nicht direkt an dieselben Unterrichtsbeobachtungen gebunden ist, sondern zunächst abstrahiert werden muss (vgl. Tab. 2). Grundsätzlich muss jedoch bei einer Beurteilung der kognitiven Aktivierung darauf geachtet werden, dass Merkmale derselben nur wenig zeitlich stabil sind und somit nur unzuverlässig in kurzen Unterrichtsausschnitten beurteilt werden können (Begrich et al., 2021). Damit stellt diese Dimension sowohl erfahrene als auch unerfahrene Beobachter*innen vor eine große Herausforderung.

### *6.4.2. Zusammenfassung der Ergebnisse*

Ähnlich zur vorangegangenen Untersuchung wurde ein offenes und exploratives Vorgehen gewählt. Die Operationalisierung der kognitiven Aktivierung basierte auf dem *SEP-Framework* und dem dazugehörigen Kodiermanual. Um die Entwicklung der professionellen Wahrnehmung innerhalb einer Fachseminarsitzung zu untersuchen, wurde die kognitive Aktivierung in zwei Vignetten in einem Prä-

Post-Design durch die teilnehmenden Referendar*innen beurteilt (Tab. 2). Dazwischen fand eine Seminarsitzung statt, die durch die Fachleiter*innen gehalten wurde. Diese war an einige Vorgaben gebunden: Es wurde eine Vignette und vier Merkmale der kognitiven Aktivierung vorgegeben, die thematisiert werden sollten.
Es hat sich gezeigt, dass die Referendar*innen in der Post-Erhebung einen Zuwachs in der Wahrnehmung relevanter Unterrichtssituationen aufwiesen. Weiterhin nahm die Anwendung der vorgegebenen Merkmale im zweiten Erhebungszeitpunkt zu. Dies zeigte sich sowohl für die Lernvignette, die direkt im Seminar eingesetzt wurde, als auch für die Transfervignette, die lediglich Teil des Tests war. Der Zuwachs in der professionellen Unterrichtswahrnehmung beschränkte sich allerdings auf zwei Merkmale, die bereits zuvor von vielen Teilnehmenden genutzt wurden: „Konstruktive Einbindung von eigenen Ideen und Schüler*innenvorstellungen in den Unterricht“ und „Aktivierung von Vorwissen“. Weiterhin hat sich gezeigt, dass die Referendar*innen in der Post-Erhebung Merkmale nutzten, die von ihren Fachleiter*innen in der Seminarsitzung angesprochen wurden, auch wenn diese nicht Teil der kognitiven Aktivierung waren.
Die Untersuchung gibt Grund zu der Annahme, dass ein Training der professionellen Unterrichtswahrnehmung besonders dann effizient ist, wenn bereits ein Grundverständnis über das betreffende Merkmal vorliegt. In einigen Fällen hat ein Gespräch im Rahmen des Fachseminars gereicht, damit ein Merkmal in der Post-Erhebung verwendet wurde. In diesen Fällen ist die Anzahl der Personen, die ein Merkmal verwenden, angestiegen. In vielen Fällen ist aber auch eine häufigere Verwendung eines Merkmals durch dieselben Personen zu beobachten. Diese beschreiben in der Post-Erhebung mehr Situationen, in denen diese Merkmale relevant sind.
Die Ergebnisse verdeutlichen die Bedeutung, auf dem aktuellen Stand aufzubauen oder auch zuerst die theoretischen Grundlagen im Fachseminar zu schaffen, bevor ein Training der professionellen Unterrichtswahrnehmung wirksam sein kann. Dies wird auch durch aktuelle Forschungsergebnisse weiter gestützt (Martin et al., 2023).
Bei der Untersuchung der Kommunikation hat sich in einem *stimulated recall* gezeigt, dass die Referendar*innen besonders beim Merkmal „Konstruktive Einbindung der Schüler*innenvorstellungen in den Unterricht“ eine hohe Überschneidung in der Interpretation aufweisen. Bei anderen Merkmalen ist die Interpretation und damit die Überschneidung im Verständnis jedoch deutlich geringer und eine eindeutige Kommunikation nicht immer gegeben. Hier wird erneut die Notwendigkeit einer einheitlichen theoretischen Grundlage deutlich.

### *6.4.3. Vernetzung der Ergebnisse*

Prinzipiell können die Strukturen des Fachseminars genutzt werden, um ein vignettenbasiertes Training der professionellen Unterrichtswahrnehmung in die Lehrkräftebildung zu integrieren. Dafür wäre es jedoch auch notwendig, dass theoretische Grundlagen bereits vorhanden und auch einheitlich sind. Dies macht den Bedarf nach einer stärkeren Vernetzung der ersten und zweiten Phase erneut deutlich (Weber & Czerwenka, 2021).
Dadurch hat sich im Folgenden die Idee einer webbasierten Plattform gefestigt, die als Schnittstelle zwischen der ersten und zweiten Phase der Lehrkräftebildung genutzt werden kann. So kann sichergestellt werden, dass dieselbe theoretische Grundlage vorliegt und passend zur Einführung von spezifischen Merkmalen ist. Dabei soll die theoretische Grundlage frühzeitig mit Unterrichtsbeispielen verknüpft und diese stetig erweitert werden. Gerade für die Effektivität von Praxisphasen ist diese explizite Vernetzung von hoher Bedeutung (Hascher, 2012).
Die Kommunikation in den Fachseminaren wurde im Rahmen des Beitrages nur mit Bezug zu spezifischen Situationen in der Seminarsitzung analysiert. Dies umfasste zum einen die Verwendung von Terminologien, die nicht eindeutig durch das *SEP-Framework* verortet werden konnten und zum anderen Situationen, in denen Vorschläge zu alternativen Handlungsmöglichkeiten generiert wurden. Hierbei hat sich bereits gezeigt, dass eine erfolgreiche Kommunikation zur Unterrichtsqualität wahrscheinlich nicht immer gelingt. Besonders bei der Verwendung unterschiedlicher theoretischer Grundlagen treten dieselben Probleme auf, die bereits im zweiten Beitrag adressiert wurden. In einer hierarchischen Struktur wie dem Fachseminar wird die Kommunikation eindeutig durch die

Fachleiter*innen bestimmt. Dies kann jedoch einen großen Einfluss auf die Vergleichbarkeit des Referendariats haben und wird deshalb ebenfalls im Rahmen der weiterführenden Diskussion zu einer stärkeren Standardisierung aufgegriffen.
Darüber hinaus haben die Referendar*innen in dieser Untersuchung teilweise Merkmale ihrer Fachleiter*innen übernommen, auch wenn diese nicht Teil der Vorgaben waren. Da die Referendar*innen nichts von den Vorgaben wussten, ist dies nicht weiter überraschend gewesen. Es verdeutlicht aber, dass die Fachleiter*innen in einer Position sind, in der sie einen großen Einfluss auf die Referendar*innen haben und sämtliche Vorgaben von außen zunächst durch diesen „Filter" laufen. Hieraus erwächst die Frage, inwiefern das aktuelle System eher einer Lehrkräftebildung im Sinne einer Ausbildung entspricht. Zwar ist besonders das Studium darauf ausgelegt, einen reflektierten Ansatz bei der Gestaltung der eigenen Lehre zu verfolgen, allerdings sollte die Reflexion der Unterrichtshandlungen auch mit dem Eintritt in die Praxis weiterverfolgt werden (Weber & Czerwenka, 2021). Es zeigt sich jedoch beim Feedback in der Praxis häufig eine Vorgabe von Handlungen, die nicht weiter reflektiert, sondern umgesetzt werden sollen (Puttick & Wynn, 2020). Dieser Punkt soll deshalb ebenfalls in der weiterführenden Diskussion thematisiert werden.
Insgesamt ließ sich durch die Studie verdeutlichen, dass eine Entwicklung der professionellen Unterrichtswahrnehmung an das Vorhandensein bestimmter Voraussetzungen gebunden ist. Auch wenn es in der Studie nicht direkt untersucht wurde, ist es naheliegend, dass es sich dabei um zugrundeliegendes theoretisches Wissen handelt, was anschlussfähig an bestehende Erkenntnisse zur Entwicklung der professionellen Unterrichtswahrnehmung ist (Meschede et al., 2017). Dieses beeinflusst sowohl die Beurteilung der Unterrichtsqualität als auch eine erfolgreiche Kommunikation zwischen den Teilnehmenden in den Fachseminaren. Dabei werden auch die Zusammenhänge zwischen unterschiedlichen Merkmalen wichtig. Im Rahmen des Beitrages wurde das ***S**ystematic and **T**ransferable **A**pproach for **R**atings of Instructional Quality* (STAR)-model entwickelt, welches das Zusammenspiel zwischen theoretischen Grundlagen und einer kriterienorientierten Entwicklung der professionellen Unterrichtswahrnehmung verdeutlicht. Da hieraus weitere Überlegungen für eine transparente Gestaltung der Lehrkräftebildung abgeleitet werden können, soll es ebenfalls in der weiterführenden Diskussion ausführlicher aufgegriffen werden. Das Modell bietet auch eine wichtige Basis für die Aufgabenstruktur der webbasierten Plattform „VirtU-net", welche im fünften Beitrag beschrieben wird.

### *6.5. Beitrag 5: Virtuelle Unterrichtshospitationen im Chemieunterricht – Eine Vernetzung der ersten und zweiten Phase der Lehrkräftebildung*

#### *6.5.1. Ausgangslage*

Um den wahrgenommenen Bruch zwischen der ersten und zweiten Phase der Lehrkräftebildung zu adressieren, hat sich die Einbindung von Videovignetten als potenzielle Gelenkstelle und gemeinsame Grundlage angeboten. Somit startete das Projekt „VirtU-net" bereits im April 2021 und wurde stetig durch die Ergebnisse der Dissertation beeinflusst. Dabei wurden vor allem die Ergebnisse des dritten und vierten Beitrages genutzt, um die vignettenbasierte Fortbildung angehender Lehrkräfte zu gestalten. Vor allem die Erkenntnisse des vierten Beitrags waren ausschlaggebend für das Design der Aufgabenstellungen. Diese wurden in mehreren Workshops und in Zusammenarbeit mit Fachleiter*innen getestet und die Rückmeldungen iterativ eingearbeitet. Videovignetten haben sich bereits in anderen Beiträgen als wichtiges Werkzeug zur Ausbildung angehender Lehrkräfte bewährt und werden in Deutschland durch ein breites Angebot unterschiedlicher webbasierter Plattformen zur Verfügung gestellt (Junker et al., 2022). Ein besonderes Anliegen von „VirtU-net Chemie" ist die Vernetzung der ersten und zweiten Phase, wodurch sich eine sinnvolle Ergänzung des bisherigen Angebots ergibt. Eine gemeinsame Plattform bietet eine einfache Möglichkeit der Vernetzung und zeitgleich wird eine gemeinsame theoretische Grundlage aufgebaut, die stetig ergänzt werden kann. Die Basis dafür wird schon im Studium gelegt und kann später im Referendariat erneut aufgegriffen werden.
Der Anspruch besteht dabei nicht darin, das zugrundeliegende *SEP-Framework* als allgemeine Vorgabe in der Lehrkräftebildung zu etablieren. Vielmehr soll eine transparente Basis geschaffen werden, die

die Möglichkeit bietet, auch andere Ansätze der Unterrichtsqualität zu verorten und damit unterschiedliche theoretische Grundlagen zusammenzuführen. Damit soll eine kollaborative Entwicklung von Ansätzen der Unterrichtsqualität verfolgt werden (Charalambous & Praetorius, 2022).

### *6.5.2. Zusammenfassung der Ergebnisse*

Der Praxisbeitrag zur Beschreibung der webbasierten Plattform umfasst keine konkrete Untersuchung und damit keine separaten Ergebnisse. Durch den Beitrag wird ein möglicher Einsatz eines Fortbildungsmoduls sowie der generelle Aufbau der Plattform beschrieben. Diese beruhen dabei auf den vorangegangenen Ergebnissen und nutzen explizit das *SEP-Framework* sowie das *STAR-Model*. Die Plattform umfasst zum aktuellen Stand 13 Videovignetten zu den Themenblöcken „Arbeiten mit Modellen“, „Chemische Reaktion“, sowie „Säure-Base“ und wird auch künftig weiter ergänzt. Eine Anmeldung ist dabei für alle Personen möglich, die in der Lehrkräftebildung tätig sind:
https://virtu-net.idn.uni-hannover.de/

### *6.5.3. Vernetzung der Ergebnisse*

Für die weitere Entwicklung von „VirtU-net Chemie“ sollten die bisherigen Erkenntnisse der Untersuchungen einbezogen werden. Dabei ist vor allem die Kommunikation ein wichtiger Aspekt, der nicht nur weiter untersucht, sondern auch aktiv bei der Ausgestaltung der Webseite berücksichtigt werden sollte. Wichtig wäre es, die Nutzenden ins Gespräch zu bringen. Damit soll nicht nur die Analyse der vorhandenen Vignetten aus möglichst vielen Perspektiven ermöglicht werden, sondern auch die theoretische Grundlage diskutiert werden. Der Vorteil bei dieser Plattform ist der Einsatzbereich, der sowohl gezielt die erste als auch die zweite Phase anspricht. Ein stärkerer Austausch zwischen Vertreter*innen beider Phasen sollte die Vernetzung in der Ausbildung voranbringen und die bisherigen Ideen erweitern (Weber & Czerwenka, 2021). Dabei geht es weniger um eine Standardisierung, als vielmehr um ein gegenseitiges Verständnis der jeweiligen Positionen mit Bezug auf eine gemeinsame Basis. Dafür wären einige konzeptuelle Erweiterungen der Plattform geeignet, die ebenfalls in der weiterführenden Diskussion aufgegriffen werden sollen.

## 7. Übergreifende Diskussion

Die Nutzung einer gemeinsamen theoretischen Grundlage für die Lehrkräftebildung sowie die Vernetzung der ersten und zweiten Phase wären direkte Konsequenzen der bisherigen Arbeit. Ein generischer Rahmen sollte schon früh genutzt werden, um Ansätze miteinander vergleichbar zu machen. Dieser sollte strukturiert und fachlich adaptiert in die Lehrkräftebildung integriert werden, wobei im Folgenden mögliche Umsetzungen und Implikationen diskutiert werden. Diese beziehen sich sowohl auf die Beurteilung von Unterrichtsqualität, als auch auf die professionelle Wahrnehmung und die Frage nach Expertise in der Lehrkräftebildung. Der Aufbau der übergreifenden Diskussion orientiert sich an der Reihenfolge der Einzelbeiträge. Dadurch wird zunächst die theoretische Perspektive der Unterrichtsforschung beleuchtet und davon ausgehend die Praxis der Lehrkräftebildung diskutiert.

### *7.1. Zusammenführung von Ansätzen der Unterrichtsqualität*

Der grundlegende Bedarf einer systematischen Zusammenführung unterschiedlicher Ansätze der Unterrichtsqualität wird zunehmend häufiger explizit geäußert, was auch aus der Diskussion zu kollaborativen Ansätzen hervorgeht (Charalambous et al., 2021). Mit der Erstellung des *Syntheseframeworks* wurde ein Vorgehen zur Zusammenführung von Beobachtungsinstrumenten vorgestellt (Praetorius & Charalambous, 2018). Dieses Vorgehen wurde im Rahmen der Dissertation aufgegriffen und adaptiert, um unterschiedliche Ansätze aus den Naturwissenschaften miteinander zu vergleichen und zu verknüpfen (Heinitz et al., 2022). Dabei wurden fünf generalisierte Kategorien von Gemeinsamkeiten und Unterschieden abgeleitet, die für weitere kollaborative Arbeiten relevant sein können. Besonders wichtig scheint dabei das Ergebnis, dass der Verwendungszweck die Hierarchie der Systematik beeinflusst. Somit können sehr spezifische Merkmale nicht direkt in andere Ansätze übertragen werden, da es ansonsten zu konzeptuellen Überschneidungen kommen kann. Das spricht

wiederum dafür, nicht einfach einen Ansatz für alle vorzugeben. Vielmehr sollte weitestgehend auf bestehende Ansätze aufgebaut und spezifische Adaptionen theoretisch begründet werden. Die generische Grundlage sollte dabei aber explizit adressiert werden, sodass der neue Ansatz darin verortet werden kann. Damit sollte es auch im Nachhinein möglich sein, spezifische Adaptionen miteinander zu vergleichen. Bei konsequenter Umsetzung könnten somit alle bestehenden und künftigen Ansätze zur Erfassung der Unterrichtsqualität über indirekte Verbindungen miteinander vernetzt werden. Die Verknüpfung der einzelnen Ansätze muss jedoch transparent zugänglich sein. Damit könnte auch abweichenden Konzeptualisierungen derselben theoretischen Konstrukte entgegengewirkt werden, die zu uneindeutigen Ergebnissen beim Vergleich von Studien zur Unterrichtsqualität führen können (Christ et al., 2022).

*Implikation 1: Die Idee eines kollaborativen Ansatzes sollte weiterverfolgt, allerdings nicht einfach eine Systematik vorgegeben werden. Vielmehr besteht das Ziel in der Zusammenführung darin, eine breite generische Basis als gemeinsame Grundlage zu nutzen und diese transparent zu adaptieren.*

### *7.2. Gemeinsame Grundlage der Unterrichtsqualität*

Bei allen Erkenntnissen zu spezifischen Verwendungszwecken und Fachspezifika stellt sich die Frage, inwiefern überhaupt eine gemeinsame Grundlage gefunden werden kann. Für eine erhöhte Transparenz wäre es gerade für angehende Lehrkräfte hilfreich. Allerdings können einige spezifische Aspekte eventuell nicht sinnvoll verortet werden (Heinitz et al., 2022). Damit würde prinzipiell eine ähnliche Problematik geschaffen, vor der auch die drei Basisdimensionen stehen (Praetorius et al., 2018). Eine Ergänzung der gemeinsamen Grundlage müsste schon von vornherein eingeplant werden. Es ist unwahrscheinlich, dass der Rahmen so umfassend sein kann, dass alle künftigen Entwicklungen des Unterrichts mitgedacht sind. Besonders strukturelle Änderungen, wie sie z.B. durch zunehmende Digitalisierung oder asynchrones Unterrichten bedingt sind, beeinflussen den Unterricht sehr stark und damit auch das Beurteilungssystem. Weiterhin müssen Fachspezifika, Perspektivenunterschiede oder spezifische Verwendungszwecke ebenso in einer gemeinsamen Grundlage berücksichtigt werden. Dadurch wird das theoretische Konstrukt jedoch sehr komplex und abstrakt, wodurch auch hier die Frage nach einer validen Messung der Unterrichtsqualität relevant wird (Praetorius et al., 2012).
Einen sehr umfassenden Rahmen wieder auf einen konkreten Verwendungszweck anzupassen, ist eine große Herausforderung. Bei der Operationalisierung vom *MAIN-Teach-Modell* (Praetorius et al., 2023) in einem Beobachtungsinstrument, wie es durch das *INSULA*[4] (Wemmer-Rogh et al., 2023) abgebildet ist, fällt vor allem auf, dass die fachlichen Anteile eher gering sind. Gerade bei einem Vergleich mit dem *SEP-Framework* (Heinitz & Nehring, 2020) wird deutlich, dass dort mehr Merkmale zumindest anteilig als fachspezifisch gekennzeichnet sind. Auch im *Syntheseframework* (Praetorius & Charalambous, 2018), welches die Grundlage für das *MAIN-Teach-Modell* und damit indirekt für das *INSULA* bildet, ist ein größerer Anteil der Merkmale als fachspezifisch gekennzeichnet. Dadurch stellt sich folglich die Frage, inwiefern Fachspezifika ausreichend abgebildet sind. Zur Beantwortung dieser Frage müsste prinzipiell erneut eine Verortung bzw. Gegenüberstellung der Ansätze erfolgen, wie es bereits zuvor gemacht wurde (Praetorius et al., 2020b). Daraus wird deutlich, dass eine einmalige Gegenüberstellung unterschiedlicher Ansätze nicht ausreicht. Die Zusammenführung unterschiedlicher Ansätze zu einer gemeinsamen Grundlage kann damit nicht als statisch angesehen werden, sondern vielmehr als stetiger iterativer Prozess.
Die Einbindung von gemeinsamen Grundlagen der Unterrichtsqualität wird teilweise in größerem Maßstab in der Praxis umgesetzt, z.B. durch die Einbindung der drei Basisdimensionen für das Unterrichtsfeedback in Baden-Württemberg (Fauth et al., 2021). Bei der gemeinsamen Grundlage muss die Parsimonität berücksichtigt werden. Für die Praxis scheint es an einigen Stellen eher angemessen, mit einem sparsameren Ansatz zu arbeiten, gerade wenn dieses breit verstanden und akzeptiert werden soll. Auch wenn damit ein kollaborativer Ansatz möglich wäre, müssen die Limitationen berücksichtigt werden, die damit einhergehen (Praetorius et al., 2020a). Die konkrete Ausgestaltung einer gemeinsamen Grundlage bleibt nach wie vor ein offener Diskussionspunkt. Die Notwendigkeit ist

[4] Instrumentarium zur Unterrichtsbeurteilung ausgerichtet auf den Lehrplan 21 im Auftrag von argev

eindeutig, allerdings muss zwischen dem Umfang und der Anwendbarkeit abgewogen werden. Aus einer theoretischen Perspektive heraus ist ein großer Umfang natürlich einfacher, wenn es um die Verortung unterschiedlicher Ansätze geht. Aus einer praktischen Perspektive sollte der Umfang nicht allzu groß werden, gerade wenn konkrete Beobachtungsinstrumente abgeleitet werden sollen. Hierbei kann jedoch mit einem modular konzipierten Ansatz ein Teil der Diskussion aufgefangen werden (Abb. 1). Dieser würde einen stetig gleichbleibenden „Kern" enthalten, der bei Bedarf erweitert werden kann. Wichtig wäre jedoch, dass diese Erweiterungen ebenfalls als gemeinsame Grundlagen konzipiert sind, sodass sie inhaltlich und theoretisch auf den „Kern" abgestimmt sind. Ein modularer Ansatz müsste die Möglichkeit bieten, die Hierarchie anzupassen, wenn spezifische Merkmale in den Fokus gerückt werden sollen. Damit könnten auch spezifische Einsatzzwecke trotz umfassender Rahmung ermöglicht werden. Weiterhin wäre es wichtig, dass fachspezifische Anteile transparent markiert werden, da diese bei der Beurteilung fachliches oder fachdidaktisches Wissen voraussetzen (Neuhaus, 2021). Darüber hinaus könnten auch direkte Zusammenhänge zu den Lernzuwächsen der Schüler*innen herausgestellt und somit Merkmale für „good teaching" und „effective teaching" (Berliner, 2005) gekennzeichnet werden.
Auch bei einem modularen Ansatz müsste beachtet werden, dass die Beurteilenden selbst einen Einfluss auf die Reliabilität des Instruments haben können (Hill et al., 2012). Selbst ein Ratertraining steigert nicht immer die Reliabilität (Praetorius et al., 2012), sodass es für die gemeinsame Grundlage wichtig wäre, wenn die Ebene mit der niedrigsten Abstraktion möglichst konkrete Indikatoren enthält. Damit sollte die Objektivität und Reliabilität bei der Anwendung begünstigt werden. Es muss jedoch beachtet werden, dass unterschiedliche Zusammensetzungen nicht zu grundlegend unterschiedlichen Beurteilungen führen und die Problematik einer Messung durch unterschiedliche Instrumente erneut aufgeworfen wird (Brunner, 2018).

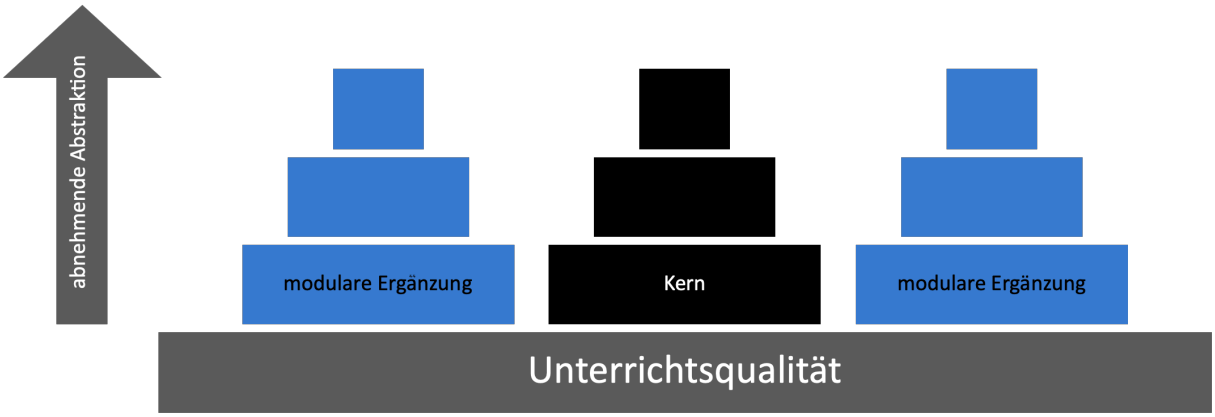



*Abbildung 1: Konzept eines modularen Ansatzes zur Unterrichtsqualitätsbeurteilung. Innerhalb eines Moduls gibt es unterschiedliche Ebenen, deren Abstraktionsgrad stetig abnimmt, sodass die Hierarchie für spezifische Verwendungszwecke nach einem Baukastensystem angepasst werden kann.*

*Implikation 2: Ein modularer Ansatz zur Unterrichtsqualitätsbeurteilung könnte einen Kompromiss zwischen Theorie und Praxis darstellen. Dafür müsste eine flexible Anpassung nach einem Baukastensystem ermöglicht werden, ohne dass die zugrundeliegenden Theorien bei einer Restrukturierung Widersprüche aufweisen.*

### *7.2.1. Eine naturwissenschaftsdidaktische Auslegung der gemeinsamen Grundlage*

Das *MAIN-Teach-Modell* bietet eine potenzielle Grundlage für einen übergreifend genutzten Ansatz. Jedoch wurde erneut die Frage aufgeworfen, inwiefern eine konkrete Operationalisierung aussehen müsste, damit sie auch für den naturwissenschaftlichen Unterricht geeignet ist. Dafür könnte eine erneute Verortung des *SEP-Frameworks* einen ersten Ansatz bieten. Das *SEP-Framework* ist grundsätzlich dazu in der Lage, andere Ansätze der Naturwissenschaftsdidaktiken abzubilden (Heinitz et al., 2022). Dabei bleibt jedoch die eigentlich vorgesehene Struktur der Ansätze nicht erhalten. Sollte es wiederum möglich sein, Grundvoraussetzungen für das Gelingen naturwissenschaftlichen Unterrichts zu formulieren, könnten diese direkt in die Lehrkräftebildung integriert werden. Hier wäre

es erneut die Idee des „Kerns“ einer gemeinsamen naturwissenschaftsdidaktischen Grundlage (Abb.1). Dieser könnte durch die Verortung des *SEP-Frameworks* im *MAIN-Teach-Modell* aus der „Kernebene“ abgeleitet werden. Der „Kern“ der gemeinsamen Grundlage sollte dabei möglichst präzise ausdifferenziert werden, um den Interpretationsrahmen gering zu halten. Dafür könnte das Kodiermanual zum *SEP-Framework* einen Ansatz bieten, sollte jedoch zunächst zu einem Beobachtungsinstrument weiterentwickelt werden. Das Kodiermanual ist erstellt worden, um Aussagen über Unterrichtsqualität zu kategorisieren. Als Basis dienten Beobachtungsinstrumente, die jedoch bei der Ausgestaltung der Systematisierung für eine Clusterung stärker abstrahiert werden mussten (Heinitz & Nehring, 2020). Durch die induktive Ergänzung des Kodiermanuals durch die Aussagen der Fachleiter*innen konnte eine weitere Ebene des Manuals ergänzt werden (Meyer, 2021). Um das *SEP-Framework* zu einem Beobachtungsinstrument weiterzuentwickeln, würde sich eine weitere induktive Ergänzung anbieten. Die Aussagen der Fachleiter*innen und Referendar*innen sind aus einer externen Perspektive heraus formuliert, sodass sie direkt genutzt werden können, um Indikatoren zu generieren. Um eine Abstufung in der qualitativen Ausprägung eines Kriteriums zu entwickeln, können die zuvor analysierten Bewertungen der jeweiligen Aussagen hinzugezogen werden. Da das Beobachtungsinstrument *INSULA* bereits konkrete Indikatoren für eine Ausprägung der Unterrichtsqualitätsmerkmale in vier Stufen enthält (Wemmer-Rogh et al., 2023), kann eine direkte Gegenüberstellung genutzt werden, um ein Beobachtungsinstrument für den naturwissenschaftlichen Unterricht aus dem *SEP-Framework* abzuleiten.

*Implikation 3: Das Kodiermanual zum SEP-Framework könnte genutzt werden, um die Indikatorenebene einer naturwissenschaftsdidaktischen Auslegung der gemeinsamen Grundlage zu erstellen. Dafür könnten weitere induktive Ergänzungen notwendig sein.*

### *7.2.2. Kommunikation auf Basis einer gemeinsamen Grundlage*

Die Verwendung unterschiedlicher Terminologien und Operationalisierungen erschwert die Zusammenführung von Forschungsansätzen zur Unterrichtsqualität (Christ et al., 2022). Dadurch muss für die Entwicklung einer gemeinsamen Grundlage von Unterrichtsqualität zunächst ein *common ground* (Clark & Schaefer, 1989) für eine erfolgreiche Kommunikation gefunden werden. Eine Verortung unterschiedlicher Ansätze der Unterrichtsqualität allein auf Basis der verwendeten Terminologien ist in vielen Fällen nicht unmittelbar möglich (Heinitz et al., 2022). Theoretische Grundlagen und die daraus resultierende Logik der Ansätze können mitunter dazu führen, dass ähnliche Inhalte unterschiedlich benannt oder unterschiedliche Inhalte gleich benannt sind. Im Prinzip müsste jede umfassende Systematik der Unterrichtsqualität den Anspruch haben, möglichst klar definierte Indikatoren zu präsentieren, um eine erfolgreiche Kommunikation zu ermöglichen. Für eine breite Anwendbarkeit ist aber gerade ein breiter Interpretationsrahmen häufig attraktiver, womit eine inhaltliche Inkonsistenz akzeptiert wird.

Diese Problematik erstreckt sich nicht nur auf die hier vorgestellten Bestrebungen zu mehr Einheitlichkeit in der Unterrichtsforschung, sondern hat auch direkte Auswirkungen auf die Ausbildung angehender Lehrkräfte im Referendariat. Diese ist nicht an feste Vorgaben gebunden, wodurch die theoretischen Grundlagen zwischen den Fachseminaren abweichen können (Kunz & Uhl, 2021). Eine gemeinsame Grundlage würde damit auch die Kommunikation in der Lehrkräftebildung vereinfachen, da ein *common ground* bereits konzeptuell vorgegeben oder zumindest stark angeleitet sein würde. Hier muss außerdem der Perspektivenunterschied zwischen Fachleiter*innen und Referendar*innen berücksichtigt werden. Abhängig von der Formulierung der Beurteilung können Merkmale unterschiedlich aufgefasst werden. Bei einem offenen Interpretationsrahmen kann das Referenzsystem uneindeutig sein und die Interpretation leichter abweichen (Fauth et al., 2020), dies könnte die Kommunikation erschweren.

Der *stimulated recall* zur Kommunikation in den Fachseminaren (Heinitz & Nehring, submitted) hat diese Problematik weiter verdeutlicht. In einigen Fällen wurden Terminologien verwendet, die an keine direkt ersichtliche theoretische Grundlage gebunden sind. Dennoch wurden diese Terminologien in den Diskussionen zur Unterrichtsqualität verwendet und durch die Teilnehmenden des Seminars nicht

weiter hinterfragt. Die anschließende Analyse hat gezeigt, dass kein eindeutiges Verständnis zu diesen Begriffen vorlag. Nur in einigen Ausnahmefällen wurden sie von den meisten Teilnehmenden eines Fachseminars gleich gedeutet. Diese unterschiedlichen Interpretationen sind jedoch nicht in den Gesprächen deutlich geworden, sodass die Kommunikation aus Sicht der Teilnehmenden eines Seminars nicht gehemmt war. Dies kann besonders dann problematisch sein, wenn Terminologien durch die Fachleiter*innen verwendet werden, die keine eindeutige Definition haben und die Referendar*innen sich nicht durch scheinbare Unwissenheit exponieren wollen. Dabei wäre gerade in diesen Situationen eine direkte Nachfrage wichtig, damit die unterschiedlichen Verständnisse und damit der fehlende *common ground* deutlich werden. Das Fachseminar hat eine hierarchische Struktur, wodurch theoretische Grundlagen in den meisten Fällen in einer klaren Richtung vorgegeben werden. Dies wird auch daran deutlich, dass Feedback für angehende Lehrkräfte häufig eher einer direkten Handlungsanweisung und weniger einem theoriegeleiteten Reflexionsansatz gleicht (Puttick & Wynn, 2020). Für eine zielführende Kommunikation in den Fachseminaren wäre es besonders wichtig, an den Stand der Referendar*innen anzuschließen.
Um die Differenz zwischen den Fachseminaren zu verringern, wäre es wichtig, die Kommunikation zwischen den Fachseminaren zu stärken. Dabei sollte ein gleichberechtigter Austausch stattfinden und nicht lediglich eine theoretische Grundlage vorgegeben werden. Die Untersuchung zur Beurteilung der kognitiven Aktivierung in den Fachseminaren hat gezeigt, dass die einfache Vorgabe einer theoretischen Rahmung nicht zu einer einheitlichen Einbindung führt (Heinitz & Nehring, submitted). Vielmehr wurden individuelle Interpretationen in die Fachseminare integriert, die implizite Theorien über guten Unterricht verdeutlicht haben (Praetorius et al., 2012). Dadurch erfolgte eine Hybridisierung der theoretischen Grundlage und ein abweichender Input für die Fachseminare. Ein möglicher Ansatz wäre, eine Plattform für den Austausch zu schaffen, bei der die theoretischen Grundlagen direkt an unterrichtspraktische Beispiele geknüpft sind. Die Plattform „VirtU-net“ bietet hierzu einen ersten Zugang, muss jedoch für eine direkte Kommunikation weiterentwickelt werden (Abschnitt 7.5).

*Implikation 4: Der Komfort eines breiten Interpretationsrahmens bei hoher Abstraktion von Unterrichtsqualitätsmerkmalen hemmt eine erfolgreiche Kommunikation und damit die Unterrichtsforschung und Lehrkräftebildung. Oberflächlich wird ein common ground suggeriert, der jedoch inhaltlich nicht vorhanden ist.*

### *7.3. Standardisierung in der Lehrkräftebildung*

Grundsätzlich haben Personen, die die Lehrkräftebildung durchlaufen, am Ende formal die gleiche Qualifikation. Es gibt zwar Spezifikationen bezogen auf Fächer und Schulformen, aber prinzipiell kann später innerhalb desselben Bereichs frei gewechselt werden. Es zeigt sich jedoch, dass es mitunter große Unterschiede zwischen den Fachseminaren gibt (Heinitz & Nehring, 2023; Heinitz & Nehring, submitted). Diese sind teilweise auf die Unterschiede der Fachleiter*innen zurückzuführen, die wiederum durch fehlende Vorgaben bedingt sind (Kunz & Uhl, 2021). Die unterschiedlichen Hintergründe der angehenden Lehrkräfte verstärken diese Varianz. Daraus leitet sich wiederum die Frage ab, ob die Lehrkräftebildung am Ende zu vergleichbaren Qualifikationen führt.
Wenn Unterrichtsqualität stärker standardisiert wäre, könnte die Lehrkräftebildung einheitlicher gestaltet werden. Dabei ist jedoch offen, in welchem Ausmaß die Standardisierung für die Lehrkräftebildung sinnvoll und wünschenswert ist. Grundsätzlich sollen die gleichen Qualifikationen erreicht werden. Eine individuelle Auslegung ist aber auch an vielen Stellen nachvollziehbar. Unterricht ist kein standardisierter Prozess, der immer nach demselben Schema abläuft. Dafür gibt es zu viele individuelle Einflussfaktoren, auf die entsprechend reagiert werden muss. Weiterhin verfolgt Unterricht nicht nur ein einziges Ziel, sodass weitere Variation entstehen kann. Die geringe Standardisierung der Lehrkräftebildung lässt sich auch international finden (z.B. Tatto, 2021), sodass die daraus folgenden Implikationen auch über den deutschsprachigen Raum hinaus relevant sind.
Eine Idee wäre es, den zuvor angesprochenen „Kern“ der Unterrichtsqualität inklusiver modularer Erweiterung (Abschnitt 7.2) herauszuarbeiten und für die Lehrkräftebildung vorzugeben. Somit würde

es die Möglichkeit geben, die Grundlage individuell zu erweitern, ohne die Vergleichbarkeit zu verlieren. Ein modulares System könnte schrittweise aus der Forschung zur Unterrichtsqualität abgeleitet werden und für jeden möglichen Fokus in der Lehrkräftebildung ein Modul zur Evaluation angeboten werden. Damit könnte auch dem Argument begegnet werden, dass es ein hoch individueller Prozess ist, weil dann für jeden Bedarf entsprechend aus einem standardisierten Pool gewählt werden kann. Damit wäre ein Kompromiss zwischen Standardisierung und Individualität möglich. Für eine vergleichbare Umsetzung müssen jedoch die Fachleiter*innen an konkrete Qualifikationen gebunden sein, die eine Vernetzung der theoretischen Grundlage mit der Unterrichtspraxis erlaubt (Weber & Czerwenka, 2021). Es müsste außerdem der individuelle Einfluss auf die Beurteilung reduziert werden. Auch mit der Vorgabe einer gemeinsamen Grundlage kann der Einfluss impliziter Theorien über guten Unterricht nicht ausgeschlossen werden, was sich auch dran zeigt, dass selbst erfahrene Beurteiler*innen diese aufzeigen (Taut & Rakoczy, 2016). Sinnvoll wäre es deshalb, gemeinsame Beurteilungen durchzuführen und eine Form von Ratertraining für die Ausbildung angehender Lehrkräfte einzuführen. Ratertrainings allein führen nicht in jedem Fall zu einer einheitlicheren Beurteilung der Unterrichtsqualität (Praetorius et al., 2012), weshalb sie durch weitere Maßnahmen unterstützt werden sollten. Hierzu könnten Ankerratings hilfreich sein, die genutzt werden können, um die Beurteilungen unterschiedlicher Rater auch im Nachhinein vergleichbar zu machen (Wind et al., 2021). Diese könnten die Bewertung einzelner Merkmale oder sogar die Benotung der gesamten Stunde relativieren. Um die Beurteilung auf eine breitere Basis zu stellen, könnte auch eine Unterstützung durch automatisierte Beurteilungen angedacht werden. Diese sind dabei nicht als eigenständige Beurteilung zu verstehen, würden aber einen zusätzlichen Input für die Diskussion liefern (Abschnitt 7.3.3). Die Idee wäre es hierbei, das Feedback auf eine breitere Basis zu stellen, sodass die daraus resultierenden Abweichungen einen produktiven Diskurs nach dem Vorbild der Unterrichtsdiagnostik erlaubt (Helmke & Lenske, 2013).

*Implikation 5: Die Lehrkräftebildung benötigt konkretere Vorgaben, sollte jedoch genug Raum für individuelle Entwicklung bieten. Ein modularer Ansatz zur Unterrichtsqualitätsbeurteilung könnte dafür eine zielführende Lösung bieten.*

### *7.3.1. Noviz*innen und Expert*innen in der Chemielehrkräftebildung*

Unterscheidungen zwischen Noviz*innen und Expert*innen sind in der Lehrkräftebildung nicht eindeutig möglich und häufig an unterschiedliche Kriterien gebunden (Palmer et al., 2005). Die Ergebnisse der Dissertation haben jedoch gezeigt, dass es sich bei Fachleiter*innen und Referendar*innen um zwei unterschiedliche Gruppen bei der Beurteilung von Unterrichtsqualität handelt (Heinitz & Nehring, 2023). Damit kann eine Unterteilung aufgrund der Unterrichtserfahrung (Berliner, 1989) zumindest teilweise nachvollzogen werden. Die Gruppen unterschieden sich prinzipiell in vier Bereichen: die Anzahl verwendeter Merkmale, die Auswahl der Merkmale, die Bewertung derselben Merkmale und die Benotung der Unterrichtsstunde. Die höhere Anzahl unterschiedlicher wahrgenommener Merkmale weist möglicherweise auf eine größere theoretische Grundlage bei den Fachleiter*innen hin. Es wäre jedoch auch denkbar, dass mehr Merkmale durch die Fachleiter*innen angesprochen wurden, weil diese durch ihre Unterrichtserfahrung Situationen schneller verarbeiten können (Blömeke et al., 2016b). Darüber hinaus kann das zugrundeliegende Wissen die Auswahl der Merkmale beeinflussen (Krepf et al., 2018). Die Unterrichtsstunde behandelt mit der Umkehrbarkeit der chemischen Reaktion ein eher grundlegendes Thema (vgl. Tab. 2). Somit werden Unterschiede im Fachwissen die Beurteilung wahrscheinlich weniger stark beeinflusst haben als Unterschiede im fachdidaktischen und pädagogischen Wissen.

Die Beurteilung derselben Merkmale könnte ebenfalls durch Unterschiede im vorhandenen Wissen beeinflusst sein. Es ist jedoch auch denkbar, dass das Setting, in dem eine Referendarin beurteilt werden sollte, auf Seite der Referendar*innen einen *protective bias* (Vazire, 2010) ausgelöst hat. Diese könnten sich selbst mit der videographierten Referendarin identifiziert und ihre Bewertung einzelner Merkmale entsprechend positiver ausgelegt haben. Da die durchschnittliche Bewertung auch bei den

Referendar*innen eine starke Korrelation zur Benotung aufwies, beeinflusst diese positivere Tendenz möglicherweise nicht nur einzelne Merkmale, sondern auch den Gesamteindruck.
Neben den Unterschieden in der Bewertung und Benotung zwischen den Gruppen ließen sich jedoch auch viele Unterschiede innerhalb der Gruppen finden. Erstaunlich war hierbei, dass selbst niedrig inferente Merkmale teilweise breite Streuungen in der Beurteilung aufgewiesen haben. Es kann folglich selbst bei Fachleiter*innen vorkommen, dass sie durch bestimmte Merkmale einen *bias* entwickeln, der dann einen Einfluss auf die Beurteilung von anderen Merkmalen hat. Dies könnte vergleichbar zu intuitiven Beurteilungen sein, die bereits bei kurzen Unterrichtsausschnitten auftreten und zeitlich stabil bestehen bleiben (Begrich et al., 2021). Diese Beurteilung könnte dann unbewusst auf andere Merkmale übertragen werden. Tendenziell sind diese intuitiven Beurteilungen treffender, je höher das zur Beurteilung benötigte Wissen ist (Begrich et al., 2020). Da es für die Unterrichtsstunde keine objektive Beurteilung gab, kann dieser Zusammenhang für die hier durchgeführte Untersuchung nicht abschließend festgestellt werden. Es ist jedoch naheliegend, dass sehr extreme Beurteilungen wahrscheinlich nicht treffend sind, wenn es eine hohe Standardabweichung in der Beurteilung gab und die intuitiven Beurteilungen in diesen Fällen das Ergebnis verfälscht haben könnten.
Eine Unterscheidung auf Basis der Bewertung derselben Merkmale zeigt in erster Linie, dass Fachleiter*innen tendenziell strenger beurteilen. Dies scheint jedoch nicht ausschließlich an einer höheren Expertise bei der Unterrichtsbeobachtung zu liegen, da ein vignettenbasiertes Training der Unterrichtsbeurteilung bei angehenden Lehrkräften dazu geführt hat, dass diese anschließend positiver beurteilten (Gabriel-Busse et al., 2020). Es könnte jedoch auch sein, dass diese Tendenz nur bis zu einem bestimmten Punkt der Expertise reicht und danach wieder abnimmt. Der *bias* an verschiedenen Stellen der Beurteilung würde erneut dafürsprechen, dass der Beurteilungsprozess oder zumindest Teile davon objektiver gestaltet werden sollten. Dafür könnte Abschnitt 7.3.3 zur Automatisierung der Beurteilung einen möglichen Ansatz bieten.
Offen bleibt im Prinzip die Frage, was eine Expertise im Bereich der Unterrichtsqualitätsbeurteilung genau ausmacht. Aus theoretischer Perspektive kann nicht eindeutig geklärt werden, was Lehrkräfte zu Expert*innen macht (Palmer et al., 2005). Fachleiter*innen sollten Expert*innen für Unterricht sein, aber müssten darüber hinaus genauso Expert*innen der Unterrichtsqualitätsbeurteilungen sein. Weber & Czerwenka (2021) fordern eine konkrete Vorgabe für die Qualifikation von Fachleiter*innen. Dabei sollte entsprechend auch die Unterrichtsqualitätsbeurteilung mitgedacht werden. Vergleichbar zu Kriterien für *quality teaching (*Blömeke et al., 2016a) sollte es auch Kriterien für Fachleiter*innen geben, um der Problematik abweichender Beurteilungen zu begegnen. Welche dafür genau geeignet wären, ist jedoch aufgrund der Komplexität der Thematik nur schwer herauszustellen.
Die theoretische Grundlage hat für die Beurteilung einen genauso großen Einfluss wie die Erfahrung mit unterschiedlichen Unterrichtssituationen. Beides kann dazu führen, dass unterschiedliche Situationen als relevant wahrgenommen werden oder dieselben Situationen unterschiedlich interpretiert werden. Steinwachs & Martens (2023) sprechen sich dafür aus, dass die professionelle Unterrichtswahrnehmung aus Sicht von Lehrkräften und Forschenden als gleichwertig angesehen wird und zur gegenseitigen Professionalisierung genutzt wird. Die Zusammenführung beider Perspektiven scheint ein sinnvolles Ziel zur Verbesserung der Lehrkräftebildung. Eine mögliche Umsetzung wird im Rahmen der Weiterentwicklung von VirtU-net Chemie diskutiert (Abschnitt 7.5).

*Implikation 6: Expertise in der Unterrichtsqualitätsbeurteilung kann nicht allein durch Lehrerfahrung erreicht werden. Fachleiter*innen sollten hierfür an eindeutige Kriterien gebunden werden.*

### *7.3.2. Direkte Handlungsanweisungen in der Lehrkräftebildung*

Der direkte Einfluss der Fachleiter*innen auf die Referendar*innen kann auch zu unerwünschten Entwicklungen führen. Bei der Untersuchung der professionellen Unterrichtswahrnehmung in Beitrag vier ist dahingehend vor allem ein Fachseminar aufgefallen. In diesem wurden in der Prä-Erhebung fast ausschließlich Merkmale der kognitiven Aktivierung verwendet und in der Post-Erhebung kamen Merkmale der „Auswahl und Thematisierung von Inhalten“ hinzu (Heinitz & Nehring, submitted). Die Vorgaben haben eine Thematisierung dieser Dimension nicht vorgesehen, dennoch schien es für die

Fachleiter*in in diesem Fall passend zu sein. Dieses Beispiel verdeutlicht zwei Punkte: Zum einen geht jede externe Vorgabe zunächst über die Fachleiter*innen und kann durch diese angepasst werden. Zum anderen haben die Fachleiter*innen einen direkten Einfluss auf die Referendar*innen und können Entwicklungen hervorrufen, die nicht direkt zur bestehenden theoretischen Grundlage der Referendar*innen passen könnten. Besonders die universitäre Phase sollte eigentlich dazu genutzt werden, eine reflektierte Herangehensweise für die Gestaltung einer Unterrichtsstunde zu entwickeln. Die zweite Phase sollte darauf aufbauen und einen reflektierten Blick auf die Praxis erlauben. Aktuell findet diese Verzahnung jedoch kaum statt (Weber & Czerwenka, 2021).
Eine Studie zur Analyse von schriftlichem Feedback für angehende Lehrkräfte in der Praxis zeigte, dass dieses häufig auf direkte Handlungsanweisungen bezogen ist, ohne konkrete Anbindungen an theoretische Grundlagen (Puttick & Wynn, 2020). Wenn die zweite Phase hauptsächlich darin besteht, das umzusetzen, was von den Fachleiter*innen vorgegeben wird, bestärkt das den wahrgenommenen Bruch zwischen den Phasen. Weiterhin kann es dazu führen, dass eine unreflektierte Übernahme von Unterrichtshandlungen normalisiert wird. Die strikte Hierarchie in den Fachseminaren kann ein reflektiertes Vorgehen hemmen und eine einfache Replikation bestehender Praxis unterstützen.
Ein besonderes Potenzial würde sich für die Referendar*innen ergeben, wenn das Vorgehen anderer Standorte transparenter eingesehen werden kann. Der Vergleich unterschiedlicher Standorte oder auch der Vergleich mit anderen Lehrkräften könnte gut als Reflexionsanlass für das eigene Handeln genutzt werden. Leichte Abweichungen zwischen den Perspektiven können auch hier wieder als Reflexionsanlässe im Sinne einer Unterrichtsdiagnostik zielführend genutzt werden (Helmke & Lenske, 2013). Auch unterschiedliche Interpretationen derselben Unterrichtsvignetten durch die Referendar*innen könnten produktiv genutzt werden, wenn Fachleiter*innen diese direkt als Lernanlass aufgreifen (Gabriel-Busse et al., 2020). Hierfür würde sich die Nutzung einer gemeinsamen Videoplattform mit der Möglichkeit zum Austausch zwischen den Fachseminaren anbieten.

*Implikation 7: Die Lehrkräftebildung sollte besonders in der zweiten Phase transparenter gestaltet werden und den Austausch zwischen unterschiedlichen Fachseminaren stärken.*

### *7.3.3. Automatisierung zur Unterstützung von Unterrichtsqualitätsbeurteilungen*

Individuelle Einflüsse werden an vielen Stellen der Lehrkräftebildung deutlich. Grundsätzlich muss berücksichtigt werden, dass gerade die zweite Phase ein individueller Entwicklungsprozess für angehende Lehrkräfte ist. Dennoch würde eine stärkere Standardisierung von grundlegenden Aspekten dazu führen, dass individuelle Faktoren gezielter berücksichtigt werden könnten, da weniger Varianz durch Rahmenbedingungen erzeugt wird.
Sofern es möglich ist eine gemeinsame Grundlage der Unterrichtsqualität herauszustellen, könnte diese wie in Abschnitt 7.2 und 7.2.1 beschrieben operationalisiert werden. Eine konkrete Operationalisierung verringert nicht nur den Interpretationsrahmen, sondern ermöglicht auch eine Automatisierung der Beurteilung. Diese ist selbstverständlich nicht als Ersatz, sondern vielmehr als Ergänzung und Unterstützung der regulären Beurteilungen zu verstehen und kann in unterschiedlichem Ausmaß erfolgen. Voraussetzung ist jedoch zunächst ein trainiertes Modell mit einem ausreichend großen Datensatz. Die automatisierte Unterstützung könnte verschiedene Aspekte umfassen. Beispielsweise könnten Aussagen zur Unterrichtsqualität frei getätigt und auf Basis der gemeinsamen Grundlage kodiert werden, um ein gemeinsames Verständnis sicherzustellen. Zusätzlich könnten nicht angesprochene Merkmale hervorgehoben werden, um das Feedback zu erweitern. Ähnlich zur zuvor angesprochenen modularen Erweiterung der Grundlage könnten Entwicklungsprofile abgeleitet werden. Diese könnten dann die Beurteilung durch das Aufzeigen von möglichen Aspekten ergänzen. Da die berechneten mittleren Bewertungen eine hohe Korrelation zu den Benotungen der Unterrichtstunde aufgewiesen haben (Heinitz & Nehring, 2023), könnte diese auch automatisch berechnet werden, um einen Vergleich für die Benotung zu erhalten. Prinzipiell könnte diese Auswertung immer weitergedacht werden, sodass auch Gespräche automatisch transkribiert und kodiert werden. Dieser Ansatz kann auch genutzt werden, um ggf. zugrundeliegende Theorien herauszustellen, oder den Entwicklungsbedarf von angehenden Lehrkräften aufzuzeigen. Eine weitere

Überlegung wäre es, Unterschiede in der Kommunikation herauszustellen, indem z.B. Fachseminarsitzungen aufgezeichnet und kodiert werden. Damit hätten die Fachleiter*innen eine direkte Übersicht, was laut Kodierung angesprochen wurde und was eigentlich intendiert war. Noch weitergedacht, könnten darauf aufbauend auch Unterrichtsvideos direkt mit Blick auf die Unterrichtsqualität analysiert werden. Es ist bereits möglich Unterrichtshandlungen automatisch zu kategorisieren, wobei auch hier die Analyse der Unterrichtsqualität als nächster Schritt herausgestellt wird (Foster et al., 2024). Hierauf aufbauend wäre auch eine Analyse des Perspektivenunterschieds interessant. Eine automatische Kodierung einer Unterrichtsaufzeichnung wäre zwar auch eine Analyse aus einer externen Perspektive, allerdings könnte der Referenzrahmen (Fauth et al., 2020) deutlich einfach eingesehen und angepasst werden. Außerdem könnte das System durch weitere Informationen ergänzt werden, sodass es möglicherweise eine Synthese der Beurteilungsperspektiven ergibt oder auch eine völlig neue Perspektive auf den Unterricht geworfen wird, die über die klassischen drei Perspektiven (Clausen, 2002) hinausgeht. Die Möglichkeiten scheinen praktisch endlos zu sein und viel Potenzial für weitere Arbeiten zu enthalten.

*Implikation 8: Eine Automatisierung in der Unterrichtsqualitätsbeurteilung bietet viele Möglichkeiten, muss jedoch als Hilfsmittel und Ergänzung und nicht als Ersatz betrachtet werden. Besonders wenn es um den individuellen Entwicklungsbedarf einer angehenden Lehrkraft geht, ist eine menschliche und damit individuelle Betrachtungsweise von hoher Bedeutung. Es kann jedoch helfen, eine automatisierte Unterstützung als neutralen Beobachter hinzuzuziehen.*

### *7.4. Entwicklung der professionellen Unterrichtswahrnehmung in den Fachseminaren*

Die professionelle Unterrichtswahrnehmung ist ein wichtiger Mediator zwischen dem Wissen einer Lehrkraft und den Lernzuwächsen der Schüler*innen (Blömeke et al., 2022). Damit nimmt die Förderung derselben einen besonderen Stellenwert für die Lehrkräftebildung ein. Hierfür sollten die Strukturen der Fachseminare entsprechend ausgelegt sein. Besonders im Kontext einer professionellen Unterrichtswahrnehmung kann die Expertise der Fachleiter*innen jedoch nicht immer vorausgesetzt werden, da diese hauptsächlich auf Basis ihrer Unterrichtserfahrung für diese Rolle ausgewählt werden (Weber & Czerwenka, 2021). Eine langjährige Unterrichtserfahrung ist nicht an bestimmte theoretische Grundlagen gebunden. Das theoretische Wissen hat jedoch einen größeren Einfluss auf die professionelle Unterrichtswahrnehmung als praktische Lehrerfahrung (Stürmer et al., 2014). Die Ergebnisse des dritten und vierten Beitrags legen nahe, dass es Abweichungen in der professionellen Unterrichtswahrnehmung zwischen den Fachseminaren gibt. Für eine vergleichbare Ausbildung wird hier wieder die Bedeutung einer gemeinsamen Grundlage deutlich. Diese sollte auch bei der professionellen Unterrichtswahrnehmung fachspezifische Aspekte berücksichtigen (Steffensky et al., 2015). Die eingesetzten Vignetten in Beitrag drei und vier enthielten klar erkennbare fachliche Inhalte, die die Beurteilung entsprechend beeinflusst haben könnten (Tab. 2). Weiterhin sollte berücksichtigt werden, dass auch generische Aspekte bei der professionellen Wahrnehmung fachspezifische Unterschiede aufweisen können (Stahnke & Friesen, 2023).

Aus der Dissertation gehen erste Möglichkeiten für eine wirksame Implementierung von Videovignetten im Referendariat hervor. Dabei wurde explizit ein Ansatz verfolgt, die professionelle Unterrichtswahrnehmung möglichst nah an den bestehenden Seminarstrukturen zu entwickeln, um die Unterrichtsforschung und -praxis zu verknüpfen. Der untersuchte Ansatz zur vignettenbasierten Entwicklung der professionellen Unterrichtswahrnehmung kann sehr leicht durch die Fachleiter*innen umgesetzt werden, zeigt jedoch Limitationen durch die individuellen Unterschiede im Fachseminar. Selbst bei einer gemeinsamen Grundlage für die Fachleiter*innen müssen die Unterschiede auf Seiten der Referendar*innen berücksichtigt werden. Deren theoretische Grundlagen haben einen großen Einfluss auf eine erfolgreiche Einbindung von Videovignetten für die Förderung der professionellen Wahrnehmung (Martin et al., 2023). Die Referendar*innen nehmen in den Fachseminaren die Rolle der Lernenden ein, daher ist es naheliegend, dass auch hier das Vorwissen einen großen Einfluss auf die Lernzuwächse hat (Simonsmeier et al., 2021). Bereits bei der Wahrnehmung muss der Einfluss des

zugrundeliegenden theoretischen Wissens berücksichtigt werden, damit bestimmte Situationen überhaupt als relevant eingestuft werden können (König et al., 2014; Wolff et al., 2016)
Die professionelle Unterrichtswahrnehmung kann auch bei einer gemeinsamen Grundlage durch implizite Theorien, individuelle Verständnisse und hierarchische Strukturen in den Seminaren beeinflusst werden. Deshalb könnten vor allem breit genutzte Onlineplattformen zur Vernetzung der Standorte ein großes Potenzial bieten. Wie zuvor beschrieben gibt es hierfür bereits ein breites Angebot (Junker et al., 2022), das durch die entwickelte Plattform VirtU-net Chemie weiter ergänzt wurde. Dabei ist die Vernetzung von erster und zweiter Phase, wie es auf VirtU-net Chemie angedacht ist, ein wichtiger Ansatzpunkt für die Entwicklung der professionellen Unterrichtswahrnehmung. Diese ist gekennzeichnet durch die Verknüpfung von theoretischem Wissen mit unterrichtspraktischen Beispielen (Sherin & van Es, 2009), was auf der Plattform direkt umgesetzt wird. Das besondere Potenzial der Onlineplattform liegt hier vor allem darin, dass die theoretische Grundlage für beide Phasen gleichbleibt und stetig ergänzt wird. Damit würde eine der großen Herausforderungen der Lehrkräftebildung in Deutschland abgemildert (Weber & Czerwenka, 2021). Weiterhin setzt die Plattform auf einen systematischen Ausbau des theoretischen Wissens, das schrittweise mit weiteren unterrichtspraktischen Beispielen verknüpft wird. Diese explizite und schrittweise Vernetzung bietet gerade für die Praxisphase ein hohes Lernpotenzial, da dies für angehende Lehrkräfte häufig schwierig ist (Korthagen & Kessels, 1999). Videovignetten bieten den Vorteil, dass die Komplexität reduziert wird (Beck et al., 2002), allerdings geht damit auch ein beschränkter Einblick in die Gesamtheit der Stunde einher. Eine Limitation beim Einsatz von Videovignetten ist, dass einige Merkmale zeitlich weniger stabil sind und somit ein kurzer Unterrichtsausschnitt nur einen beschränkten Einblick in die Ausprägung eines Merkmals für die gesamte Unterrichtsstunde bieten kann (Begrich et al., 2020, 2021). Dies verdeutlicht die Bedeutung eines kriterienorientierten Ansatzes, wie er auf VirtU-net umgesetzt wird (Heinitz & Nehring, 2024). Dabei geht deutlich hervor, dass die Ausprägung eines Merkmals oder einer Dimension der Unterrichtsqualität aus mehreren Aspekten besteht, die getrennt voneinander wahrgenommen werden können. VirtU-net Chemie bietet darüber hinaus die Möglichkeit, den weiteren Verlauf einer Stunde in zusätzlichen Vignetten anzusehen, wodurch ein Mittelweg zwischen Reduktion und Komplexität angeboten wird.

*Implikation 9: Die professionelle Unterrichtswahrnehmung ist für Lehrkräfte von hoher Bedeutung. Ein gezieltes Training im Referendariat setzt jedoch eine stärkere Vernetzung der ersten und zweiten Phase voraus, wofür VirtU-net Chemie und vergleichbare Plattformen einen Ansatz bieten können.*

### *7.5. Ausbau von VirtU-net Chemie zu einer Lernplattform*

Die bisherigen Überlegungen zu einer stärkeren Standardisierung und Kommunikation innerhalb der Lehrkräftebildung sollen auch bei der weiteren Entwicklung von VirtU-net Chemie berücksichtigt werden. Unter Einbezug einer automatisierten Unterrichtsqualitätsbeurteilung (Abschnitt 7.3.3) könnte VirtU-net Chemie zu einem Selbstlernangebot weiterentwickelt werden. Dabei würden die Nutzer*innen die Unterrichtsvignetten frei beurteilen, was dann wiederum automatisch kodiert werden könnte. Dies könnte sowohl auf Ebene der Wahrnehmung als auch auf Ebene der Interpretation geschehen, sodass eine Abstufung in der Komplexität der Aufgabe möglich wäre. Dies würde an das bisherige Konzept anschließen, bei welchem die Merkmale schrittweise eingeführt werden. Aus einem Abgleich mit allen relevanten Merkmalen einer Vignette könnte dann ein Entwicklungsangebot abgeleitet werden. Prinzipiell würde dies aber auch ohne eine automatisierte Auswertung möglich sein. Dafür müssten den Nutzer*innen die relevanten Merkmale zur Verfügung gestellt werden, woraus diese dann eigenständig oder angeleitet weitere Entwicklungsschritte ableiten könnten.
Die Plattform könnte auch die Beurteilungen der Nutzer*innen zusammenführen, wodurch ein direkter Vergleich unterschiedlicher Sichtweisen ermöglicht wird. Diese könnten bei Unterschieden als Lerngelegenheit genutzt werden (Gabriel-Busse et al., 2020) oder auch als Ankerbeispiele, um die Beurteilung unterschiedlicher Personen vergleichbarer zu machen (Wind et al., 2021). Ein somit herausgestellter *bias* könnte auch direkt an die Nutzenden zurückgemeldet und als Lerngelegenheit genutzt werden. Die Beurteilung könnte dafür als Annotationen in die Videos eingebunden werden,

was auch für die Kommunikation genutzt werden kann. Damit würde die Möglichkeit bestehen, auch fachseminarübergreifend über die Bedeutung von spezifischen Merkmalen zu reden und das Verständnis der Unterrichtsqualität zu vereinheitlichen.
An vielen Stellen ist in der Lehrkräftebildung bereits ein kriterienorientierter Ansatz zur Entwicklung der professionellen Unterrichtswahrnehmung angedacht. Dennoch ist er häufig nicht direkt als solcher gekennzeichnet. Die Idee des *Systematic and Transferable Approach for Ratings of Instructional Quality* (STAR)-Model (Heinitz & Nehring, submitted) ist es jedoch gerade diesen Ansatz transparent abzubilden. Im Rahmen der Lernmodule auf VirtU-net ist der Ansatz bereits integriert. Hieraus könnte sich auch ein Potenzial ergeben, die Unterrichtsqualitätsbeurteilung durch unterschiedliche Personen theoriebasiert gegenüberzustellen. Das Modell könnte für die Plattform weiter interaktiv aufbereitet werden, damit die jeweiligen Qualitätsmerkmale direkt mit theoretischen Aspekten verknüpft werden. Damit könnte eine einfache Möglichkeit zur Verknüpfung von Unterrichtsforschung und -praxis erreicht werden.
Darüber hinaus könnte das Training der professionellen Wahrnehmung durch den bereits angedachten Einsatz von 360°-Unterrichtsvignetten erweitert werden. Die Besonderheit wäre hierbei die freie Fokussierung im Raum und damit eine erhöhte Voraussetzung der Fähigkeit, auf relevante Merkmale zu fokussieren. Damit könnten angehende Lehrkräfte auf reale Unterrichtssituationen vorbereitet werden, in denen häufig parallel relevante Situationen ablaufen. Ein Beispiel hierfür wären Experimentierphasen, bei welchen die Sicherheit der Schüler*innen im Blick behalten und gleichzeitig auf individuelle Rückfragen eingegangen werden muss. Das volle Potenzial von 360°-Unterrichtsvignetten für die Lehrkräftebildung muss jedoch noch untersucht werden und ist ein aktuelles Thema in der Unterrichtsforschung (z.B. Daltoè et al., 2024).

*Implikation 10: Die Plattform VirtU-net Chemie bietet einen Ansatz für die stärkere Vernetzung von Unterrichtsforschung und -praxis. Dabei gibt es jedoch noch viel Potenzial zur Weiterentwicklung.*

## 8. Limitationen

Die Limitationen der Einzelbeiträge werden in den entsprechenden Artikeln (im Anhang) aufgeführt. Grundlegend muss jedoch der explorative Charakter von Beitrag drei und vier beachtet werden. Die daraus abgeleiteten Implikationen sind bewusst vorsichtig formuliert und werden an vielen Stellen durch aktuelle Erkenntnisse aus der Forschung unterstützt. Dennoch müssen die Studien mit größeren Stichproben repliziert werden, bevor eine eindeutige Aussage möglich ist. Die Dissertation ist durch qualitative Methoden geprägt, da bisher wenige Erkenntnisse für die Beurteilung der Unterrichtsqualität im Referendariat vorliegen, sodass zunächst ein Einblick geschaffen werden musste.

## 9. Ausblick und Weiterführung der bisherigen Arbeit

Im Rahmen der Diskussion wurden einige Aspekte aufgegriffen, die für die Weiterentwicklung der bisherigen Arbeit und die Verbesserung der Lehrkräftebildung relevant werden. Zentrales Element ist dabei immer die allgemeine Nutzung einer gemeinsamen Grundlage für Unterrichtsqualität. Dabei muss es nicht ein perfektes Framework geben, das für jeden Bereich genutzt wird, aber alle bestehenden Ansätze sollten sich transparent zu bestehenden Systematiken positionieren, sodass sie untereinander vergleichbar bleiben. Die Pluralität der Ansätze in der Forschung erschwert die einheitliche Arbeit in der Lehrkräftebildung, was entsprechend wahrnehmbare negative Auswirkungen auf diese hat. Damit sie grundlegend verbessert werden kann, sollten Forschungsansätze und Praxisansätze weniger als Gegenteil, sondern vielmehr als unterschiedliche Perspektiven auf denselben Gegenstand angesehen werden. Auch wenn dies allzu selten explizit kommuniziert wird, ist die unterschiedliche Ansicht häufig präsent und unbestreitbar hinderlich. Die Kommunikation zwischen der ersten und zweiten Phase, sowie der Unterrichtsforschung, stellt einen weiteren wichtigen Punkt dar, der konsequent weiterentwickelt werden müsste. Hierfür können Onlineplattformen mit Unterrichtsbeispielen (z.B. VirtU-net Chemie) einen wichtigen Beitrag leisten, allerdings müssen sie dafür auch in der Breite implementiert werden. Grundsätzlich wäre hier ein möglichst zentraler Ansatz gut, da die einzelnen Angebote ansonsten nicht immer sichtbar sind. Die Kommunikation müsste

darüber hinaus aktiv gesucht werden und auch über Onlineplattformen hinausgehen. Dafür sind wiederum gezielte Veranstaltungen für die Vernetzung von Unterrichtsforschung und -praxis von besonderer Bedeutung.
Neben diesen grundlegenden Erkenntnissen wurden auch einige Punkte aufgegriffen, die das System verbessern könnten, jedoch eher zukunftsorientiert gedacht sind. Hier ist vor allem die stärker automatisierte Auswertung von Unterrichtsqualitätsbeurteilungen, aber auch von Unterricht selbst denkbar. Dabei muss jedoch bedacht werden, dass das System nie vollständig standardisiert oder automatisiert werden kann. Die Ausbildung von angehenden Chemielehrkräften ist an vielen Stellen durch individuelle Bedürfnisse geprägt. Diese stammen dabei sowohl von angehenden Lehrkräften als auch den Schüler*innen und Ausbildenden. Die individuellen Bedürfnisse sind nachvollziehbar, da es sich um Menschen handelt, die nie vollständig standardisiert werden können. Entsprechend müssen individuelle Aspekte berücksichtigt werden, damit das System stets flexibel auf neue Herausforderungen reagieren kann. Dabei ist jedoch ein transparentes und nachvollziehbares Vorgehen wichtig, um Verständnis bei allen Beteiligten zu garantieren. Schlussendlich wird ein gemeinsames Ziel verfolgt.

## 10. Literatur


Allen, L., O'Connell, A. and Kiermer, V. (2019), How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. Learned Publishing, 32: 71-74. https://doi.org/10.1002/leap.1210

Beck, R. J., King, A., & Marshall, S. K. (2002). Effects of videocase construction on preservice teachers' observations of teaching. *Journal of Experimental Education* 70(4), 345–361. https://doi.org/10.1080/00220970209599512

Begrich, L., Fauth, B., Kunter, M. *et al.* (2017). Wie informativ ist der erste Eindruck? Das Thin-Slices-Verfahren zur videobasierten Erfassung des Unterrichts. *Z Erziehungswiss* **20** (Suppl 1), 23–47. https://doi.org/10.1007/s11618-017-0730-x

Begrich, L., Fauth, B., & Kunter, M. (2020). Who sees the most? Differences in students' and educational research experts' first impressions of classroom instruction. *Social Psychology of Education*, *23*(3), 673–699. https://doi.org/10.1007/s11218-020-09554-2

Begrich, L., Kuger, S., Klieme, E., & Kunter, M. (2021). At a first glance – How reliable and valid is the thin slices technique to assess instructional quality? *Learning and Instruction*, *74*. https://doi.org/10.1016/j.learninstruc.2021.101466

Berliner. (1989). New Directions for Teacher Assessment. In Pfleiderer J.(eds) Proceedings of the 1988 ETS Invitational Conference. Educational Testing Service. University of Michigan.

Berliner, D.C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, 56(3), 205–213. https://doi.org/10.1177/0022487105275904.

Blömeke, S., Gustafsson, J. E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie / Journal of Psychology* (Vol. 223, Issue 1, pp. 3–13). Hogrefe Publishing. https://doi.org/10.1027/2151-2604/a000194

Blömeke, S., Olsen, R.V., Suhl, U. (2016a). Relation of Student Achievement to the Quality of Their Teachers and Instructional Quality. In: Nilsen, T., Gustafsson, JE. (eds) Teacher Quality, Instructional Quality and Student Outcomes. IEA Research for Education, vol 2. Springer, Cham. https://doi.org/10.1007/978-3-319-41252-8_2

Blömeke, S., Busse, A., Kaiser, G., König, J., & Suhl, U. (2016b). The relation between content-specific and general teacher knowledge and skills. *Teaching and Teacher Education*, 56, 35–46. https://doi.org/10.1016/j.tate.2016.02.003

Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progress. *Learning and Instruction*, 79. https://doi.org/10.1016/j.learninstruc.2022.101600

Blomberg, G., Stürmer, K., & Seidel, T. (2011). How pre-service teachers observe teaching on video: Effects of viewers' teaching subjects and the subject of the video. *Teaching and Teacher Education*, 27(7), 1131–1140. https://doi.org/10.1016/j.tate.2011.04.008

Blomberg, G., Renkl, A., Sherin, G., Borko, H., & Seidel, T. (2013). Five research-based heuristics for using video in pre-service teacher education. *Journal for Educational Research*, 90–114.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, 31(4), 416–440. https://doi.org/10.3102/0162373709353129

Brunner, E. (2018). Quality of Mathematics Teaching: A Question of Perspective. Journal Fur Mathematik-Didaktik, 39(2), 257–284. https://doi.org/10.1007/s13138-017-0122-z

Camburn, E., & Barnes, C. A. (2004). Assessing the validity of a language arts instruction log through triangulation. *Elementary School Journal*, 105(1), 49–73. https://doi.org/10.1086/428802

Charalambous, C. Y., & Praetorius, A. K. (2020). Creating a forum for researching teaching and its quality more synergistically. *Studies in Educational Evaluation*, 67. https://doi.org/10.1016/j.stueduc.2020.100894

Charalambous, C. Y., Praetorius, A.-K., Sammons, P., Walkowiak, T., Jentsch, A., & Kyriakides, L. (2021). Working more collaboratively to better understand teaching and its quality: Challenges faced and possible solutions. *Studies in Educational Evaluation*, 71, 101092. https://doi.org/10.1016/j.stueduc.2021.101092

Charalambous, C. Y., & Praetorius, A. K. (2022). Synthesizing collaborative reflections on classroom observation frameworks and reflecting on the necessity of synthesized frameworks. *Studies in Educational Evaluation*, 75. https://doi.org/10.1016/j.stueduc.2022.101202

Christ, A. A., Capon-Sieber, V., Grob, U., & Praetorius, A.-K. (2022). Learning processes and their mediating role between teaching quality and student achievement: A systematic review. *Studies in Educational Evaluation*, 75, 101209. https://doi.org/10.1016/j.stueduc.2022.101209

Clark, H. H., & Schaefer, E. F. (1989). Contributing to Discourse. *Cognitive Science*, *13*(2), 259–294. https://doi.org/10.1207/s15516709cog1302_7

Daltoè T., Ruth-Herbein E., Brucker B., Jaekel A.-K., Trautwein U., Fauth B., Gerjets P. & Göllner R., (2024). Immersive insights: Unveiling the impact of 360-degree videos on preservice teachers' classroom observation experiences and teaching-quality ratings, *Computers & Education*, https://doi.org/10.1016/j.compedu.2023.104976.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy*, 24(2), 267–329. https://doi.org/10.1177/0895904808330173

Döbrich, P., & Storch, H. (2012). Pädagogische Entwicklungsbilanzen mit Studien-SEMinaren oder: Lehrerausbildung ohne Bilanzierung? http://www.gfpf.info

Dorfner, T., Förtsch, C., & Neuhaus, B. J. (2017). Die methodische und inhaltliche Ausrichtung quantitativer Videostudien zur Unterrichtsqualität im mathematisch-naturwissenschaftlichen Unterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, 23(1), 261–285. https://doi.org/10.1007/s40573-017-0058-3

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives.

*Zeitschrift für Pädagogik*. Beiheft, 66(1), 138–155. https://doi.org/10.25656/01:25870

Fauth, B., Herbein, E., Maier, J.L. (2021). Beobachtungsmanual zum Unterrichtsfeedbackbogen Tiefenstrukturen https://ibbw.kultus-bw.de/site/pbs-bw-km-root/get/documents_E-523136125/KULTUS.Dachmandant/KULTUS/Dienststellen/ibbw/Empirische%20Bildungsforschung/Programme-und-Projekte/Unterrichtsfeedbackbogen/IBBW_Unterrichtsfeedbackbogen_Manual_Juni%202021.pdf

Foster, J. K., Korban, M., Youngs, P., Watson, G. S., & Acton, S. T. (2024). Automatic classification of activities in classroom videos. Computers and Education: Artificial Intelligence, 6, 100207. https://doi.org/10.1016/j.caeai.2024.100207

Gabriel-Busse, K., Groß-Mlynek, L., Feldhoff, T. & Harring, M. (2020). Eine Unterrichtssequenz – unterschiedliche Einschätzungen. Analyse videografierter Unterrichtssequenzen als Bestandteil einer evidenzbasierten Lehrer/innenausbildung. In I. Gogolin, B. Hannover & A. Scheunpflug (Hrsg.), *Zeitschrift für Erziehungswissenschaft. Band 4: Evidenzbasierung in der Lehrkräftebildung.* (S. 291-314). Wiesbaden: Springer

Gabriel-Busse, K., & Lipowsky, F. (2021). 90 Minuten Mathematikunterricht bei gleichbleibender Unterrichtsqualität? – Analysen zur zeitlichen Stabilität und Generalisierbarkeit von Ratings zur Unterrichtsqualität im 2. Schuljahr. *Unterrichtswissenschaft*, 49(1), 137–163. https://doi.org/10.1007/s42010-020-00086-4

Goodwin, C. (1994). Professional Vision. American Anthropologist, New Series, 96(3), 606–633.
Heinitz, B., & Nehring, A. (2020). Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung. *Unterrichtswissenschaft* (Vol. 48, Issue 3, pp. 319–360). Springer VS. https://doi.org/10.1007/s42010-020-00074-8

Hascher, T. (2012). Forschung zur Bedeutung von Schul- und Unterrichtspraktika in der Lehrerinnen- und Lehrerbildung. *Beiträge zur Lehrerbildung* 30, S. 87-98 - DOI: 10.25656/01:13805

Heinitz, B., & Nehring, A. (2020). Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung. *Unterrichtswissenschaft* (Vol. 48, Issue 3, pp. 319–360). Springer VS. https://doi.org/10.1007/s42010-020-00074-8

Heinitz, B., Szogs, M., Förtsch, C., Korneck, F., Neuhaus, B. J., & Nehring, A. (2022). Unterrichtsqualität in den Naturwissenschaften. Eine vergleichende Gegenüberstellung von Ansätzen zwischen Fachspezifik und Generik. *Zeitschrift Für Didaktik Der Naturwissenschaften*, 28(1), 10. https://doi.org/10.1007/s40573-022-00146-5

Heinitz, B., & Nehring, A. (2023). Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors. International Journal of Science Education. https://doi.org/10.1080/09500693.2023.2213382

Heinitz & Nehring (2024). Virtuelle Unterrichtshospitationen im Chemieunterricht – Eine Vernetzung der ersten und zweiten Phase der Lehrkräftebildung. *Der mathematische und naturwissenschaftliche Unterricht : MNU, 182-190.*

Heinitz & Nehring (submitted). Perception and Interpretation of Cognitive Activation in Science Pre-Service Teacher Education – The Necessity of Establishing a Common Approach

Hellermann, C., Gold, B., & Holodynski, M. (2015). Förderung von Klassenführungsfähigkeiten im Lehramtsstudium: Die Wirkung der Analyse eigener und fremder Unterrichtsvideos auf das strategische

wissen und die professionelle Wahrnehmung. *Zeitschrift für Entwicklungspsychologie Und Pädagogische Psychologie*, 47(2), 97–109. https://doi.org/10.1026/0049-8637/a000129

Helmke, A., & Lenske, G. (2013). Unterrichtsdiagnostik als Voraussetzung für Unterrichtsentwicklung (Vol. 31, Issue 2).

Helmke, & Schrader. (2008). Merkmale der Unterrichtsqualität - Potenzial Reichweite und Grenzen SEMINAR 3-2008 S.17-47. 17–47.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. In *Educational Researcher* (Vol. 41, Issue 2, pp. 56–64). https://doi.org/10.3102/0013189X12437203

Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity. *Frontiers in Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.00721

Jacobs, J. K., & Morita, E. (2002). Japanese and American Teachers' Evaluations of Videotaped Mathematics Lessons. *Journal for Research in Mathematics Education* (Vol. 33, Issue 3).

Jentsch, A., Schlesinger, L., Heinrichs, H., Kaiser, G., König, J., & Blömeke, S. (2021). Erfassung der fachspezifischen Qualität von Mathematikunterricht: Faktorenstruktur und Zusammenhänge zur professionellen Kompetenz von Mathematiklehrpersonen. *J Math Didakt*, 42, 97–121. https://doi.org/10.1007/s13138-020-00168-x.

Junker, R., Zucker, V., Oellers, M., Rauterberg, T., Konjer, S., Meschede, N., & Holodynski, M. (2022). *Lehren und Forschen mit Videos in der Lehrkräftebildung*. Waxmann.

Kang, H., & van Es, E. A. (2019). Articulating Design Principles for Productive Use of Video in Preservice Education. *Journal of Teacher Education*, 70(3), 237–250. https://doi.org/10.1177/0022487118778549

Kang, J. (2020). Interrelationship Between Inquiry-Based Learning and Instructional Quality in Predicting Science Literacy. *Research in Science Education*. https://doi.org/10.1007/s11165-020-09946-6

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring Usable Knowledge: Teachers' Analyses of Mathematics Classroom Videos Predict Teaching Quality and Student Learning. *American Educational Research Journal*, 49(3), 568–589. https://doi.org/10.3102/0002831212437853

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In J. Baumert, et al. (Hrsg.), PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Wiesbaden: VS. https://doi.org/10.1007/978-3-322-97590-4_12

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I. „Aufgabenkultur" und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), TIMSS-Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente (S. 43–57). Bonn: Bundesministerium für Bildung und Forschung.

König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A., Kaiser G. (2014) Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach, *Teaching and Teacher Education*, 38, 76-88, https://doi.org/10.1016/j.tate.2013.11.004.

Korneck, F., Krüger, M., & Szogs, M. (2017). Professionswissen, Lehrerüberzeugungen und Unterrichtsqualität angehender Physiklehrkräfte unterschiedlicher Schulformen. In E. Sumfleth, & H. Fischler (Eds.), Professionelle Kompetenzen von Lehrkräften der Chemie und Physik. *Studien zum Physik- und* Chemielernen, Bd. 200 (pp. 1–21). Logos.

Korthagen, F. A. J., & Kessels, J. P. A. M. (1999). Linking Theory and Practice: Changing the Pedagogy of Teacher Education. *Educational Researcher*. 28(4).

Kramer, C., König, J., Kaiser, G., Ligtvoet, R., & Blömeke, S. (2017). Der Einsatz von Unterrichtsvideos in der universitären Ausbildung: Zur Wirksamkeit video- und transkriptgestützter Seminare zur Klassenführung auf pädagogisches Wissen und situationsspezifische Fähigkeiten angehender Lehrkräfte. *Zeitschrift Fur Erziehungswissenschaft*, 20, 137–164. https://doi.org/10.1007/s11618-017-0732-8

Krepf, M., Plöger, W., Scholl, D., & Seifert, A. (2018). Pedagogical content knowledge of experts and novices—what knowledge do they activate when analyzing science lessons? *Journal of Research in Science Teaching*, 55(1), 44–67. https://doi.org/10.1002/tea.21410

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., & Ewald, S. (2016). Bedingungen und Effekte von Unterricht: Aktuelle Forschungsperspektiven aus der pädagogischen Psychologie. McElvany, N., Bos, W., Holtappels, H. G., Gebauer, M.M., Schwabe, F. (Hrsg.): Bedingungen und Effekte guten Unterrichts. Münster; New York: Waxmann S.9-31.

Kunter, M., & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV (S. 85–113).

Kunz, H. & Uhl S. (2021). Allgemeine Ziele, Aufbau und Struktur des Vorbereitungsdienstes in den Bundesländern. In (Peitz J. & Harring M.) Das Referendariat. Ein systematischer Blick auf den schulpraktischen Vorbereitungsdienst (S. 15 – 27). Waxmann.

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, 36, 143–152. https://doi.org/10.1016/j.tate.2013.07.010

Labudde, P., Viiri, J., Fischer, H. E., & Neumann, K. (2014). Summary and Discussion. In H. E. Fischer, P. Labudde, K. Neumann, & J. Viiri (Eds.), Quality of instruction in physics. Comparing Finland, Switzerland and Germany (pp. 111–127). Waxmann.

Lortie, D. (1975). Schoolteacher: A Sociological Study. London: University of Chicago Press.

Martin, M., Farrell, M., Seidel, T., Rieß, W., Könings, K. D., van Merriënboer, J. J. G., & Renkl, A. (2023). Knowing what matters: Short introductory texts support pre-service teachers' professional vision of tutoring interactions. *Teaching and Teacher Education*, 124. https://doi.org/10.1016/j.tate.2023.104014

Mayring, P. (2014). Qualitative Content Analysis Theoretical Foundation, Basic Procedures and Software Solution. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173

Meschede, N., Fiebranz, A., Möller, K., & Steffensky, M. (2017). Teachers' professional vision, pedagogical content knowledge and beliefs: On its relation and differences between pre-service and in-service teachers. *Teaching and Teacher Education*, 66, 158–170. https://doi.org/10.1016/j.tate.2017.04.010

Meyer, J. (2021). Weiterentwicklung und Evaluation eines Instruments zur Erfassung der Qualität von naturwissenschaftlichem Unterricht (Masterarbeit, Institut für Didaktik der Naturwissenschaften, Leibniz Universität Hannover)

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D.G. (2009). Preferred reporting items for systematic reviews and Meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, 151(4), 264–270.

Müller, M.M., Gold, B. (2022). Videobasierte Erfassung wissensbasierten Verarbeitens als Teilprozess der professionellen Unterrichtswahrnehmung – Analyse eines geschlossenen und offenen Verfahrens. *Z Erziehungswiss* **26**, 7–29 (2023). https://doi.org/10.1007/s11618-022-01128-6

Neuhaus, B. J. (2021). Quality of instruction in biology education. Unterrichtswissenschaft, 49(2), 273–283. https://doi.org/10.1007/s42010-021-00114-x

Palmer, D. J., Stough, L. M., Burdenski, T. K., and Gonzales, M. (2005). Identifying teacher expertise: an examination of researchers' decision making. *Educ. Psychol.* 40, 13–25. https://doi.org/10.1207/s15326985ep4001_2

Niedersächsisches Kultusministerium (2022). Kerncurriculum für das Gymnasium – gymnasiale Oberstufe die Gesamtschule – gymnasiale Oberstufe das Fachgymnasium das Abendgymnasium das Kolleg – Chemie. https://cuvo.nibis.de/cuvo.php?p=download&upload=140

Palmer, D. J., Stough, L. M., Burdenski, T. K., & Gonzales, M. (2005). Identifying teacher expertise: An examination of researchers' decision making. *Educational Psychologist*, 40(1), 13–25. https://doi.org/10.1207/s15326985ep4001_2

Praetorius, A. K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, 22(6), 387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002

Praetorius, A. K. (2013). Einschätzung von Unterrichtsqualität durch externe Beobachterinnen und Beobachter. Eine kritische Betrachtung der aktuellen Vorgehensweise in der Schulpraxis. *Beiträge zur Lehrerbildung* 31. DOI:10.25656/01:13845

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. *ZDM - Mathematics Education*, 50(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM - Mathematics Education*, 50(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A.-K., Rogh, W., & Kleickmann, T. (2020a). Blinde Flecken des Modells der drei Basisdimensionen von Unterrichtsqualität? Das Modell im Spiegel einer internationalen Synthese von Merkmalen der Unterrichtsqualität. *Unterrichtswissenschaft*, 48(3). https://doi.org/10.1007/s42010-020-00072-w

Praetorius, A. K., Herrmann, C., Gerlach, E., Zülsdorf-Kersting, M., Heinitz, B., & Nehring, A. (2020b). Teaching quality in different subject matters in German-speaking countries—Inbetween genericness and subject-specificity. *Unterrichtswissenschaft*, 48(3), 409–446. https://doi.org/10.1007/s42010-020-00082-8

Praetorius, A.-K., Charalambous, C., Wemmer-Rogh, W., Gossner, L., Herrmann, C., Ufer, S., Gräsel, C. & Keller, S. (2023). MAIN-Teach-Modell. *Zenodo*. https://doi.org/10.5281/zenodo.8280389

Puttick, S., & Wynn, J. (2020). Constructing 'good teaching' through written lesson observation feedback. *Oxford Review of Education*, *47*(2). https://doi.org/10.1080/03054985.2020.1846289

Schlesinger, L., Jentsch, A. Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education* 48, 29–40 (2016). https://doi.org/10.1007/s11858-016-0765-0

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

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499. https://doi.org/10.3102/0034654307310317

Sherin, M. G. (2001). Developing a professional vision of classroom events: Teaching elementary school mathematics. *Beyond classical pedagogy* (S. 75–93). Erlbaum.

Sherin, M. G. (2007). The development of teachers' professional vision in video clubs. In *Video research in the learning sciences* (pp. 383-395). Erlbaum.

Sherin, M. G., & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, 60(1), 20–37. https://doi.org/10.1177/0022487108328155

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching of the new reform. Harvard *Educational Review*, 57, 1–22.

Simonsmeier, B.A., Flaig, M., Deiglmayr, A., Schalk, L. & Schneider, M. (2022) Domain-specific prior knowledge and learning: A meta-analysis, *Educational Psychologist*, 57:1, 31-54, DOI: 10.1080/00461520.2021.1939700

Stahnke R and Friesen M (2023) The subject matters for the professional vision of classroom management: an exploratory study with biology and mathematics expert teachers. *Front. Educ.* 8:1253459. https://doi.org/10.3389/feduc.2023.1253459

Star, J. R., & Strickland, S. K. (2008). Learning to observe: Using video to improve preservice mathematics teachers' ability to notice. *Journal of Mathematics Teacher Education*, 11(2), 107–125. https://doi.org/10.1007/s10857-007-9063-7

Steffensky, M., Gold, B., Holdynski, M., & Möller, K. (2015). Professional Vision of Classroom Management and Learning Support in Science Classrooms—Does Professional Vision Differ Across General and Content-Specific Classroom Interactions? In*ternational Journal of Science and Mathematics Education*, 13(2), 351–368. https://doi.org/10.1007/s10763-014-9607-0

Steinwachs, J., Martens, H. (2023). Praktiken der Unterrichtswahrnehmung hinsichtlich des Umgangs mit anthropomorphen und teleologischen Schülervorstellungen im Evolutionsunterricht. *ZfDN* 29, 12. https://doi.org/10.1007/s40573-023-00161-0

Strong, M., Gargani, J., & Hacifazlioğlu, Ö. (2011). Do We Know a Successful Teacher When We See One? Experiments in the Identification of Effective Teachers. *Journal of Teacher Education*, 62(4), 367–382. https://doi.org/10.1177/0022487110390221

Stürmer, K., Könings, K. D., & Seidel, T. (2013). Declarative knowledge and professional vision in teacher education: Effect of courses in teaching and learning. *British Journal of Educational Psychology*, 83(3), 467–483. https://doi.org/10.1111/j.2044-8279.2012.02075.x

Stürmer, K., Könings, K. D., & Seidel, T. (2014). Factors Within University-Based Teacher Education Relating to Preservice Teachers' Professional Vision. *Vocations and Learning*, 8(1), 35–54. https://doi.org/10.1007/s12186-014-9122-z

Sunder, C., Todorova, M., & Möller, K. (2016). Kann die professionelle Unterrichtswahrnehmung von Sachunterrichtsstudierenden trainiert werden? – Konzeption und Erprobung einer Intervention mit Videos aus dem naturwissenschaftlichen Grundschulunterricht. *Zeitschrift Für Didaktik Der Naturwissenschaften*, 22(1), 1–12. https://doi.org/10.1007/s40573-015-0037-5

Tatto, M. T. (2021). Teacher Education in the United States of America: An Overview of the Policies, Pathways, Issues and Relevant Research. *Teacher Education Policy and Research* (177–194). Springer Singapore. https://doi.org/10.1007/978-981-16-3775-9_13

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction*, 46, 45–60. https://doi.org/10.1016/j.learninstruc.2016.08.003

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, 98(2), 281–300. https://doi.org/10.1037/a0017908

Weber K. E. & Czerwenka K. (2021). Anschlussfähigkeit und Kooperation der ersten und zweiten Phase der Lehrkräftebildung. In (Peitz J. & Harring M.) Das Referendariat. Ein systematischer Blick auf den schulpraktischen Vorbereitungsdienst (S. 255 – 264). Waxmann.

Wemmer-Rogh, W., Gossner, L., Wehrli, F., & Praetorius, A.-K. (2023). Instrumentarium zur Unterrichtsbeurteilung ausgerichtet auf den Lehrplan 21 in Auftrag der argev. Validierte Version auf Basis des MAIN-Teach-Modells (INSULA 2.0). *Zenodo*. https://doi.org/10.5281/zenodo.8280334

Wiernik, A. (2020). Guter Unterricht in der zweiten Phase der Lehramtsausbildung. Eine qualitativ-rekonstruktive Studie zum impliziten Unterrichts- und Professionsverständnis von Seminarleitenden. Bad Heilbrunn: Verlag Julius Klinkhardt DOI:10.25656/01:20578

Wind, S. A., Jones, E., & Bergin, C. (2021). Principals' severity affects teacher evaluation: Statistical adjustments mitigate effects. *School Effectiveness and School Improvement*, 32(3), 413–429. https://doi.org/10.1080/09243453.2021.1892773

Wolff, C. E., Jarodzka, H., van den Bogert, N., & Boshuizen, H. P. A. (2016). Teacher vision: expert and novice teachers' perception of problematic classroom management scenes. *Instructional Science*, 44(3), 243–265. https://doi.org/10.1007/s11251-016-9367-z

Yin, R. K. (2018). Case Study Research and Applications: Design and Methods (6th ed.). Thousand Oaks, CA: Sage.

# 11. Anhang

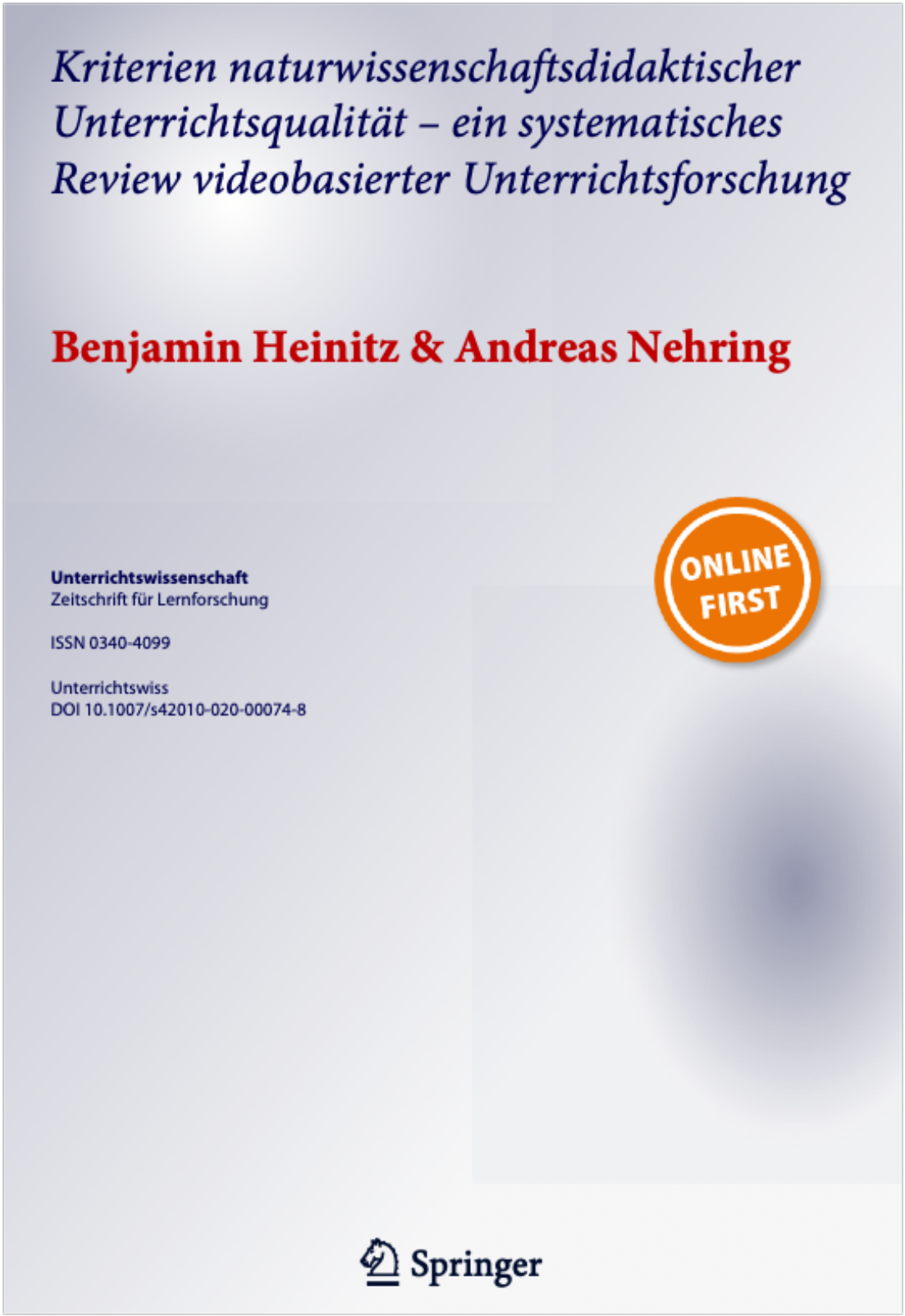
Im Folgenden sind die Einzelbeiträge mit den dazugehörigen Publikationen eingebunden. Die Zusammenfassungen/Abstracts sind direkt aus den Publikationen entnommen, die Beschreibung der Eigenleistung orientiert sich an der *Contributor Roles Taxonomy* (CRediT, Allen et al., 2019).

## *11.1. Beitrag 1: Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung*

Heinitz, B., & Nehring, A. (2020). Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung. In Unterrichtswissenschaft (Vol. 48, Issue 3, pp. 319–360). Springer VS. https://doi.org/10.1007/s42010-020-00074-8

Zusammenfassung:
Im Kontext des Themenheftes „Die Verortung von Merkmalen der Unterrichtsqualität zwischen Generik und Fachspezifik“ bestehen die Ziele des vorliegenden Beitrags darin, einen systematischen Überblick über die in quantitativen, naturwissenschaftsdidaktischen Videostudien angewendeten Kriterien zur Konzeptualisierung und Messung von Unterrichtsqualität zu geben und inhaltliche Beziehungen zu generisch konzeptualisierten und formulierten Kriterien der allgemeinen Unterrichtsforschung herzustellen und zu reflektieren. Im Rahmen eines systematischen Reviews wurden dazu die Kriterien, Operationalisierungen und Beobachtungsitems von 28 Videostudien analysiert, ordnend zusammengefasst und in Beziehung zu einem generisch orientierten Syntheseframework zur Unterrichtsqualität gesetzt, das drei Ebenen unterschiedlicher Auflösungsgrade umfasst. Es wurde ein Kodierverfahren umgesetzt, das wiederum zwischen drei Graden der Übereinstimmung differenziert. Die Objektivität der Kodierungen wurde mit einer Doppelkodierung abgesichert ($0{,}66 < \kappa < 0{,}78$). Die Ergebnisse zeigen, dass aus naturwissenschaftsdidaktischen Videostudien insgesamt 388 Kriterien extrahiert und in Beziehung zum Syntheseframework gesetzt werden konnten. Insgesamt lässt sich, auf der allgemeinsten Ebene des Frameworks, ein hoher Grad an Überschneidungen zwischen naturwissenschaftsdidaktischen und eher generisch formulierten Perspektiven herstellen. Auf detaillierteren Ebenen zeigt sich, dass zahlreiche Kriterien spezifisch operationalisiert und in Form von „fachspezifischen Perspektivierungen“ dem Framework zugeordnet werden können. Auf dieser Grundlage präsentiert der Beitrag eine Systematisierung generischer und naturwissenschaftsdidaktischer Perspektiven auf Unterrichtsqualität. In der Diskussion wird aufgezeigt, dass diese Systematisierung disziplinenübergreifende Kommunikation ermöglicht und neue Impulse für die weitere naturwissenschaftsdidaktische Unterrichtsqualitätsforschung setzt.

CRediT Author Statement zur Eigenleistung:
Benjamin Heinitz: Conceptualization**,** Writing — original draft, review & editing, Investigation, Data Curation, Formal analysis, Visualization


Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung
Benjamin Heinitz & Andreas Nehring
ONLINE FIRST
Unterrichtswissenschaft
Zeitschrift für Lernforschung
ISSN 0340-4099
Unterrichtswiss
DOI 10.1007/s42010-020-00074-8
Springer

Unterrichtswiss
https://doi.org/10.1007/s42010-020-00074-8

ALLGEMEINER TEIL

# Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung

**Benjamin Heinitz · Andreas Nehring**



**Zusammenfassung** Im Kontext des Themenheftes „Die Verortung von Merkmalen der Unterrichtsqualität zwischen Generik und Fachspezifik“ bestehen die Ziele des vorliegenden Beitrags darin, einen systematischen Überblick über die in quantitativen, naturwissenschaftsdidaktischen Videostudien angewendeten Kriterien zur Konzeptualisierung und Messung von Unterrichtsqualität zu geben und inhaltliche Beziehungen zu generisch konzeptualisierten und formulierten Kriterien der allgemeinen Unterrichtsforschung herzustellen und zu reflektieren. Im Rahmen eines systematischen Reviews wurden dazu die Kriterien, Operationalisierungen und Beobachtungsitems von 28 Videostudien analysiert, ordnend zusammengefasst und in Beziehung zu einem generisch orientierten Syntheseframework zur Unterrichtsqualität gesetzt, das drei Ebenen unterschiedlicher Auflösungsgrade umfasst. Es wurde ein Kodierverfahren umgesetzt, das wiederum zwischen drei Graden der Übereinstimmung differenziert. Die Objektivität der Kodierungen wurde mit einer Doppelkodierung abgesichert ($0{,}66 < \kappa < 0{,}78$). Die Ergebnisse zeigen, dass aus naturwissenschaftsdidaktischen Videostudien insgesamt 388 Kriterien extrahiert und in Beziehung zum Syntheseframework gesetzt werden konnten. Insgesamt lässt sich, auf der allgemeinsten Ebene des Frameworks, ein hoher Grad an Überschneidungen zwischen naturwissenschaftsdidaktischen und eher generisch formulierten Perspektiven herstellen. Auf detaillierteren Ebenen zeigt sich, dass zahlreiche Kriterien spezifisch operationalisiert und in Form von „fachspezifischen Perspektivierungen“ dem Framework zugeordnet werden können. Auf dieser Grundlage präsentiert der Beitrag eine Systematisierung generischer und naturwissenschaftsdidaktischer Perspektiven

**Zusatzmaterial online** Zusätzliche Informationen sind in der Online-Version dieses Artikels (https://doi.org/10.1007/s42010-020-00074-8) enthalten.

B. Heinitz (✉) · A. Nehring
Institut für Didaktik der Naturwissenschaften, Leibniz Universität Hannover, Am Kleinen Felde 30, 30167 Hannover, Deutschland
E-Mail: heinitz@idn.uni-hannover.de

Published online: 04 May 2020

auf Unterrichtsqualität. In der Diskussion wird aufgezeigt, dass diese Systematisierung disziplinenübergreifende Kommunikation ermöglicht und neue Impulse für die weitere naturwissenschaftsdidaktische Unterrichtsqualitätsforschung setzt.

**Schlüsselwörter** Unterrichtsqualität · Generik · Fachspezifik · Videostudien · Systematisches Review

## Quality of instruction in science education—a systematic review including goals, contents, and methods of science education

**Abstract** Aiming at systematizing science education and rather generic perspectives on the quality of instruction, this article presents a systematic review of quantitative video studies from science education. Based on a description of the specifics of goals, contents and methods of science education and a literature research, using the PRISMA statement, the items and criteria of 28 video studies were systematized and compared to an interdisciplinary-oriented generic framework. That framework originated from generic classroom research and distinguishes between dimensions and subdimensions of quality of instruction. A coding method was implemented that differentiates between three degrees of congruence. The objectivity of the encoding was verified with a double encoding ($0.66 < \kappa < 0.78$). The results show that a total of 388 criteria could be extracted and compared to the framework. Overall, we found a high degree of comparability between science education and generic perspectives, even though terminologies and content-related emphases can turn out to be different. The article discusses the extent to which the extended framework of generic and science education dimensions of teaching quality can facilitate interdisciplinary communication and which research perspectives emerge from the mutual reflection of the disciplines as well as from the interplay of subject-specific goals, contents and methods.

**Keywords** Quality of instruction · Generic framework · Content-specifity · Video studies · Systematic review

## 1 Einleitung

### 1.1 Forschungsbedarf

Die Bemühungen um ein tieferes und systematisches Verständnis von Unterrichtsqualität gehören zweifelsohne zum Kern naturwissenschaftsdidaktischer Forschungen[1] (Brovelli 2018; Rehm 2018; Wilhelm 2018). Dabei konnte sich seit den TIMSS-Videostudien aus den Jahren 1995 (Stigler et al. 1999) eine For-

[1] Um deutlich zu machen, dass es sich bei den in diesem Beitrag berücksichtigten Disziplinen (Didaktik der Biologie, der Chemie, der Physik und des Sachunterrichts sowie Science Education) um kein homogenes Forschungsfeld ist, wird der Terminus „naturwissenschaftsdidaktische Forschungen“ im Plural verwendet.

schungslinie empirisch-fundierter Unterrichtsforschung etablieren, die insbesondere auf die Analyse videographierter Unterrichtsstudien fokussiert. Darin werden Unterrichtsqualitätsmerkmale in Form von Kriterien und konkreten Beobachtungsitems operationalisiert und zur Beurteilung von Unterricht angewendet. Facetten von Unterrichtsqualität werden anhand dieser Kriterien explizierbar (Dorfner et al. 2017).

Wenngleich in naturwissenschaftsdidaktischen Forschungen durchaus Bezüge zur allgemeinen Unterrichtsforschung hergestellt werden, ist ein systematischer und studienübergreifender Abgleich von Kriterien aus naturwissenschaftsdidaktischen Studien und generischen bzw. generisch formulierten Kriterien bisher weitgehend ausgeblieben. Dabei hat ein solcher Abgleich das Potenzial, einen Beitrag zu einem umfassenderen Bild von Unterrichtsqualität zu leisten und Kommunikation zwischen den eher naturwissenschaftsdidaktisch und den eher generisch orientierten Disziplinen auf eine gemeinsame theoretische Grundlage zu stellen.

Vor diesem Hintergrund präsentiert der vorliegende Beitrag Einblicke in ausgewählte Spezifika der Ziele, Inhalte und Methoden der naturwissenschaftlichen Unterrichtsfächer sowie ein systematisches Review der, in quantitativen Videostudien, genutzten Kriterien für die Beschreibung und Beurteilung von naturwissenschaftlichem Unterricht. Die in diesen Studien verwendeten Kriterien werden in Beziehung zu einem Syntheseframework gesetzt, das generische und fachspezifisch zu operationalisierende Kriterien der Unterrichtsqualität abbildet (Praetorius und Charalambous 2018; Praetorius et al. eingereicht; Tab. 4 dieses Beitrags). Damit soll herausgearbeitet werden, in welchem Umfang sich Beziehungen zwischen Kriterien der naturwissenschaftsdidaktischen Forschungen sowie dem Syntheseframework herstellen lassen, welche naturwissenschaftsdidaktischen Spezifikationen notwendig sind und auf welchen Ebenen Ergänzungen des Syntheseframeworks zielführend sein können.

### 1.2 Verortung des Beitrags im Themenheft

Der vorliegende Beitrag ist Teil des Themenheftes „Die Verortung von Merkmalen der Unterrichtsqualität zwischen Generik und Fachspezifik“. Darin werden Abgleiche für weitere Fächer vorgenommen und in einem gemeinsamen Synthesebeitrag zusammengeführt (Praetorius et al. eingereicht). Das Syntheseframework dient als kommunikative und konzeptuelle Gelenkstelle dieser Abgleiche. Es wurde im Rahmen eines vorangegangenen Reviewprozesses erarbeitet (Praetorius und Charalambous 2018), wobei wiederum generische, fachspezifische und „hybride“, d. h. sowohl fachspezifisch als auch generisch konzeptualisierte, Frameworks zum Mathematikunterricht eingeflossen sind. Bedeutsam für diesen Beitrag ist dabei, dass die Formulierung der Unterrichtsqualitätskriterien nicht allein mathematikspezifisch, sondern auf den Ebenen der Dimensionen, Subdimensionen und Indikatoren auf einem übergreifenden Level ohne fachspezifische Operationalisierungen vorgenommen wurden. Somit bietet sich die Möglichkeit, die formulierten Dimensionen, Subdimensionen und Indikatoren auf weitere Fächer zu übertragen und zu überprüfen, inwiefern Überschneidungen der Konzeptualisierung von Unterrichtsqualität vorliegen und inwiefern Zuordnungen möglich sind. Im Rahmen des Einleitungsbeitrages des The-

menhefts (Praetorius et al. eingereicht) wird eine deutsche Übersetzung des Syntheseframeworks vorgestellt, welche explizit generische Formulierungen verwendet, sodass eine Übertragung des Syntheseframeworks auf andere Fächer erleichtert wird. Diese deutsche Übersetzung des Syntheseframeworks wird für den abschließenden Vergleich zwischen den Fächern, aber auch zur Darstellung der Ergebnisse dieses Beitrags verwendet. Für die grundlegende Auswahl des Syntheseframeworks gegenüber anderen Möglichkeiten zur strukturierten Erfassung von Unterrichtsqualitätskriterien sei an dieser Stelle auf den Einleitungsbeitrag verwiesen, in dem diese Frage ausführlich erläutert wird.

### *1.2.1 Vergleich von Kriterien für Unterrichtsqualität*

Bei der Abbildung von Kriterien zur Erfassung der Unterrichtsqualität wird in den meisten Ansätzen eine hierarchische Form der Strukturierung genutzt. Kriterien werden dabei unter bestimmten Begriffen zusammengefasst, wobei die Anzahl der Strukturierungsebenen zwischen einzelnen Ansätzen variieren kann. Ein Kriterium für Unterrichtsqualität kann wiederum selbst als Sammelbegriff für einzelne Beobachtungsitems verwendet werden, sodass dieses Kriterium nicht automatisch die unterste Ebene der Hierarchie bilden muss. Für den Vergleich von Kriterien der naturwissenschaftsdidaktischen Forschung mit dem Syntheseframework, muss deshalb zunächst eine gemeinsame Bezugsebene gefunden werden, auf der eine zielführende Gegenüberstellung möglich ist. Die Ebenen, die zur strukturierten Abbildung von Unterrichtsqualität genutzt werden, unterscheiden sich in der Anzahl der jeweils untergeordneten Elemente und somit in ihrem Grad der Abstraktion. Hierbei ist es grundsätzlich so, dass eine „höhere“ Ebene einen stärkeren Grad der Abstraktion aufweist, da eine größere Anzahl unterschiedlicher Konzepte unter einem Begriff versammelt ist (geringerer Auflösungsgrad). Die „unteren“ Ebenen werden dagegen, durch eine abnehmende Anzahl untergeordneter Konzepte, zunehmend konkretisiert und dadurch weniger abstrakt (höherer Auflösungsgrad). Verallgemeinert kann zunächst festgehalten werden, dass ein Beobachtungsitem, das sich auf genau einen Aspekt einer Unterrichtsbeobachtung bezieht, die „unterste“ und konkreteste mögliche Ebene der hierarchischen Struktur darstellt. Die darüber liegende Ebene, die mehrere dieser Items zu einem didaktischen Konzept zusammenfasst, wird im Folgenden als ein „Kriterium“ für Unterrichtsqualität bezeichnet. Zwischen den einzelnen Ebenen kann es durch unterschiedliche theoretische Grundlagen dazu kommen, dass terminologische Übereinstimmung vorliegen, auch wenn diese inhaltlich nicht gegeben ist. Andersherum kann es vorkommen, dass eine unterschiedliche Terminologie verwendet wird, auch wenn inhaltlich dasselbe Konzept beschrieben wird. Aus diesem Grund stellt ein Vergleich auf einer möglichst differenzierten „unteren“ Ebene für eine Gegenüberstellung unterschiedlicher Ansätze zur Beschreibung von Unterrichtsqualität den besten Ansatz dar, wobei die inhaltliche Ausrichtung eines Kriteriums auch im Fall einer einheitlichen Terminologie beachtet werden muss. Um eine Handhabbarkeit dieser Vergleiche zu gewährleisten, erfolgt die Darstellung in diesem Beitrag auf „höheren“, d. h. allgemeineren Ebenen. Über Online-Supplements sind die Vergleiche auf den untersten, d. h. am stärksten detailliertesten Ebenen zugänglich gemacht.

### 1.3 Verortung des Beitrags im Feld der Unterrichtsqualitätsforschung

Die Fragen danach, was Unterrichtsqualität ist, sind keinesfalls einfach zu beantworten und in hohem Maße davon abhängig, mit welchem Ziel, vor welchem theoretischen Hintergrund, mit welchem Grad der Ausdifferenzierung konzeptualisiert und operationalisiert oder inwiefern auf Tiefen- und Oberflächenstrukturebene analysiert wird (Givvin et al. 2005; Kunter und Ewald 2016). Wenngleich mit der Unterscheidung zwischen „good teaching", „effective teaching" und „quality teaching" eine grundlegende Unterscheidung für die Ableitung von Zielkriterien vorliegt (Berliner 2005; Brunner 2018), bestehen nach wie vor zahlreiche Herausforderungen für die Forschung. Dazu gehören nicht nur methodische für eine valide Messung von Unterrichtsqualität durch externe Beobachtende (z. B. Praetorius et al. 2014), sondern auch theoretische Herausforderungen. Die lerntheoretische Verortung von Unterrichtsqualitätskriterien, die Frage der Hierarchisierung verschiedener Kriterien oder die Frage des Verhältnisses von fachspezifischen oder generischen Kriterien von Unterrichtsqualität lassen sich dazu zählen (Schlesinger und Jentsch 2016).

Auf einer strukturellen Ebene ist Unterrichtsforschung ein disziplinenübergreifendes Unterfangen, das sowohl im Rahmen der erziehungswissenschaftlichen, pädagogisch-psychologischen (Kunter und Ewald 2016) als auch der fachdidaktischen Forschung (Sumfleth und Fischer 2013) umgesetzt wird. So zeigen Dorfner et al. (2017) in ihrem systematischen Review auf, dass sich in zahlreichen Videostudien aus dem mathematisch-naturwissenschaftsdidaktischen Bereich Bezüge zu den drei Basisdimensionen von Unterrichtsqualität finden lassen (siehe auch Einleitungsbeitrag in diesem Themenheft), jedoch die konkreteren Konzeptualisierungen und Operationalisierungen sehr verschieden sein können. Ein Grund hierfür kann darin bestehen, dass naturwissenschaftsdidaktische Forschungen vor dem Hintergrund der Spezifik des jeweils beforschten Fachunterrichts oder gar Themenfeldes stattfinden, jedoch kooperativ und interdisziplinär ausgerichtet sein können. Das Modell der drei Basisdimensionen stellt in der Forschung zur Unterrichtsqualität einen häufig verwendeten Ansatz der Strukturierung dar, wird jedoch im Rahmen dieses Themenheftes nicht als Bezugsrahmen für den Vergleich der fachspezifischen Qualitätskriterien verwendet. Trotz der vielfachen Verweise auf das Modell der drei Basisdimensionen in der Forschung zur Unterrichtsqualität wird häufig der fachspezifische (aber teilweise auch der generische) Ergänzungsbedarf der Dimensionen um weitere Aspekte herausgestellt. Im Einleitungsbeitrag dieses Themenheftes wird der Bedarf nach der Ergänzung ausführlich dargestellt, weshalb an dieser Stelle darauf verwiesen sei. Exemplarisch sei jedoch die Untersuchung von Szogs et al. (2017) genannt, in deren Rahmen die *fachliche Korrektheit* und die *fachliche Transparenz* als Ergänzungen der drei Basisdimensionen genannt werden oder die Untersuchung von Brunner (2018) in deren Rahmen die *fachliche Korrektheit* ergänzt wird. Das verwendete Syntheseframework von Praetorius und Charalambous (2018) beinhaltet die zentralen Aspekte der drei Basisdimensionen und erweitert diese um weitere Punkte, die bereits in der empirischen Forschung Anwendung finden, wie die genannten Punkte zur fachlichen Ergänzung. Es erscheint für diesen Beitrag als auch für das Themenheft zielführender, eine solche Erweiterung zu nutzen und heraus-

zuarbeiten, ob bei dieser erweiterten Systematisierung von Unterrichtsqualität nach wie vor Ergänzungsbedarf besteht.

### *1.3.1 Abgrenzung von Generik und Fachspezifik*

Da im Folgenden eine Unterscheidung zwischen generischen und fachspezifischen Qualitätskriterien vorgenommen wird, soll diese Abgrenzung zunächst näher erläutert werden. Grundlegend orientiert sich diese Unterscheidung, wie auch im Einleitungsbeitrag beschrieben, daran, ob ein Kriterium auf andere Fächer übertragen werden kann, oder ob es spezifisch nur in einem Fach angewendet werden kann. Hierbei muss jedoch bedacht werden, dass die Formulierung einer Dimension oder Subdimension durchaus generisch erfolgen kann, auch wenn die darunterliegenden Ebenen erst durch eine Konkretisierung der Operationalisierung einen fachspezifischen Fokus erhalten. Aus diesem Grund sei auf die Unterscheidung zwischen Generik und Fachspezifik von Wüsten (2010) verwiesen, die auch im Rahmen des Reviews zu mathematisch-naturwissenschaftsdidaktischen Videostudien von Dorfner et al. (2017) angewendet wurde, woran dieser Beitrag anschließt und die somit auch im Folgenden Anwendung findet. Nach dieser Unterscheidung setzt ein generisches bzw. fachunabhängiges Kriterium weder fachliches, noch fachdidaktisches Wissen für eine Beurteilung voraus. Fachspezifische Kriterien hingegen sind abhängig vom unterrichteten Fachinhalt und setzten somit für eine Beurteilung ein fachliches und/oder fachdidaktisches Wissen voraus. Durch diese Unterscheidung wird berücksichtigt, dass eine Dimension, Subdimension oder eine darunterliegende Organisationsebene fachunspezifisch formuliert und somit auf andere Fächer übertragen werden kann, aber trotzdem als fachspezifisch wahrgenommen wird, wenn es um die konkrete Operationalisierung geht. Auf das Syntheseframework bezogen bedeutet dies, dass die Dimensionen und Subdimensionen grundsätzlich generisch formuliert sind und somit auf andere Fächer übertragen werden können. Sobald jedoch ein Aspekt beurteilt wird, der fachspezifisches Wissen voraussetzt, wie z. B. die „Akkuratheit und Korrektheit der thematisierten Inhalte sowie Fachmethoden“, wird die Subdimension als fachspezifisch betrachtet, auch wenn die grundlegende Konzeption eines korrekten Inhalts problemlos auf andere Fächer übertragen werden kann. Andere Dimensionen setzten wiederum kein fachspezifisches Wissen voraus und sind somit fächerübergreifend formuliert und auch inhaltlich generisch, wie z. B. das „Verhaltensmanagement“. Während einige Kriterien als klar fachspezifisch oder generisch beschrieben werden können, lassen sich andere als Konglomerat von fachspezifischen und generischen Anteilen beschreiben, so z. B. das Kriterium der „Klarheit“ (Brunner 2018) oder der „kognitiven Aktivierung“ (Praetorius und Charalambous 2018), da zu einer Beurteilung dieser Kriterien sowohl fachliches oder fachdidaktisches Wissen notwendig ist, als auch ein grundsätzliches pädagogisches Wissen.

Die in Studien genutzten Terminologien, Konzeptualisierungen und Operationalisierungen können sowohl generische als auch naturwissenschaftsdidaktische Anteile tragen (siehe Beispiele im Ergebnisteil). Das zieht es nach sich, dass Kriterien in Videostudien angewendet werden können, die eine terminologische Nähe zu generischen Ansätzen tragen, aber auf der Ebene der Konzeptualisierung und Ope-

rationalisierung unterschiedlich sind. Um diesen Zustand zu bezeichnen, nutzt der vorliegende Beitrag den Begriff der „Perspektivierung“ und meint damit, dass generisch orientierte Konzepte und Terminologien mit besonderem naturwissenschaftsdidaktischem Fokus verknüpft sein und interpretiert werden können. Dabei ist eine Fokussierung von Kriterien in generischer Terminologie auf spezifische Elemente des naturwissenschaftlichen Unterrichts möglich, wie z. B. im Falle der „Klarheit und Strukturiertheit des Experiments“ (Schulz 2011). Aber auch eine Entwicklung von fachspezifischen Kriterien und Terminologien, die wiederum eine konzeptuelle Passung zu generischen Merkmalen enthalten, erscheint möglich.

### *1.3.2 Explizite und implizite Bezüge zur Unterrichtsqualität*

Im Kontext naturwissenschaftsdidaktischer Forschungen können die Bezüge zwischen konkreten Studien und Diskursen zur Unterrichtsqualität sowohl expliziter als auch impliziter Natur sein. So existieren empirische und theoretische Arbeiten, die den expliziten Anspruch erheben, naturwissenschaftsspezifische Unterrichtsqualität abzubilden. Zur Konkretisierung seien hier exemplarisch die Videostudie von Börlin (2012) oder die Sammelbände „Wirksamer Fachunterricht“ (Brovelli 2018; Wilhelm 2018; Rehm 2018) genannt. In diesen Arbeiten werden entweder auf einer kriterienorientiert-empirischen oder auf einer diskursiv-theoretischen Ebene Merkmale von Unterrichtsqualität expliziert.

Für eine Zusammenführung von Unterrichtsqualitätskriterien wäre es jedoch unproduktiv anzunehmen, dass naturwissenschaftsdidaktische Arbeiten – ohne expliziten Bezug zum Diskurs der Unterrichtsqualitätsforschung – keine Kriterien formulierten und anwendeten und damit nicht anschluss- und aussagefähig zur Unterrichtsqualität wären. Anders formuliert: Vermutlich behaupteten nur sehr wenige bis keine Naturwissenschaftsdidaktikerinnen oder Naturwissenschaftsdidaktiker, ihre Arbeiten leisteten keinen Beitrag zu einem Verständnis von Unterrichtsqualität – auch wenn darin Studien zur Unterrichtsqualität nur randständig zitiert werden. Exemplarisch seien die Videostudien zur Basiskonzeptorientierung im Biologieunterricht (Förtsch et al. 2018) oder zur Umsetzung naturwissenschaftlicher Denk- und Arbeitsweisen erwähnt (Nehring et al. 2016). Auch wenn diese Studien sich weitgehend außerhalb von expliziten Diskursen zur Unterrichtsqualität verorten, bestehen implizite Bezüge zur Unterrichtsqualität. Im konkreten Fall dieser beiden Studien lässt sich ableiten, dass ein „guter“ naturwissenschaftlicher Unterricht an den fachspezifischen Basiskonzepten orientiert ist, die im Rahmen der Nationalen Bildungsstandards etabliert worden sind, und naturwissenschaftliche Denk- und Arbeitsweisen epistemologisch adäquat in den Unterricht integriert werden sollten. Für das Inbezugsetzen naturwissenschaftsdidaktischer Kriterien mit dem Syntheseframework ist es dabei wichtig, dass eine Aussage zur Unterrichtsqualität abgeleitet werden kann. Die genaue Abgrenzung, ab wann ein Kriterium eine vergleichbare Aussage zur Unterrichtsqualität macht und somit für den Abgleich mit dem Syntheseframework verwendet werden konnte, wird im Rahmen der methodischen Beschreibung des Reviews dargestellt (Abschn. 3.3.3).

## 2 Ziele und Fragestellungen

Da generisch und naturwissenschaftsdidaktisch orientierte Unterrichtsforschungen bisher selten systematisch in Bezug gesetzt wurden, kann disziplinenübergreifende Kommunikation und die Vergleichbarkeit von Studien erschwert werden. Dies behindert nicht nur Kooperationen in zukünftigen Forschungen, sondern auch die Aggregierung des Standes der Forschung in Reviews oder Meta-Analysen.

An diesem Punkt verortet sich der vorliegende Beitrag. Seine Ziele bestehen darin, eine Übersicht über Kriterien für die Beschreibung und Beurteilung von naturwissenschaftlichem Unterricht zu erarbeiten und in Bezug auf das, auf Praetorius und Charalambous (2018) zurückgehende, Syntheseframework zu setzen. Dazu werden exemplarische Einblicke in ausgewählte Spezifika der Ziele, Inhalte und Methoden der naturwissenschaftlichen Unterrichtsfächer und ein systematisches Review quantitativer Videostudien dargestellt und herausgearbeitet, welche naturwissenschaftsdidaktischen Perspektivierungen und Ergänzung des Syntheseframeworks gewinnbringend sein können.

Angesichts der konzeptuellen Verzahnung naturwissenschaftsdidaktischer Forschungen zur allgemeinen Unterrichtsforschung bei gleichzeitiger fachspezifischer Konkretisierung wird eine vergleichsweise hohe terminologische Passung zwischen naturwissenschaftsdidaktischen Kriterien und dem Syntheseframework erwartet. Diese Erwartung ist insofern begründet, als das Syntheseframework grundsätzlich viele generische Anteile aufweist und für die Beschreibung fachspezifischer Kriterien ebenfalls generische Formulierungen verwendet werden, da diese auf einer eher abstrakten Ebene betrachtet werden. Wird im Bereich naturwissenschaftsdidaktischer Forschung auf einer ebenso abstrakten Ebene gearbeitet, ist eine ähnlich generische Formulierung zu erwarten. Sofern die gereviewten Studien ihre Kriterien auch in der konkretesten fachspezifischen Operationalisierung, z. B. in Form von Beobachtungsitems, präsentieren, wird davon ausgegangen, dass Fachspezifika auftreten können, die auf allgemeiner terminologischer Ebene nicht deutlich werden. Diese Fachspezifika sollen bei der Gegenüberstellung zwischen dem Syntheseframework und der naturwissenschaftsdidaktischen Forschung berücksichtigt werden.

Folgende konkrete Forschungsfragen werden dabei fokussiert:

1. Welche Kriterien von Unterrichtsqualität werden in quantitativen Videostudien naturwissenschaftsdidaktischer Forschungen verwendet und in welchem Umfang lässt sich ein Vergleich der empirisch verwendeten Kriterien zum eher generischen Syntheseframework herstellen?
2. Inwiefern lassen sich aus einem Einblick in Spezifika der Ziele, Inhalte und Methoden der naturwissenschaftlichen Unterrichtsfächer Hinweise dafür ableiten, dass eine weitere Spezifikation des Syntheseframeworks vorgenommen werden kann?

# 3 Methoden

## 3.1 Das Vorgehen im Überblick

Zur Erfassung der Qualitätskriterien naturwissenschaftsdidaktischer Unterrichtsforschung wird der Untersuchung ein systematisches Review quantitativer Videostudien vorangestellt. Mit der Fokussierung auf Videostudien wird eine Eingrenzung vorgenommen, die folgendermaßen zu begründen ist: Bei den quantitativen Videostudien handelt es sich um einen dominanten und vielfach umgesetzten Bereich deutschsprachiger naturwissenschaftsdidaktischer Unterrichtsforschungen. Insbesondere die fachlichen Anteile naturwissenschaftlicher Lehr-Lern-Prozesse lassen sich aufgrund ihrer Komplexität und Vielschichtigkeit (siehe Ausführungen in 4.2 „Ausgewählte Spezifika naturwissenschaftlicher Lernprozesse“) nur eingeschränkt in Befragungen von Schülerinnen und Schülern beurteilen, da für diese Beurteilung ein Wissen notwendig wäre, das über die Stundeninhalte hinausgeht. So existieren Befragungen von Schülerinnen und Schülern im naturwissenschaftsdidaktischen Bereich, z. B. in Kontexten des Schwimmens und Sinkens, die sich auch als prädiktiv valide für den Lernerfolg der Schülerinnen und Schüler erwiesen haben (z. B. Fauth et al. 2014). Jedoch sind die Items dieser Befragungen, hier im Falle der kognitiven Aktivierung dargestellt, die als die fachspezifische Dimension der drei Basisdimensionen beschrieben wird, vergleichsweise allgemein („In our science class, we are working on tasks that I have to think about very thoroughly.“, Fauth et al., S. 8), so dass sie sich für die Zwecke des vorliegenden Beitrags eher weniger eignen. Aspekte, wie die Sachstruktur des Unterrichts oder der epistemologisch adäquate Modelleinsatz, lassen sich trivialerweise erst beurteilen, wenn Lehr-Lern-Prozesse durchschritten sind, die weit über die Inhalte einzelner Stunden hinausgehen, was einer Beurteilung durch Schülerbefragungen entgegensteht. Darüber hinaus operationalisieren quantitative Videostudien Unterrichtsqualitätsmerkmale in Form objektiv vergleichbarer Kriterien. Dabei werden Unterrichtsmerkmale in kurzer, teilweise gar stichpunktartiger Form in Listen oder Katalogen zusammengefasst und für die Beurteilung von Unterricht aufbereitet. Teilweise eignen sich die Kriterien bereits für eine Beurteilung von Unterricht, z. B. in einem Ratingverfahren, teilweise werden sie auch in Form von Beobachtungsitems weiter operationalisiert und konkretisiert. Diese Kriterien sind besonders anschlussfähig an das Syntheseframework von Praetorius und Charalambous (2018), das selbst eine Systematisierung von objektivierbaren Kriterien darstellt. Schließlich wird, für die naturwissenschaftsdidaktische Forschung wertvoll, eine Weiterführung der bisherigen Systematisierung des Standes der naturwissenschaftsspezifischen Unterrichtsforschung erreicht. Um die Erfassung der Qualitätskriterien aus dem Review naturwissenschaftsdidaktischer Videostudien theoretisch abzusichern, werden in einem ersten Schritt aus den beschriebenen Spezifika der Ziele, Inhalte und Methoden naturwissenschaftsspezifische Unterrichtsqualitätsmerkmale abgeleitet (siehe Abschn. 3.2). Aufbauend auf das systematische Review zur „[...] Ausrichtung quantitativer Videostudien [...]“ von Dorfner et al. (2017), werden dann in einem zweiten Schritt die Kriterien quantitativer naturwissenschaftsdidaktischer Videostudien herausgearbeitet und diese Kriterien in eine vergleichende Beziehung zum, auf Praetorius und Charalambous (2018) zurückgehenden, Synthe-

seframework gesetzt (siehe 3.3). Diese Gegenüberstellung der Kriterien mit dem Syntheseframework wird zusätzlich mit einer Doppelkodierung abgesichert (siehe 3.4).

Zur abschließenden Gegenüberstellung der Kriterien aus naturwissenschaftsdidaktischen Videostudien und dem Syntheseframework wird der Begriff der „Perspektivierung“ eingeführt (siehe 3.5). Damit soll deutlich gemacht werden, dass eine Abgrenzung zwischen Generik und Fachspezifik erst auf einer Ebene der stärkeren Operationalisierung deutlich wird, auch wenn bei einer abstrakteren Betrachtung auf Ebene der Dimensionen und Subdimensionen eine terminologische Überschneidung auftritt. Diese Perspektivierung stellt somit den Übergang von der fächerübergreifenden Beschreibung generischer und fachspezifischer Merkmale zu einer fachspezifischen Auslegung derselben dar. Die Übereinstimmungen, die im Rahmen des Vergleichs der Spezifika und empirisch angewendeten Kriterien mit dem Syntheseframework untersucht werden, stellen somit keine Gleichheit der beiden Perspektiven dar, sondern verdeutlichen, dass ein Vergleich beider Perspektiven auf Basis eines gemeinsamen Konzeptes zur Unterrichtsqualität möglich ist. Bei dieser Übereinstimmung ist jedoch nicht auszuschließen, dass bei einer differenzierteren Betrachtung Fachspezifika deutlich werden, was durch die Perspektivierung als Bindeglied zwischen Generik und Fachspezifik deutlich gemacht werden soll.

### 3.2 Berücksichtigung der Spezifika der Ziele, Inhalte und Methoden der naturwissenschaftlichen Fächer

Bei der Identifikation und Ableitung von Qualitätskriterien aus den Spezifika der Ziele, Inhalte und Methoden handelt es sich um ein theoriebasiertes, deduktives Vorgehen, das in den konkreten Fällen der Ergänzung von Kriterien begründet wird. Dabei werden die Spezifika beschrieben, aus denen sich ein Kriterium für Unterrichtsqualität ableiten kann und vor dem Hintergrund ihres Potentials zur Ermöglichung von Lernerfolg im naturwissenschaftlichen Unterricht beurteilt. Anschließend werden sie mit dem Syntheseframework abgeglichen.

Dabei sei explizit darauf hingewiesen, dass es sich – allein schon aus Platzgründen – um ein exemplarisches Vorgehen handelt. Ein theoriebasierter Abgleich von Kriterien mit dem Syntheseframework kann Gegenstand einer eigenen Arbeit sein. Die Einblicke in die Spezifika der Ziele, Inhalte und Methoden der naturwissenschaftlichen Fächer dienen vielmehr dazu, exemplarisch herauszuarbeiten, inwiefern eine naturwissenschaftsdidaktische Theoriearbeit mit dem Syntheseframework weitere Unterrichtsqualitätskriterien gewinnbringend zu Tage fördern könnte. Darüber hinaus soll den Lesenden des vorliegenden Themenheftes, die nicht im Bereich der Naturwissenschaftsdidaktiken tätig sind, eine terminologische Grundlage für das Verständnis des systematischen Reviews gegeben werden. Aus pragmatischen Gründen und um unnötige Redundanzen zu vermeiden, werden die Merkmale beschrieben, die geeignet erscheinen, Dimension oder Subdimension im Syntheseframework zu ergänzen und die nicht schon durch Videostudien benannt werden.

## 3.3 Systematisches Review

### *3.3.1 Literatursuche*

Das systematische Review wurde unter Berücksichtigung der strukturellen Schritte des PRISMA Statements (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Moher et al. 2009) durchgeführt. Hierzu wurde die Checkliste zum PRISMA Statement (Ziegler et al. 2011; Anhang 1) abgearbeitet und die für dieses Review relevanten Punkte systematisch expliziert, wobei für die vollständige Darstellung des Vorgehens auf diesen Beitrag verwiesen wird. Als eine erste Grundlage für die Literaturrecherche diente das systematische Review von Dorfner et al. (2017), welches Videostudien im mathematisch-naturwissenschaftlichen Fachbereich erfasste und deren methodische und inhaltliche Ausrichtung untersuchte. Das Review konnte als Grundlage für das in diesem Artikel beschriebene systematische Review genutzt werden, da sich durch den Fokus auf quantitative Videostudien im (mathematisch-)naturwissenschaftlichen Fachbereich eine gemeinsame Untersuchungsbasis ergab, auch wenn in der weiteren Analyse mit einem anderen Fokus gearbeitet wurde. Die im Artikel von Dorfner et al. (2017) berichteten Publikationen zu Videostudien bildeten somit den ersten Datensatz für das hier beschriebene Review. Weiterhin wurde, um anschlussfähig an diesen Stand der Forschung zu bleiben, dasselbe Suchraster wie im Review von Dorfner et al. (2017) genutzt und die darin berichteten Studien um Studien aus den Jahren 2016 bis 2019 ergänzt (Tab. 1).

Publikationen aus dem Jahr 2016 wurden ebenfalls in der Suche zur Erweiterung des Reviews berücksichtigt, um eine lückenlose Abbildung des Erhebungszeitraumes zu ermöglichen, da das Review von Dorfner et al. (2017) bereits Mitte 2016 erstmalig eingereicht wurde. Studien aus dem Bereich der Mathematik, die bei Dorfner et al. (2017) berücksichtigt waren, wurden für die weitere Analyse ausgeschlossen. Zusätzlich wurde eine Einschränkung auf Forschungsarbeiten aus dem deutschsprachigen Raum vorgenommen, wodurch insgesamt 22 Videostudien bzw. 85 Publikationen zu diesen Videostudien aus dem Review von Dorfner et al. (2017) in die weitere Analyse aufgenommen werden konnten.

Die Einschränkung auf den deutschsprachigen Raum wurde vorgenommen, da die teils unterschiedlichen fachdidaktischen Konzeptionen in den verschiedenen Ländern einen einheitlichen Vergleich mit dem Syntheseframework und auch zwischen den in diesem Themenheft berichteten anderen Fachbereichen erschweren würde. Das kann sich u. a. in unterschiedlichen Schwerpunktsetzungen in den Curricula ausdrücken, was wiederum andere Schwerpunktsetzungen in den Qualitätskriterien bzw. andere Orientierungsrahmen für Qualitätskriterien nach sich ziehen kann, die wiederum in einem Beitrag berücksichtigt werden müssten. Es müsste zunächst auch abgesichert werden, dass durch Übersetzungen keine terminologischen Abweichungen entstehen, die Fehlinterpretationen hervorbringen würden. Diese Perspektive kann in Anschlussarbeiten durchaus weiterverfolgt werden, würde aber den Umfang des vorliegenden Beitrags übersteigen.

Auch wenn die Einbettung des Beitrags in das vorliegende Themenheft eine Fokussierung auf den deutschsprachigen Stand der Forschung nahelegt, wurden ebenso internationale Studien in die erweiterte Suche eingeschlossen und erst in einem letz-

ten Schritt aus dem Suchergebnis entfernt. Dies ermöglicht es auf Basis des Suchergebnisses weiterführende Untersuchungen durchzuführen, bedeutet jedoch auch, dass diese in der quantitativen Darstellung der Suchergebnisse ebenfalls abgebildet sind, was bei einer Replikation der Suche beachtet werden muss. Eine Darstellung der Ergebnisse des internationalen Raums kann in diesem Themenheft, aus Gründen der Vergleichbarkeit zwischen den Beiträgen, zwar nicht umgesetzt werden, ist aber somit perspektivisch möglich.

### *3.3.2 Eingrenzung des Suchergebnisses*

Beim ersten Suchdurchlauf zur Erweiterung des Reviews konnten insgesamt 1174 Publikationen in den in Tab. 1 aufgeführten Datenbanken gefunden werden. Dieses umfangreiche Ergebnis konnte jedoch beim ersten Screening stark eingeschränkt werden. Hierbei wurden zunächst Publikationen aussortiert, die auf Grund ihres Titels oder des Abstracts ausgeschlossen werden konnten. Durch die Übernahme des Suchrasters von Dorfner et al. (2017) wurden beispielsweise auch Publikationen gefunden, die in anderen Fächern als dem für diesen Artikel gesetzten naturwissenschaftsdidaktischen Rahmen durchgeführt wurden (z. B. Mathematik). Um die Suche nicht durch zusätzliche Suchbegriffe zu beeinflussen, wurden diese Artikel erst nach dem ersten Suchdurchlauf aussortiert, sodass nach dem ersten Screening 457 Publikationen in Hinblick auf ihre angewendete Methodik näher betrachtet wur-

**Tab. 1** Suchraster aus Dorfner et al. (2017). Für das Review auf den Zeitraum 2016–2019 angepasst

| Recherchequelle | Suchbegriffe | Trefferzahl |
|---|---|---|
| Pedocs | Frei-/Volltext = Videostudie ODER<br>Frei-/Volltext = Videoanalyse ODER<br>Frei-/Volltext = Unterrichtsbeobachtung<br>Seit 2016 | 133 |
| ERIC | Publication Date: „since 2016“<br>abstract: classroom video analysis OR abstract: video study OR abstract: video tape classroom OR abstract: video data classroom<br>Peer reviewed | 97 |
| PsycINFO | Publication Date: „2016–2019“<br>AB classroom video analysis OR AB video study OR AB video tape classroom<br>Peer reviewed | 11 |
| Taylor & Francis Online | Publication Date: „2016–2019“<br>Publication Title:<br>„video study“ OR Abstract: „classroom video analysis“ OR Abstract: „video study“ OR Abstract: „video tape classroom“ OR Abstract: „video data classroom“ | 20 |
| Wiley Online Library | Publication Date: „2016–2019“<br>Abstract: „classroom video analysis“ OR Abstract: „video study“ OR Abstract: „video tape classroom“ OR Abstract: „video data classroom“ | 60 |
| Springer Link | Find Resources with all of the words: „video“ AND „study“ AND „classroom“<br>with at least one of the words: „tape“ OR „data“ OR „analysis“<br>Show documents published between „2016“ and „2019“<br>Within „Science Education“ | 853 |

den. Neben der bereits genannten Fokussierung auf naturwissenschaftliche Fächer wurden weitere Kriterien herangezogen, um die gefunden Publikationen auf ihre Passung für das Review zu überprüfen: Hierzu gehörte die Bedingung, dass es sich um eine quantitative Videostudie handelt, die in einer Schule durchgeführt wurden (Ausschluss von Exkursionen, Laborsettings, universitärer Lehrer, Kindergärten und Vorschulen). Das Verhalten der Lehrkraft sollte fokussiert werden und es sollte sich um eine vollständig ausgebildete Lehrkraft handeln (Ausschluss von Studierenden und ReferendarInnen, sofern sie nicht die Rolle einer vollständig ausgebildeten Lehrkraft einnahmen und eine reale Unterrichtsstunde bewertet werden konnte). Weiterhin wurden Videostudien ausgeschlossen, in denen das Unterrichtsvideo nicht der Fokus der Untersuchung war, sondern lediglich als unterstützende Quelle zur Bestätigung von Interviewdaten o. ä. genutzt wurde. Viele dieser Kriterien konnten bereits beim zweiten Screening mit einem Blick auf die jeweiligen angewandten Methoden überprüft werden, sodass 111 Publikationen für ein abschließendes Screening des Gesamttextes verblieben. Bei diesem abschließenden Screening wurden die bereits genannten Kriterien erneut angelegt und der Gesamttext der Publikation betrachtet, falls sich bei der Untersuchung der Methoden noch keine konkrete Aussage über die Eignung für das Review treffen ließ. Weiterhin gewährte dieses Vorgehen eine Absicherung der zuvor getroffenen Auswahl. Nach der Betrachtung des Gesamttextes verblieben 51 Publikationen zu quantitativen Videostudien, die den zuvor genannten Kriterien entsprachen. Wie bereits beschrieben wurde das Suchergebnis für die Erhebung der Qualitätskriterien für diese Untersuchung auf den deutschsprachigen Raum begrenzt, was 27 Publikationen einschloss. Durch die Überschneidung des Zeitraums des Reviews von Dorfner et al. (2017) und der erweiterten Suche dieses Beitrags mussten in einem letzten Schritt gefundene Duplikate ausgeschlossen werden. Für den Vergleich zwischen den Kriterien der Videostudien und dem Syntheseframework von Praetorius und Charalambous (2018) verblieben somit 10 Publikationen für den Zeitraum 2016–2019, in denen über 6 zuvor nicht erfasst Videostudien berichtet wurde. In zwei dieser Fälle wird von einer Anbindung an bereits erfasst Videostudien berichtet, da die Publikationen jedoch über den Rahmen einer einzelnen Videostudie hinaus berichten, wurden sie gesondert aufgeführt. Für eine detaillierte Auflistung aller Videostudien und der dazugehörigen Publikationen, sowie der fachlichen Ausrichtung – siehe Anhang 2. Unter Berücksichtigung des Reviews von Dorfner et al. (2017) ergab dies eine Gesamtzahl von 28 Videostudien mit 95 Publikationen, die in das Review eingingen.

### *3.3.3 Herausarbeitung der Qualitätskriterien*

Für die Beantwortung der ersten Fragestellung nach den empirisch angewendeten Qualitätskriterien mussten zunächst die in den jeweiligen Publikationen angegebenen Kriterien identifiziert und herausgearbeitet werden. Hierbei wurden alle Kriterien berücksichtigt, die für die Datengenerierung im Rahmen der Beurteilung von Unterricht Anwendung fanden (z. B. in Form von Kodiermanualen) und somit eine direkte Anbindung an die empirische Forschung aufweisen. Kriterien, die sich auf die Oberflächenstruktur von Unterricht beziehen und keine Bezüge zur Unterrichtsqualität oder zur Nutzung von Unterrichtsangeboten aufwiesen, z. B. „Redeanteile

innerhalb einer Stunde“ oder die Art der „Sachbegegnung“ (Alltagsgegenstände, Computersimulation, Chemische Geräte, Gedankenexperimente, Mischung; Schulz 2011), wurden nicht berücksichtigt. Bei den Beispielen werden Aspekte der Untersuchung benannt, die in der genannten Publikation ohne eine weitere Aussage zur Unterrichtsqualität aufgeführt wurden. Kriterien mit einer inhaltlichen Überschneidung oder ähnlichen Terminologie können durchaus in anderen Publikationen oder in einem anderen fachlichen Diskurs eine Anbindung zur Unterrichtsqualität aufweisen. Wenn dies jedoch nicht in der untersuchten Publikation geschah, wurden sie für das Review nicht als Qualitätskriterien erfasst. Die Erfassung von Oberflächenstruktur ohne eine Anbindung zur Unterrichtsqualität konnte nicht zielführend mit Kriterien verglichen werden, die eine Anbindung zur Unterrichtsqualität aufweisen. Beispielsweise trifft ein Kriterium wie die „Verwendung von Materialien aus dem Alltag“ (Widodo und Duit 2004) eine eindeutige Aussage, welche Art der von Schulz (2011) beschriebenen „Sachbegegnung“ zu bevorzugen ist. Auch für den späteren Vergleich mit dem Syntheseframework, welches auf Unterrichtsqualität fokussiert, ist eine eindeutige Aussage zur Unterrichtsqualität notwendig, damit ein Kriterium beispielsweise mit der Beschreibung „Content is explicitly presented (e.g., by selecting appropriate examples)“ vergleichbar ist. Eine einfache Auflistung der verwendeten Zugänge weist zwar eine inhaltliche Ähnlichkeit auf, es wird jedoch hierbei keine Aussage in Bezug auf einen „guten“ Unterricht getroffen. Wurden Aspekte der Oberflächenstruktur in einem Kontext der Unterrichtsqualität oder Angebotsnutzung verortet, wurden sie – wie z. B. „Organisationformen im Unterricht“ mit einem Hinweis auf die Lernförderlichkeit von kooperativem Lernen (Schulz 2011) – einbezogen. Wichtig für eine Berücksichtigung eines Kriteriums ohne einen expliziten Bezug zur Unterrichtsqualität ist, dass sich eine Handlungs- oder Verhaltensempfehlung für den Unterricht daraus ableiten lässt.

Wurden in den Publikationen einer Studie keine Kodiermanuale zugänglich gemacht, wurden die Kriterien berücksichtigt, die im Text der Publikationen benannt wurden. Zur Aufnahme in die Kodierung war es jedoch notwendig, dass die genannten Qualitätskriterien im Rahmen einer Unterrichtsaufzeichnung Anwendung fanden. Im Rahmen des Reviews wurde jede gefundene Publikation zu naturwissenschaftsdidaktischen Videostudien erfasst und die darin berichteten Qualitätskriterien tabellarisch erfasst (siehe Anhang 3[2]). Da die im Rahmen des Reviews gefunden Videostudien eine sehr unterschiedliche Anzahl an Publikationen aufweisen, wurde die Einschränkung vorgenommen, dass ein Qualitätskriterium für den an das Review anschließenden Vergleich mit dem Syntheseframework lediglich einmalig pro Videostudie erfasst wurde. Sollten Qualitätskriterien innerhalb einer Videostudie mit einer ähnlichen Terminologie, aber inhaltlicher Abweichung in unterschiedlichen Publikationen auftreten, wurden sie für die Videostudie vollständig erfasst. Beispielsweise weisen ein „motivational unterstützender Unterricht“ (Seidel et al. 2006a) und eine „Exploration der Interessen […] der Schüler“ (Widodo und Duit 2004) eine inhaltli-

[2] Anhang 3 kann auf Grund des Umfangs der Tabelle (111 × 124 Felder, inklusive Kommentare) nicht in einer angemessenen Darstellung an diesen Artikel angehängt werden. Um die vollständige Tabelle dennoch für die Lesenden verfügbar zu machen, kann sie auf Nachfrage von den Autoren individuell zur Verfügung gestellt werden.

che Überschneidung auf, nähern sich diesem Aspekt der Unterrichtsqualität jedoch aus unterschiedlichen Richtungen. Beide Beispiele wurden in diesem Fall Publikationen entnommen, die im Rahmen der IPN-Videostudie veröffentlicht wurden und beide Qualitätskriterien wurden für den Vergleich mit dem Syntheseframework genutzt. Sollte hingegen in unterschiedlichen Publikationen zu derselben Videostudie unterschiedliche Terminologien für dasselbe inhaltliche Konstrukt verwendet werden, wurde lediglich eine der Ausführungen erfasst – dieser Fall ist vor allem dann eingetreten, wenn die Ergebnisse einer Videostudie mehrsprachig veröffentlicht wurden. Als Beispiel hierfür kann die „lebensweltliche Einbettung" (Börlin 2012) genannt werden, welche in einer englischen Übersetzung als „embedment of everyday-life" (Börlin und Labudde 2014) ebenfalls mit Bezug zur QuIP-Studie beschrieben wird, aber auch in der Variation „everyday-life context" (Beerenwinkel und Arx 2016) im Rahmen derselben Videostudie verwendet wird. In diesem Fall wurde lediglich die „lebensweltliche Einbettung" als Qualitätskriterium der QuIP-Studie mit dem Syntheseframework verglichen. Die Anzahl der Veröffentlichungen zu einer Videostudie unterschied sich in einem Rahmen von einer bis 26 Publikationen. Innerhalb einer Videostudie weisen die Publikationen häufige Überschneidungen und Querverweise zu Publikationen oder Kodiermanualen derselben Videostudie auf. Bei einer quantitativen Darstellung der erhobenen Qualitätskriterien und deren Verteilung auf das Syntheseframework würde das Gesamtbild somit durch Videostudien mit besonders vielen Publikationen sehr stark geprägt. Um zu vermeiden, dass eine einzelne Videostudie das Gesamtbild beim Vergleich der Videostudien mit dem Syntheseframework zu stark beeinflusst, wurde die Einschränkung vorgenommen, dass pro Videostudie lediglich ein Qualitätskriterium aufgeführt wird, wenn es eine vollständige inhaltliche Überschneidung zwischen mehreren Qualitätskriterien gibt.

#### *3.3.4 Vergleich der Qualitätskriterien mit dem Syntheseframework*

Der Beitrag steht vor der grundlegenden Herausforderung Kriterien zur Erfassung der Unterrichtsqualität aus unterschiedlichen Bereichen gegenüberzustellen. Für einen Vergleich von Qualitätskriterien aus der naturwissenschaftsdidaktischen Forschung mit dem Syntheseframework ist jedoch zunächst eine Analyse der jeweiligen Hierarchisierungen notwendig. Das Syntheseframework besteht aus Dimensionen („höchste und allgemeine Ebene"), Subdimensionen („mittlere Ebene") und einer Indikatorenebene („unterste und detaillierteste Ebene"). Im Fall des Syntheseframeworks bedeutet dies, dass ein Vergleich auf der Indikatorenebene den besten Zugang bietet und für die naturwissenschaftsdidaktische Forschung ist dies, abhängig von der jeweiligen Publikation, würde jedoch im Optimalfall die Betrachtung von Beobachtungsitems bedeuten. Wie bereits in Abschn. 1.2.1 angedeutet, können bei diesem Vergleich jedoch Schwierigkeiten auftreten. Zunächst kann es vorkommen, dass in einer Publikation zu einer Videostudie keine Beobachtungsitems, sondern direkt Kriterien zur Erfassung der Unterrichtsqualität genannt werden. In diesem Fall wurde mit der Beschreibung in der jeweiligen Publikation gearbeitet, um die inhaltliche Ausrichtung des Kriteriums herauszustellen und einen Vergleich mit dem Syntheseframework zu ermöglichen. Damit eine einheitliche Darstellung auf der Seite der Naturwissenschaftsdidaktik erfolgen konnte, wurde der Vergleich zum Synthese-

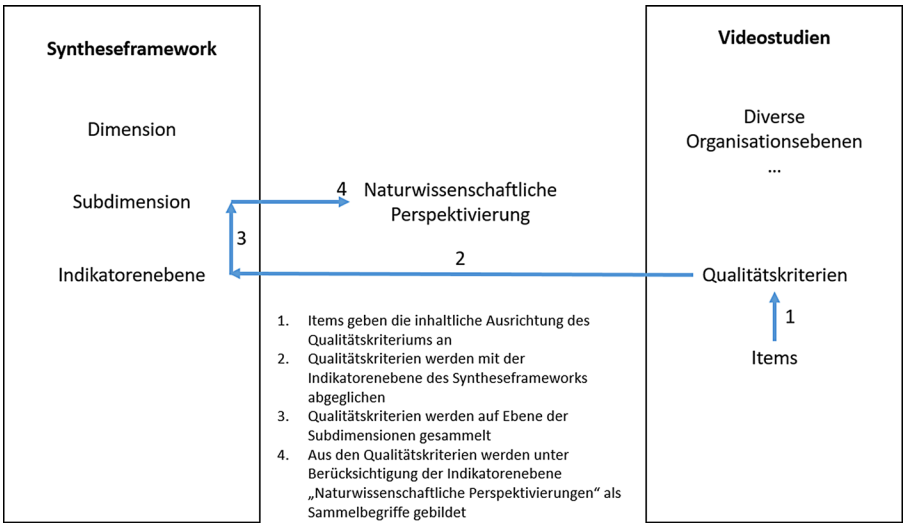



**Abb. 1** Ebenen des Vergleichs zwischen Syntheseframework und Qualitätskriterien naturwissenschaftsdidaktischer Videostudien

framework grundsätzlich mit den jeweils verwendeten Kriterien dargestellt. Sofern Beobachtungsitems vorlagen, gaben sie die inhaltliche Ausrichtung eines Qualitätskriteriums an und wurden somit direkt berücksichtigt.

Eine weitere mögliche Schwierigkeit besteht dann, wenn ein Kriterium inhaltlich einen größeren Bereich abdeckt als die Indikatoren des Syntheseframeworks. Das Syntheseframework wird deshalb zwar grundlegend auf der Indikatorenebene betrachtet, es wird jedoch die Möglichkeit geboten Kriterien aus der naturwissenschaftsdidaktischen Forschung auch auf Ebene der Subdimensionen zu vergleichen, sofern dies durch den verwendeten Grad der Abstraktion notwendig ist. Um abschließend einen einheitlichen Vergleich der zugeordneten Kriterien zu ermöglichen, werden die Ergebnisse auf Ebene der Subdimensionen zusammengefasst, welche auch für den fächerübergreifenden Vergleich genutzt werden.

Insgesamt wird somit eine einheitliche Ebene der Abstraktion zur Beschreibung von Unterrichtsqualität geboten, die einen möglichst detaillierten Vergleich zwischen den unterschiedlichen Ansätzen erlaubt, ohne dabei zwischen den hierarchischen Ebenen zu wechseln.

Für den Vergleich der in den Videostudien verwendeten Qualitätskriterien mit dem generischen Syntheseframework wurde, ausgehend von der Beschreibung des Syntheseframeworks (Praetorius und Charalambous 2018) gearbeitet. Die herausgearbeitete inhaltliche Ausrichtung eines naturwissenschaftsdidaktischen Qualitätskriteriums wurde mit den Subdimensionen bzw. der Indikatorenebene des Syntheseframeworks abgeglichen und zunächst tabellarisch gegenübergestellt (siehe Anhang 3).

Abb. 1 stellt die Schritte zum Vergleich der Kriterien in ihrer Gesamtheit dar. Im Fall von fehlenden Beobachtungsitems entfiel Schritt 1 und im Fall eines höheren Abstraktionsgrades eines Qualitätskriteriums beinhaltet Schritt 2 eine direkte Zu-

ordnung auf Subdimensionsebene. Da bei der Beschreibung von Qualitätskriterien der naturwissenschaftsdidaktischen Forschungen davon ausgegangen werden kann, dass unter anderem fachspezifische Formulierungen verwendet werden, sollen diese abschließenden in Schritt 4 ebenfalls berücksichtigt werden. Die Beschreibung zur Bildung dieser naturwissenschaftlichen Perspektivierung wird in Abschn. 3.5 dargestellt.

Für den Vergleich wurde die englische Version des Syntheseframeworks von Praetorius und Charalambous (2018) verwendet, weshalb diese auch in Tab. 3 dargestellt wird. Eine deutsche Übersetzung lag zu Beginn der Arbeit nicht vor, wird jedoch in einer erweiterten Darstellung in Tab. 4 und im Rahmen des Einleitungsbeitrags berichtet. Da die deutsche Übersetzung bereits eine inhaltliche Erweiterung der Dimensionen und Subdimensionen beinhaltet, wird für die Darstellung des ersten Vergleichs mit dem Syntheseframework die englische Version verwendet. Bei der Übersetzung und inhaltlichen Erweiterung sind bereits Ergebnisse des ersten Vergleichs mit dem Syntheseframework eingeflossen und Abweichungen, die zuvor zwischen dem Syntheseframework und den naturwissenschaftsdidaktischen Qualitätskriterien bestanden, wurden dadurch teilweise behoben, wodurch die in Tab. 3 dargestellten Abweichungen nicht vollständig auf das Erweiterte Syntheseframework übertragen werden können.

Die Gegenüberstellung der Qualitätskriterien mit dem Syntheseframework erfolgte anhand eines dreistufigen Kategoriesystems. Hierbei wurde der Grad der Vergleichbarkeit zum Syntheseframework bei jedem Qualitätskriterium analysiert. Die drei Stufen der Vergleichbarkeit wurden als „vollständige Übereinstimmung“, „teilweise Übereinstimmung“ und „starke Abweichung“ festgelegt.

Eine „vollständige Übereinstimmung“ schließt alle Qualitätskriterien ein, die inhaltlich im Syntheseframework wiederzufinden sind oder einen Teil der Indikatorenebene einer Subdimension des Syntheseframeworks erfüllen. Somit werden in dieser Kategorie auch Qualitätskriterien erfasst, die ein weniger komplexes Konstrukt beschreiben, als es in einer Subdimension des Syntheseframeworks der Fall ist – z. B. weist „goal-clarity“ (Börlin und Labudde 2014) eine „vollständige Übereinstimmung“ mit „Presenting the content in a structured way“ auf, da die Indikatorenebene der Subdimension des Syntheseframeworks „Lesson objectives are clear“ (siehe Anhang 3) enthält. Das Syntheseframework und die Beschreibung des Qualitätskriteriums stimmen hierbei in ihrer inhaltlichen Auslegung einer Zielklarheit als Teil der guten Struktur einer Unterrichtsstunde überein, auch wenn im Syntheseframework in diesem Bereich noch weitere Kriterien benannt werden.

Eine „teilweise Übereinstimmung“ umfasst alle Fälle, in denen Qualitätskriterien aus Studien entweder durch mehrere unterschiedliche Subdimensionen des Syntheseframeworks abgedeckt werden, und damit nicht klar zuzuordnen sind oder zusätzliche Aspekte beinhalten, die im Syntheseframework in dieser Form nicht abgedeckt sind. So könnten in einigen Fällen Items oder der beschriebene Inhalt eines Qualitätskriteriums auch auf mehrere Subdimensionen des Syntheseframeworks verteilt werden, so dass mehrfache Zuordnungen vorgenommen werden müssten. Als Beispiel können hierfür die „Anker und Integrationshilfen“ (Herweg 2008) genannt werden, die eine Übereinstimmung zu der Subdimension „Presenting the content in a structured way“ (Syntheseframework) durch die Verknüpfung zu „Lernzielen“

und dem „roten Faden“ aufweisen. Aber auch eine Übereinstimmung zu „Teacher facilitation of students’ cognitive activity“ (Syntheseframework) aufweisen, da eine „Verknüpfung mit Vorwissen“ stattfinden soll. In diesen Fällen wurde das Qualitätskriterium leidglich der Subdimension zugeordnet, bei der die meisten Items eine Passung aufwiesen bzw. in der die Beschreibung innerhalb der Publikation am ehesten treffend schien. Im zweiten Fall der „teilweisen Übereinstimmung“ kann ein Qualitätskriterium zwar prinzipiell einer Subdimension zugeordnet werden, enthält jedoch weitere Aspekte, die nicht im Syntheseframework abgebildet sind. Diese können hierbei sowohl fachspezifisch, als auch generisch sein, geben jedoch in beiden Fällen einen Hinweis auf eine mögliche Ergänzungsstelle des Syntheseframeworks. Als Beispiel für diesen Fall kann das Qualitätskriterium „Vernetzung“ (Glemnitz 2007) genannt werden, bei dem sich zwar eine Überschneidung zu „Presenting the content in a structured way“ (Syntheseframework) ergibt, jedoch zusätzlich zur einfachen Verbindung der Inhalte auch das Niveau der Vernetzung untersucht wird, was im Syntheseframework nicht weiter ausgeführt wird.

Um keine artifizielle Erhöhung der Anzahlen „teilweiser Übereinstimmung“ zu erzeugen, wurde für jedes Qualitätskriterium der Videostudien nur eine Zuordnung auf Subdimensionsebene vorgenommen. Das betreffende Qualitätskriterium wurde dann innerhalb der Kodiertabelle mit einem Hinweis auf nicht enthaltene Inhalte oder auf Inhalte aus anderen Subdimensionen versehen. Die Frage nach der Fachspezifik von Kriterien wurde im Rahmen der „Perspektivierung“ von Kriterien weiterverfolgt.

Die dritte Abstufung „starke Abweichung“ weist wiederum auf explizite Ergänzungsstellen hin. Kriterien, die dieser Abstufung zugeordnet werden, können in keiner der Subdimensionen des Syntheseframeworks verortet werden und wurden lediglich auf Dimensionsebene eingeordnet. Als ein Beispiel hierfür kann der „Evolutionärer Umgang mit Schülervorstellungen“ (Cauet 2015) genannt werden. Hierbei wird im Gegensatz zum Syntheseframework nicht eine Korrektur bestehender Vorstellung beschrieben, sondern ein aufbauender Prozess, der an die bestehenden Vorstellungen anknüpft. Dieser fachspezifische Prozess ist in dieser Form nicht im Framework enthalten, wodurch sich die inhaltliche Abweichung ergibt.

Dieser Vergleich auf den Ebenen der „Übereinstimmung“ bildet somit ab, welche Qualitätskriterien in Videostudien der naturwissenschaftlichen Fächer angewendet werden und welche sich bereits inhaltlich im Syntheseframework wiederfinden lassen.

### 3.4 Absicherung der Objektivität durch Doppelkodierung im systematischen Review

Um sicherzustellen, dass – angesichts des interpretativen Anteils der Zuordnungen – die Verortung der Qualitätskriterien im Syntheseframework mit einer ausreichenden Objektivität erfolgte, wurde eine Doppelkodierung durch einen zweiten Rater durchgeführt. Die Doppelkodierung umfasste 17 % der insgesamt erfassten Kriterien, was 65 Kriterien entspricht. Die Auswahl der Kriterien erfolgte hierfür zufällig, wobei jeweils ganzheitliche Publikationen zu verschiedenen Videostudien aus verschiedenen Fachbereichen gewählt wurden. Dies umfasste 3 Publikationen aus den Fachbereichen Chemie, Physik und Sachunterricht, welche in Anhang 2 markiert

**Tab. 2** Ergebnis der Doppelkodierung

| | Anzahl der Übereinstimmungen | Anzahl der Abweichungen | Anzahl der Kriterien | Prozentuale Übereinstimmung (%) | Anzahl der möglichen Zuordnungs-Dimensionen | Cohens Kappa |
|---|---|---|---|---|---|---|
| Dimension | 54 | 11 | 65 | 83,1 | 7 | 0,78 |
| Sub-Dimension | 45 | 20 | 65 | 69,2 | 20 | 0,66 |
| Abstufung | 45 | 20 | 65 | 69,2 | 47 | 0,67 |

sind. Als Maße der Objektivität wurden sowohl die relative Beurteilerübereinstimmung als auch Cohens Kappa betrachtet. Nach Wirtz und Caspar (2002) weisen die Richtwerte für Cohens Kappa zwar je nach Autor unterschiedliche Werte auf, können jedoch auf die Werte ($0{,}60 < \kappa < 0{,}75$) für den Bereich guter Übereinstimmung und ($\kappa > 0{,}75$) für den Bereich sehr guter Übereinstimmung zusammengefasst werden. Die Ergebnisse der Doppelkodierung (Tab. 2) lassen darauf schließen, dass die Einordnung der Kriterien in das Syntheseframework mit einer angemessenen Objektivität erfolgte. Die Doppelkodierung wurde für eine präzisere Analyse auf drei Facetten der Zuordnungen betrachtet. Zuerst wurde die Übereinstimmung bezüglich der Zuordnung von Kriterien auf Ebene der sieben Dimensionen des Syntheseframeworks überprüft, anschließend die Zuordnung auf Ebene der Subdimensionen und abschließend die Übereinstimmung bei den zuvor beschriebenen Abstufungen der Zuordnungen (3.3.4). Bei den beschriebenen Ebenen ist zu beachten, dass eine gemeinsame Zuordnung zu einer tieferen Ebene z. B. „Abstufung in der Übereinstimmung“ voraussetzt, dass eine gemeinsame Zuordnung in den darüber liegenden Ebenen (Dimension & Subdimension) stattgefunden hat. Die Rater können nicht dieselbe „Abstufung der Übereinstimmung“ gewählt haben, wenn sie zuvor nicht dieselbe Subdimension gewählt haben und ebenso nicht dieselbe Subdimension, wenn sie zuvor nicht dieselbe Dimension gewählt haben.

Der Wert für Cohens Kappa lag im ersten Fall der Zuordnung der Qualitätskriterien (Dimensionen) bei 0,78 woraus geschlossen werden kann, dass die Rater die Dimensionen in weiten Teilen gleich interpretierten. Im zweiten (Subdimensionen) und dritten Fall (Abstufungen) ließ sich zunächst feststellen, dass die absolute Anzahl der Übereinstimmungen in beiden Fällen gleich groß war. Hieraus folgt, dass die Rater bei einer einheitlichen Zuordnung zu einer Subdimension auch dieselbe Abstufung wählten, da eine gemeinsame Abstufung voraussetzt, dass zuvor dieselbe Subdimension gewählt wurde.

Die Zuordnung zu einer Subdimension liegt mit einem Wert für Cohens Kappa von 0,66 in einem Bereich guter Übereinstimmung, woraus hierfür ebenfalls eine weitgehend einheitliche Interpretation der Subdimensionen abgeleitet werden kann. Dies kann ebenfalls für die Zuordnung zu einer Abstufung mit einem Wert für Cohens Kappa von 0,67 angenommen werden. Die Abweichungen zwischen den Ratern wurden für eine bessere Analyse der Doppelkodierung detailliert betrachtet. Hierbei stellte sich heraus, dass in 18 von 20 Fällen einer Abweichung durch mindestens einen der beiden Rater die Abstufung „teilweise Übereinstimmung“ oder „starke Abweichung“ gewählt wurde. Aus den Definitionen dieser Abstufungen (3.3.4) geht

hervor, dass diese einen interpretativen Anteil enthalten, welcher von den Ratern dementsprechend unterschiedlich ausgelegt werden kann. Trotz dieser interpretativen Anteile lässt das Ergebnis der Doppelkodierung auf eine grundlegende Objektivität der Zuordnung schließen. Abweichungen zwischen den Ratern wurden im Anschluss an die Doppelkodierung diskutiert, wobei ein besonderer Fokus auf den interpretativen Anteil der Abstufungen und das Verständnis der Subdimensionen des Syntheseframeworks gelegt wurde. Abweichende Interpretationen wurden hierbei herausgearbeitet und in einem Konsensverfahren vereinheitlicht, sodass die gemeinsame Interpretation beider Rater in nachfolgenden Kodierungen berücksichtigt werden konnte.

### 3.5 Bildung „naturwissenschaftsdidaktischer Perspektivierungen“ von Unterrichtsqualität

Die Bildung von „naturwissenschaftsdidaktischen Perspektivierungen“ wurde, angesichts der in 1.3.1 beschriebenen möglichen terminologischen und konzeptuellen Unterschiede und Gemeinsamkeiten, als abschließender Schritt der Gegenüberstellung der Qualitätskriterien aus den Videostudien mit dem Syntheseframework durchgeführt. Die grundlegende Idee der Perspektivierung ist es, ein Bindeglied zwischen den Kriterien der naturwissenschaftsdidaktischen Unterrichtsforschung und der generisch formulierten Systematisierung des Syntheseframeworks zu bilden. Durch diese Verbindung soll eine bessere Vergleichbarkeit beider Perspektiven hergestellt und die Kommunikation erleichtert werden. Für die Bildung der naturwissenschaftlichen Perspektivierung wurde die Indikatorenebene des Syntheseframeworks genutzt und die Formulierungen überarbeitet, um die fehlenden Aspekte der naturwissenschaftsdidaktischen Qualitätskriterien zu erfassen. Dies terminologische Überarbeitung berücksichtigte sowohl die Formulierungen der Qualitätskriterien aus den Videostudien, als auch die Spezifika der naturwissenschaftlichen Fächer, die in Abschn. 4 näher erläutert werden. Die Formulierungen wurden durch einen ersten Rater erstellt und anschließend durch einen zweiten Rater überarbeitet. Die abschließend gewählte Formulierung der naturwissenschaftlichen Perspektivierung wurde dann in einem Konsensverfahren festgelegt.

Für das genaue Vorgehen bedeutet dies, dass an das systematische Review angeschlossen wurde, indem die Zuordnung der Qualitätskriterien zur Indikatorenebene des Syntheseframeworks (siehe Anhang 3) als Basis für die inhaltliche Sammlung der Kriterien genutzt wurde. In dieser Tabelle wurden die Qualitätskriterien aus den Studien bereits unter bestimmten Sammelbegriffen des Syntheseframeworks zusammengefasst, wie z. B. „Logische Sequenzierung der Unterrichtsinhalte“ (Syntheseframework), die teilweise bereits eine Überschneidung zu den (generischen) Terminologien der in den Videostudien verwendeten Qualitätskriterien aufwiesen. Für eine möglichst umfangreiche Abbildung der Qualitätskriterien aus den Videostudien, wurden auch im Fall eher generischer Kriterien neue Formulierungen für die Perspektivierung gewählt, damit alle Aspekte der Qualitätskriterien abgedeckt werden können, wie z. B. „Strukturierter Ablauf und Sequenzierung der Stunde“ (Perspektivierung). Diese eher generischen Perspektivierungen können somit von der Terminologie des Syntheseframeworks abweichen, auch wenn die dazugehörigen Kriterien

zuvor mit einer „vollständigen Übereinstimmung“ in das Syntheseframework eingeordnet wurden. Darüber hinaus wurden die Qualitätskriterien, die zuvor mit einer „teilweisen Übereinstimmung“ oder „starken Abweichung“ kodiert wurden, genauer betrachtet, um die fehlenden Aspekte des Syntheseframeworks terminologisch in die Perspektivierung aufzunehmen. Es wurden somit Begriffe gewählt, die naturwissenschaftsdidaktische Aspekte hervorheben, die zuvor nicht ausreichend durch die generische Formulierung des Syntheseframeworks abgedeckt wurden, wie z. B. „Auswahl naturwissenschaftlicher Denk- und Arbeitsweisen“ (Perspektivierung), im Vergleich zu „Auswahl von bedeutungsvollen, dem Lernstand angemessenen Inhalten sowie Fachmethoden“ (Syntheseframework) oder „Konstruktive Einbindung von eigenen Ideen und Schülervorstellungen in den Unterricht“ (Perspektivierung) im Vergleich zu „Entwicklung und Revision von Konzepten“ (Syntheseframework).

Hierarchisch liegt die Indikatorenebene, auf der die naturwissenschaftsdidaktischen Qualitätskriterien zunächst gesammelt wurden, unter der Subdimensionsebene. Die Perspektivierung wird jedoch bewusst neben der Subdimensionsebene dargestellt, um zu verdeutlichen, dass die Qualitätskriterien der naturwissenschaftsdidaktischen Forschung zwar eine grundsätzliche Vergleichbarkeit zum Syntheseframework aufweisen, aber nicht generell vollständig durch dasselbe abgebildet werden können und somit nicht einfach eine fachliche Auslegung eines generischen Aspekts darstellen. Die naturwissenschaftsdidaktische Perspektivierung hat hierbei nicht den Anspruch das Syntheseframework zu ersetzen oder eine Erweiterung der Indikatorenebene vorzunehmen, sondern Bezüge zwischen der generischen und der naturwissenschaftsdidaktischen Perspektive herzustellen.

Um abzubilden, in welchem Umfang die naturwissenschaftsdidaktischen Perspektivierungen fachspezifische Aspekte beinhalten, wurde für jede Perspektivierung eine Markierung des Verhältnisses zwischen Fachspezifik und Generik vorgenommen. Diese Markierung erfolgte vergleichbar zum Syntheseframework von Praetorius und Charalambous (2018) und enthält die Kategorien:

1. Generisch (G): Die Aspekte der Perspektivierung sind eher generisch und können ohne fachspezifisches Wissen umgesetzt oder bewertet werden (z. B. Allgegenwärtigkeit).
2. Fachspezifisch (S): Die Aspekte der Perspektivierung sind eher fachspezifisch und setzen ein fachspezifisches Wissen zur Umsetzung oder Bewertung voraus (z. B. Auswahl von Fachinhalten).
3. Generisch mit fachspezifischen Anteilen (G + S): Die Aspekte der Perspektivierung sind großteils generisch, fachspezifische Anteile unterstützen die Umsetzung oder Bewertung jedoch (z. B. Zielklarheit).
4. Verknüpfung generischer und fachspezifischer Aspekte (G*S): Die Aspekte der Perspektivierung haben sowohl generische, als auch fachspezifische Anteile und eine Berücksichtigung beider Perspektiven ist zur Umsetzung oder Bewertung notwendig (z. B. Auswahl herausfordernder Lerngelegenheiten)

Zur Zuordnung der Perspektivierungen zu den unterschiedlichen Kategorien der Fachspezifik wurde zunächst die Kodierung der Subdimensionen aus dem Syntheseframework von Praetorius und Charalambous (2018) auf die Perspektivierung übertragen. Die Perspektivierungen einer Subdimension erhielten somit in diesem Schritt

jeweils dieselbe Kategorie der Fachspezifik. Um die Perspektivierungen innerhalb einer Subdimension präziser zuzuordnen und ggf. Abweichungen zur Zuordnung im Syntheseframework herauszustellen, wurde in einem nächsten Schritt, ausgehend von den zugeordneten Kriterien zu einer Perspektivierung der Grad der Fachspezifik ermittelt. Dies erfolgt auf Basis der Beschreibungen oder Items, die zu den jeweiligen Kriterien innerhalb der Videostudien vorlagen. Für den Fall, dass auf Basis der Kriterien keine genaue Zuordnung zu einer der 4 Kategorien der Fachspezifik getroffen werden konnte, wurde mit Hilfe der theoretischen Ausarbeitung zu den Spezifika der naturwissenschaftlichen Fächer eine Auswahl getroffen. Als Grundlage für die Unterscheidung zwischen Generik und Fachspezifik diente die in Abschn. 1.3.1 beschriebene Abgrenzung. Die Zuordnung zu den Kategorien der Fachspezifik wurde zunächst durch einen Rater getroffen und anschließend mit einem zweiten Rater besprochen. Die abschießende Zuordnung der Perspektivierungen zu diesen Kategorien erfolgte dann in einem Konsensverfahren zwischen beiden Ratern.

Bei der Zuordnung und Bestimmung des fachspezifischen Anteils der Perspektivierungen ergaben sich unterschiedliche Typisierungen des Verhältnisses von Fachspezifik und Generik. Diese werden im Diskussionsteil (6.2) weiter ausgeführt, wurden jedoch für die Darstellung in Tab. 4 und Anhang 4 nicht berücksichtigt.

Damit der Prozess der Bildung und Kategorisierung der Perspektivierung nachvollziehbar und transparent bleibt, werden in Anhang 4 sämtliche Kriterien berichtet, die unter einer Perspektivierung zusammengefasst wurden. Die naturwissenschaftsdidaktischen Perspektivierungen, sowie die genaue Anzahl der Qualitätskriterien, die zu einer Perspektivierung zusammengefasst wurden sind Tab. 4 zu entnehmen.

## 4 Einblicke in ausgewählte Spezifika der Ziele, Inhalte und Methoden der naturwissenschaftlichen Unterrichtsfächer

### 4.1 Zieldimensionen naturwissenschaftlichen Unterrichts

Zentral für aktuelle nationale und internationale Diskurse über die Ziele schulischen naturwissenschaftlichen Lernens ist der Begriff „scientific literacy“ (Bybee 1997; Gräber et al. 2002; Osborne 2007). Zu dessen Kern gehören das Verfügen und Anwenden fachlich adäquater Vorstellungen, die aktive und auf naturwissenschaftlichem Verständnis beruhende Teilhabe an gesellschaftlich relevanten Diskussionen und Entscheidungsprozessen, Kenntnisse epistemologischer Merkmale der Naturwissenschaften, die Fähigkeit, zu einer ökonomisch bedeutsamen Innovationsfähigkeit eines Landes beizutragen wie auch positive Einstellungen gegenüber den Naturwissenschaften selbst und naturwissenschaftlich-technischen Berufen (siehe dazu insb. Kind 2013; oder Gräber et al. 2002).

Eine grundlegende Kategorisierung von vier Zieldimensionen naturwissenschaftlichen Unterrichts wurde, vor diesem Hintergrund, durch Hodson (2014, S. 2537) vorgelegt:

1. *Learning science:* Der Erwerb von Wissen und Entwicklung naturwissenschaftlich adäquater Vorstellungen über Theorien und Modelle auf einer theoretischen Ebene als Zieldimension von Unterricht.
2. *Learning about science:* Die Entwicklung von epistemologisch adäquaten Vorstellungen über das Wesen der Naturwissenschaften als Zieldimension von Unterricht. Dazu gehören Einsichten in die Sicherheit, aber auch die Grenzen und Veränderbarkeit naturwissenschaftlichen Wissens und dessen Generierung und Überprüfung in naturwissenschaftlicher Forschung. Im internationalen Diskurs hat sich hierfür der Begriff „nature of science" weitgehend durchgesetzt.
3. *Doing science:* Der Erwerb von Fähigkeiten zur Umsetzung naturwissenschaftlicher Denk- und Arbeitsweisen. Dazu zählt die aktive und eigenständige Umsetzung von Experimenten, Beobachtungen oder die Nutzung von Modellen als Forschungswerkzeuge. Schülerinnen und Schüler sollen danach in die psychomotorischen Aktivitäten eingebunden werden und selbst naturwissenschaftlich untersuchend tätig sein („hands on"), aber auch damit verbundene Denkprozesse (z. B. beim Bilden von Hypothesen oder bei der Interpretation von Daten) nachvollziehen („minds on").
4. *Addressing socio-scientific issues (SSIs):* Die Entwicklung einer naturwissenschaftlich fundierten Kritikfähigkeit an gesellschaftlichen, ökonomischen, ökologischen, politischen und persönlichen Problemstellungen und Entscheidungsprozessen als Zieldimension von Unterricht. Die Schülerinnen und Schüler sollen insbesondere an Schnittstellen zwischen Naturwissenschaften, Technik und Gesellschaft die Bedeutung von Wissen kennenlernen und für fundierte Entscheidungen nutzbar machen.

In Anlehnung an die Diskussionen um „scientific literacy" hat sich in Deutschland der Begriff der naturwissenschaftlichen Grundbildung weitgehend etabliert und wurde in den nationalen Bildungsstandards der KMK (2005a, 2005b, 2005c) verankert, wobei sich eine weitgehende Kongruenz zwischen den vier Zieldimensionen und den deutschen Kompetenzbereichen feststellen lässt, auf die aus Platzgründen nicht eingegangen werden kann.

### 4.2 Ausgewählte Spezifika naturwissenschaftlicher Lernprozesse

Zwar lassen sich Inhalte der naturwissenschaftlichen Fächer anhand konkreter, lebensweltlicher oder fachlicher Phänomene motivieren und ganzheitlich erlebbar machen (Muckenfuss 2001; Wagenschein 1976). Das Wesen und der Erfolg der Naturwissenschaften als wissenschaftliche Disziplinen bestehen ja gerade aber darin, von diesen Phänomenen zu abstrahieren (Talanquer 2011; Treagust et al. 2003) und eine Nutzung von naturwissenschaftlichen Modellvorstellungen anzubahnen, die diese Phänomene beschreiben, erklären oder vorhersagen können. Für diese Unterscheidung existieren eine Fülle von Begrifflichkeiten in den verschiedenen naturwissenschaftsdidaktischen Disziplinen, wie z. B. die der „Erfahrungswelt", den Sinnen zugänglichen Bereichen der Naturphänomene, und der „Modellwelt", den Denkbereichen der naturwissenschaftlichen Theorien, wie sie von Mikelski-Seifert und Fischler (2003) vorgeschlagen wurden.

Häufig steht naturwissenschaftlicher Unterricht vor der Anforderung, an phänomenorientierten Erfahrungen der Schülerinnen und Schüler anzuknüpfen oder diese Erfahrungen, z. B. in Versuchen oder Experimenten, gezielt zu stiften („Erfahrungswelt"), gleichzeitig aber so davon zu abstrahieren, dass fachlich tragfähiges Theoriewissen („Modellwelt") auf- und ausgebaut werden kann – wobei dieser Auf- und Ausbau wiederum den Umgang mit fachspezifischen und teilweise komplexen Repräsentationen einfordert. Mit Blick auf die Frage der exemplarischen Erweiterung des Syntheseframeworks sei auf eine Auswahl von weiteren Charakteristika verwiesen, die angesichts der Fülle an Facetten und Forschungsarbeiten nicht abschließend sein kann:

1. Die Schülerinnen und Schüler kommen mit zahlreichen Erfahrungen und individuellen Vorstellungen über die Phänomene der Natur in den Unterricht. Diese Erfahrungen bilden den Ausgangspunkt ihrer Lernprozesse (Gropengießer und Kattmann 2013) und bleiben bei Nichtberücksichtigung im Unterricht weiterbestehen. Verstärkt werden diese Prozesse durch die Problematik der Polysemie, bei der Fachwörter andere Bedeutungen haben als im Alltag („Wärme", „Reaktion", „Teilchen"). In der Erforschung und Weiterentwicklung von Schülervorstellungen finden naturwissenschaftsdidaktische Forschungen eines ihrer zentralen Themenfelder; für jede der drei naturwissenschaftlichen Fächer liegen Überblickswerke zu Schülervorvorstellungen vor (Kattmann 2015; Barke 2006; Schecker et al. 2018). Vor dem Hintergrund eines konstruktivistischen Lernverständnisses greift guter naturwissenschaftlicher Unterricht diese Schülervorvorstellungen auf und macht sie, wie z. B. im Rahmen der Arbeiten zur didaktischen Rekonstruktion (Duit et al. 2012) beschrieben, zum Gegenstand der Planung und Durchführung.
2. Phänomene bilden häufig den Ausgangspunkt für die Erarbeitung von fachlich adäquaten Vorstellungen der „Modellwelt", jedoch können verschiedene Phänomene in den Unterricht eingebunden werden, die diesen Zugang ermöglichen. Mit der Entscheidung für oder gegen ein spezifisches Phänomen ergeben sich Konsequenzen für die Sachstruktur des Unterrichts, für mögliche Schülervorstellungen oder gar den weiterführenden Verlauf des Unterrichts über einen längeren Zeitraum hinweg. Bestimmte Experimente eignen sich besser oder schlechter, um z. B. spezifische Fragestellungen aufzuwerfen oder einen spezifischen Effekt in den Unterricht zu integrieren (Sommer und Pfeifer 2018). Die im Syntheseframework beschriebene „Auswahl und Thematisierung von Inhalten und Fachmethoden" ist in besonderer Weise auch durch die Auswahl von (z. B. experimentellen) Zugängen auf Phänomenebene gekennzeichnet.
3. Die „Erfahrungswelt" und die „Modellwelt" lassen sich in verschiedene Ebenen differenzieren, deren Durchdenken und Verknüpfung vertieftes fachliches Verständnis und naturwissenschaftliche Kompetenz ermöglichen (Klein et al. 2018):
   - Physik (nach Mikelski-Seifert und Fischler 2003): Alltagserfahrungen, Experimente, ikonische Modelle, verbale Modellaussagen, simulative Modelle
   - Chemie (nach Johnstone 1991): Makro-, Submikro-, Symbolebene (für einen Überblick der zahlreichen Ausdifferenzierungen: Sjöström und Talanquer 2014; Talanquer 2011)

– Biologie (nach Hammann 2019): Biosphäre, Ökosystem, Lebensgemeinschaft, Organismus, Organsysteme und Organe, Gewebe, Zellen, Organellen, Moleküle.

Der Blick für Unterrichtsqualität erhält auf diesen Ebenen wesentliche Impulse dadurch, dass ein Phänomen der Erfahrungswelt aus unterschiedlichen Ebenen der Modellwelt erklärt und vorhergesagt werden kann. Vertieftes Lernen und Verständnis eines Phänomens vollziehen sich häufig in einem Durchschreiten dieser Ebenen. Die Fähigkeit, fachlich sinnvolle Bezüge zwischen diesen Ebenen herstellen zu können, ist zentral für das Verständnis. Auch Beschreibung von zunehmender naturwissenschaftlicher Expertise durch den Umgang zunehmend vernetzter und komplexer Sachverhalte (Bernholt et al. 2009; Kauertz et al. 2010) basiert auf dieser grundlegenden Idee. Im Sinne von Spiralcurricula kann das Unterrichtsangebot so aufgebaut sein, dass Phänomene zu mehreren Zeitpunkten thematisiert werden, aber jeweils auf neuen Ebenen oder auf Grundlage zunehmender Komplexität betrachtet werden können. Naturwissenschaftsdidaktische Forschungen verweisen dabei auch darauf, dass Zusammenhänge zwischen dem Komplexitätsniveau des Unterrichtsangebotes und der Schülerinnen und Schüler existieren können (Podschuweit et al. 2016) und gerade die Passung dazwischen ein bedeutsamer Faktor ist (Lau 2011).

4. Gleichzeitig lassen sich jedoch in verschiedenen naturwissenschaftlichen Themenfeldern und Theorien grundlegende Ideen bzw. Denkweisen identifizieren, die als häufig wiederkehrend und inhaltlich fundamental für naturwissenschaftliches Verständnis angesehen werden. Zu solchen „central ideas“ oder „core ideas“ gehören z. B. der diskontinuierliche Aufbau der Materie, die Erhaltung von Energie und der Masse (Talanquer 2015). Die Bildungsstandards der KMK (2005a, 2005b, 2005c) definieren Basiskonzepte, zu denen in verschiedenen Inhalten der Curricula Bezüge hergestellt werden sollen (z. B. Biologie: Struktur und Funktion; Chemie: Struktur-Eigenschaft; Physik: Wechselwirkung). Die oben angesprochenen Spiralcurricula können anhand derartiger zentraler Ideen oder Basiskonzepten organisiert sein.
5. Häufig diskutiert ist das im Mittel abfallende Interesse an Naturwissenschaften mit zunehmender Beschulungsdauer, das bereits in Deutschland in den 1990er Jahren systematisch nachgewiesen wurde (Gräber 1992) und auch aktuell zu bestehen scheint (Reiss et al. 2016). Während aktuelle Forschung auf ein komplexes Wechselspiel zwischen Wissen und Interesse hinweist (Höft et al. 2019), kann auf affektiver Ebene nicht nur das Interesse, sondern auch emotionale Reaktionen wie Angst, Respekt oder Ekel eine Rolle spielen, so dass eine inhaltsbezogene emotionale Unterstützung notwendig wird. Beispielhaft sei hier auf das Sezieren von Organen oder die Arbeit mit lebendigen Tieren, wie Mäusen oder Schaben, in der Biologie verwiesen (Polte und Wilde 2018).

### 4.3 Ausgewählte Spezifika naturwissenschaftlicher Methoden

Die Stiftung von Erfahrungen an Phänomenen verläuft häufig im Kontext von Beobachtungen oder Experimenten. Während mit dem Begriff „Experimentieren“ in

einer fachdidaktisch unbedarften Verwendung jedwede Aktivität an konkreten naturwissenschaftlichen Phänomenen (Gyllenpalm und Wickman 2011) und auch unter „Beobachten“ ein „Schauen“, „Gucken“ oder „Betrachten“ gemeint sein kann, werden im naturwissenschaftsdidaktischen Diskurs diese Begriffe vor allem genutzt, um eigenständige naturwissenschaftliche Arbeitsweisen zu bezeichnen (Gropengießer 2009; Nehring et al. 2016; Wellnitz und Mayer 2013).

Das bedeutet, dass mit den Experimenten oder Beobachtungen ein wissenschaftstheoretisch fundiertes Vorgehen zur Generierung neuen Wissens bezeichnet wird. Schülerinnen und Schüler sollen an Experimenten nicht nur Merkmale naturwissenschaftlicher Phänomene erlernen oder naturwissenschaftliches Wissen aus der „Modellwelt“ erarbeiten und anwenden („knowing science“/in den Standards der KMK: Kompetenzbereich Fachwissen), sondern auch die Umsetzung dieser Arbeitsweisen nachvollziehen und verstehen („knowing about science“/in den Standards der KMK: Kompetenzbereich Erkenntnisgewinnung) sowie befähigt werden, diese Arbeitsweisen selbstständig und problemlösend umzusetzen („doing science“/in den Standards der KMK: Kompetenzbereich Erkenntnisgewinnung). Das Experimentieren kann daher mit dem Ziel des „knowing science“ implementiert werden – es wäre dann eine Methode zum Erwerb von Fachwissen – aber auch zum Erlernen der Arbeitsweise des Experimentierens an sich. Es kann aber auch derart im Mittelpunkt von Unterrichtsphasen stehen, dass es selbst zum genuinen Inhalt des Unterrichts wird (Gyllenpalm und Wickman 2011). Das Experimentieren ist dann Ziel des Unterrichts.

Zahlreiche Studien verweisen darauf, dass für eine Förderung des „doing science“ explizit reflexive Unterrichtsphasen mit Fokus auf die Denk- und Arbeitsweisen besonders lernwirksam sind (Schwichow et al. 2016b; Vorholzer et al. 2018). Gleichzeitig sollten die Denk- und Arbeitsweisen, auch bei einer didaktischen Reduktion wissenschaftstheoretisch so angemessen sein, dass kein stark vereinfachtes oder verzerrtes Bild der Naturwissenschaften vermittelt wird („knowing about science“). Inwiefern das mit welchen Ergebnissen der Fall ist und was wissenschaftstheoretisch adäquat ist, ist wiederum Gegenstand eigenständiger und umfangreicher naturwissenschaftsdidaktischer Debatten auf nationaler (z. B. Höttecke und Rieß 2015) und internationaler Ebene (z. B. Chinn und Malhotra 2002). Dass die grundlegende Voraussetzung hierfür die Beachtung von Sicherheitsregeln, sowie die Durchführung einer Gefahrstoffprüfung im Falle eines Einsatzes von Geräten und Chemikalien, ist, sei an dieser Stelle selbstverständlich auch erwähnt (Unfallkasse NRW 2018).

## 5 Ergebnisse

### 5.1 Ergebnisse des systematischen Reviews, Verortung von Kriterien im Syntheseframework und Erweiterung des Syntheseframeworks um naturwissenschaftsdidaktische Subdimensionen

Zur Beantwortung der ersten Forschungsfrage sei an dieser Stelle zunächst auf die vollständige Übersicht der Qualitätskriterien in Anhang 3 verwiesen, welche die Grundlage für die weitere Auswertung darstellt. Um den Umfang der Zuordnung

in ein übersichtliches Format zu überführen, wurde eine vollständige Auszählung der verorteten Kriterien vorgenommen, die im Folgenden näher erläutert wird und der Frage nach dem Vergleich der empirisch verwendeten Kriterien mit dem Syntheseframework nachgeht. Für eine exemplarische Verortung einzelner Kriterien mit Hilfe der Abstufungen der Übereinstimmung siehe Abschn. 3.4.

Bei der Verortung der Qualitätskriterien aus den Videostudien im Syntheseframework ergab sich eine Verteilung der naturwissenschaftsspezifischen Kriterien auf das Syntheseframework (Tab. 3) mit einer „vollständigen Übereinstimmung“ bei 72 %, einer „teilweisen Übereinstimmung“ bei 21 % und einer „starken Abweichung“ bei 7 % der Kriterien. Besonders auffällig ist hierbei, dass der Großteil der Kriterien in den Dimensionen *I. Content selection and presentation, II. Cognitive activation* und *VII. Classroom and time management*, verortet werden konnten.

Auf Ebene der Subdimensionen treten einige Stellen hervor, an denen ein besonders hoher Anteil „teilweiser Übereinstimmung“ beobachtet werden kann. *Time management, Presenting the content in mathematically accurate and correct ways, Teacher supports students solidify their procedural knowledge/skills, Forming an environment that nurtures productive habits* und *Enhancing participation and the active engagement of all students* besitzen zwar einen hohen Anteil „teilweiser Übereinstimmung“, beinhalten jedoch nur einen verhältnismäßig geringen Anteil der insgesamt verorteten Kriterien, wodurch diese Abweichung etwas relativiert wird. Dennoch bieten diese Subdimensionen einen wichtigen Ansatzpunkt für eine Erweiterung des Syntheseframeworks. Auffälliger sind die Subdimensionen *Potential for cognitive activation (a & b)* und *Teacher facilitation of students' cognitive activity*, die jeweils 12 % der im Syntheseframework verorteten Kriterien beinhalten und somit bereits knapp ein Viertel der insgesamt verorteten Kriterien abbilden. Die Subdimension *Presenting the content in a structured way* bietet einen weiteren interessanten Ansatz, da hier ein großer Anteil der Kriterien in einer einzelnen Subdimension verortet wurde (19 %). Der Anteil „teilweise Übereinstimmung“ aller drei genannten Subdimensionen befindet sich jedoch im erwarteten Bereich von ca. 20 %. Der Anteil der Kriterien mit einer „starken Abweichung“ konnte vollständig der Dimension *II. Cognitive activation* zugeordnet werden und bietet einen direkten Hinweis für eine fachspezifische Ergänzungsstelle.

Wie in 3.5 beschrieben, erfolgte eine Bildung von Perspektivierungen zur Abbildung von Qualitätskriterien aus naturwissenschaftsdidaktischen Videostudien. Tab. 4 stellt den Vergleich zwischen dem erweiterten generischen Syntheseframework und der naturwissenschaftsdidaktischen Perspektive dar.

Die Definition der Sammelbegriffe in der naturwissenschaftsdidaktischen Perspektive erlaubte es, – in einer erneuten Kodierung der Kriterien aus den Videostudien in die Sammelbegriffe – einen Teil der Kriterien mit „teilweise Übereinstimmung“ oder „starker Abweichung“ zuzuordnen. Aus pragmatischen Gründen wird der Vergleich zum Syntheseframework direkt in dieser Version dargestellt, eine ausführliche Darstellung der Verortung der Qualitätskriterien in den Sammelbegriffen befindet sich in Anhang 4. Insgesamt stieg die Anzahl der vollständigen Übereinstimmungen damit auf 95 % Prozent und die Anzahl der Kriterien in den unterschiedlichen Subdimensionen kann damit zwischen Tab. 3 und 4 differieren.

**Tab. 3** Ergebnis des Vergleichs naturwissenschaftsdidaktischer Qualitätskriterien mit dem Syntheseframework nach Praetorius und Charalambous (2018)

| | Anzahl der kodierten naturwissenschaftsdidaktischen Kriterien | | | | Verteilung | | |
|---|---|---|---|---|---|---|---|
| Dimensionen und Subdimensionen der Unterrichtsqualität nach Praetorius und Charalambous (2018) | Vollständige Übereinstimmung | Teilweise Übereinstimmung | Starke Abw | ∑ | RH T. Ü. (%) | Anteil T. Ü. gesamt (%) | RH Ges (%) |
| *I. Content selection and presentation* | | | | | | | |
| Selecting mathematically worthwhile and developmentally appropriate content | 25 | 4 | – | 29 | 13,8 | 4,9 | 7,5 |
| Motivating the content | 20 | 3 | – | 23 | 13,0 | 3,7 | 5,9 |
| Presenting the content in a structured way | 57 | 16 | – | 73 | 21,9 | 19,5 | 18,8 |
| Presenting the content in mathematically accurate and correct ways | 7 | 5 | – | 12 | 41,7 | 6,1 | 3,1 |
| *II. Cognitive activation* | | | 26 | 26 | | | |
| Potential for cognitive activation through: | | | | | | | |
| a) Teacher's selection of challenging tasks which respond to students' cognitive level | 19 | 3 | – | 22 | 13,6 | 3,7 | 5,7 |
| b) Teacher's use of mathematically rich practices | 18 | 5 | – | 23 | 21,7 | 6,1 | 5,9 |
| Teacher facilitation of students' cognitive activity | 42 | 8 | – | 50 | 16,0 | 9,8 | 12,9 |
| Teacher supports students' meta-cognitive learning from cognitively activating tasks | 16 | 4 | – | 20 | 20,0 | 4,9 | 5,2 |
| *III. Practicing* | | | | | | | |
| Teacher supports students solidify their procedural knowledge/skills | 3 | 2 | – | 5 | 40,0 | 2,4 | 1,3 |
| Teacher procedural remediation of students' difficulties and errors in practicing | 0 | 0 | – | 0 | 0,0 | 0,0 | 0,0 |
| *IV. (Formative) Assessment* | | | | | | | |
| Assessment is aligned with learning objectives | 1 | 0 | – | 1 | 0,0 | 0,0 | 0,3 |
| Teacher regularly checks for understanding | 2 | 0 | – | 2 | 0,0 | 0,0 | 0,5 |

**Tab. 3** (Fortsetzung)

| | Anzahl der kodierten naturwissenschaftsdidaktischen Kriterien | | | | Verteilung | | |
|---|---|---|---|---|---|---|---|
| Dimensionen und Subdimensionen der Unterrichtsqualität nach Praetorius und Charalambous (2018) | Vollständige Übereinstimmung | Teilweise Übereinstimmung | Starke Abw | ∑ | RH T. Ü. (%) | Anteil T. Ü. gesamt (%) | RH Ges (%) |
| Quality of feedback for students | 7 | 0 | – | 7 | 0,0 | 0,0 | 1,8 |
| Teacher capitalizes on formative assessment information to guide next instructional steps | 1 | 0 | – | 1 | 0,0 | 0,0 | 0,3 |
| *V. Cutting-across instructional aspects aiming to maximize student learning* | | | | | | | |
| Forming an environment that nurtures productive habits | 8 | 6 | – | 14 | 42,9 | 7,3 | 3,6 |
| Differentiation and adaptation | 9 | 1 | – | 10 | 10,0 | 1,2 | 2,6 |
| Enhancing participation and the active engagement of all students | 6 | 8 | – | 14 | 57,1 | 9,8 | 3,6 |
| *VI. Socio-emotional support* | | | | | | | |
| Teacher-student relationships | 7 | 2 | – | 9 | 22,2 | 2,4 | 2,3 |
| Student-student relationships | 0 | 0 | – | 0 | 0,0 | 0,0 | 0,0 |
| *VII. Classroom and time management* | | | | | | | |
| Behavior management | 22 | 5 | – | 27 | 18,5 | 6,1 | 7,0 |
| Time management | 10 | 10 | – | 20 | 50,0 | 12,2 | 5,2 |
| ∑ | 280 | 82 | 26 | 388 | | | |
| **Relative Häufigkeit** | **72 %** | **21 %** | **7 %** | – | – | – | – |

*T.Ü.* teilweise Übereinstimmung, *RH* Relative Häufigkeit

Neben der theoriebasierten Ergänzung von Kriterien durch die Reflexion der Ziele, Inhalte und Methode der naturwissenschaftlichen Fächer ergeben sich aus der Befundlage im Bereich der videostudienbasierten Kriterien mit teilweiser Übereinstimmung, zusätzlich zur fachspezifischen Auslegung der Subdimensionen, weitere Ergänzungen. Dabei handelt es sich um:

## II. Kognitive Aktivierung

- *Kooperatives Arbeiten zur zielführenden Aktivierung der Schüler/-innen:* Einzelne Aspekte des kooperativen Arbeitens, wie es in den Videostudien beschrieben wird, können zwar im Syntheseframework gefunden werden (z. B. „[…] individu-

**Tab. 4** Erweitertes Syntheseframework generischer und naturwissenschaftsdidaktischer Perspektivierungen auf Unterrichtsqualität

| Unterrichtsqualität in Perspektivierungen der allgemeinen Unterrichtsforschung: Dimensionen und Subdimensionen | Unterrichtsqualität in Perspektivierungen naturwissenschaftsdidaktischer Forschung: Subdimensionen formuliert als Sammelbegriffe für Qualitätskriterien (Verhältnis von Generik und Fachspezifik) (Anzahl der Kriterien die unter einem Begriff gesammelt wurden) |
|---|---|
| I. Auswahl und Thematisierung von Inhalten und Fachmethoden | |
| Auswahl von bedeutungsvollen, dem Lernstand angemessenen Inhalten sowie Fachmethoden | Auswahl und Einbindung von Fachinhalten (S) (10 Kriterien)<br>Auswahl und Einbindung naturwissenschaftlicher Denk- und Arbeitsweisen (S) (23 Kriterien) |
| Motivierung von Inhalten sowie Fachmethoden | Lernen im Kontext und Verknüpfung zu gesellschaftsrelevanten Problemstellungen (G*S) (19 Kriterien)<br>Motivierende Einbettung der Inhalte (G*S) (5 Kriterien) |
| Strukturierung der thematisierten Inhalte sowie Fachmethoden | Zielklarheit (G + S) (7 Kriterien)<br>Strukturieren und Fokussieren von Schlüsselaspekten (G*S) (6 Kriterien)<br>Horizontale und vertikale Vernetzung (S) (17 Kriterien)<br>Strukturierter Ablauf und Sequenzierung der Stunde (G*S) (18 Kriterien)<br>Progression innerhalb der Stunde (G*S) (1 Kriterium)<br>Angemessene Repräsentation der Inhalte (G*S) (9 Kriterien) |
| Akkuratheit und Korrektheit der thematisierten Inhalte sowie Fachmethoden | Nutzung präziser Fachsprache (S) (6 Kriterien)<br>Vermeidung inhaltlicher Fehler (S) (2 Kriterien)<br>Fachlich adäquate Einbindung von Inhalten und Denk- und Arbeitsweisen (S) (5 Kriterien) |
| *Ergänzungen aus fachspezifischen Untersuchungen* | Adäquate didaktische Reduktion unter Berücksichtigung zukünftiger fachlicher Lernschritte (theoriebasiert ergänzt aus Fachspezifika) |
| II. Kognitive Aktivierung | |
| Auswahl fachlich gehaltvoller und auf das kognitive Niveau der Schüler*innen abgestimmter Aufgaben | Auswahl herausfordernder Lerngelegenheiten (G*S) (18 Kriterien) |
| Einsatz fachlich gehaltvoller Aufgaben | Problemlösendes Lernen (G*S) (8 Kriterien)<br>Nutzung multipler Repräsentationen und Lösungswege (G + S) (3 Kriterien)<br>Einsatz komplexer und vernetzender Aufgaben (G*S) (9 Kriterien)<br>Kognitiv aktivierender Einsatz naturwissenschaftlicher Denk- und Arbeitsweisen (S) (11 Kriterien) |
| Unterstützung der kognitiven Aktivität der Schüler*innen | Aktivierung von Vorwissen (G*S) (7 Kriterien)<br>Kognitiv-konstruktive Fehlerkultur (G*S) (6 Kriterien)<br>Explizierung von Denkprozessen (G*S) (9 Kriterien)<br>Konstruktive Einbindung von eigenen Ideen und Schülervorstellungen in den Unterricht (G*S) (22 Kriterien)<br>Unterstützung kognitiv aktivierender Prozesse (G + S) (14 Kriterien) |
| Unterstützung des metakognitiven Lernens der Schüler*innen anhand kognitiv aktivierender Aufgaben | Unterstützung und Einbindung metakognitiver Prozesse (G*S) (12 Kriterien)<br>Explizierung von naturwissenschaftlichen Denk- und Arbeitsweisen (G*S) (3 Kriterien) |
| *Ergänzungen aus fachspezifischen Untersuchungen* | Kooperatives Arbeiten zur zielführenden Aktivierung der Schüler/innen (G*S) (13 Kriterien) |

**Tab. 4** (Fortsetzung)

| Unterrichtsqualität in Perspektivierungen der allgemeinen Unterrichtsforschung: Dimensionen und Subdimensionen | Unterrichtsqualität in Perspektivierungen naturwissenschaftsdidaktischer Forschung: Subdimensionen formuliert als Sammelbegriffe für Qualitätskriterien (Verhältnis von Generik und Fachspezifik) (Anzahl der Kriterien die unter einem Begriff gesammelt wurden) |
|---|---|
| III. Unterstützung des Übens | |
| Unterstützung bei der Festigung von Lern- und Anwendungsprozessen | Wiederholende Anwendung von Fachinhalten und Methoden (G*S) (4 Kriterien) |
| Konstruktiver Umgang mit Fehlern und Schwierigkeiten von Schüler*innen beim Üben | – |
| IV. Formatives Assessment | |
| Klare Ausrichtung der Beurteilung auf die zu erlernenden Kompetenzen | Unterrichtsbezogene Rückmeldung (G + S) (1 Kriterium) |
| Regelmäßige Überprüfung des Verständnisses der Schüler*innen | Regelmäßige Überprüfung des schülerseitigen Verständnisses (G + S) (3 Kriterien) |
| Qualitativ hochwertiges Feedback an die Schüler*innen | Konstruktives Feedback (G*S) (6 Kriterien) |
| Nutzung des Feedbacks als Grundlage für die Ausrichtung des weiteren Unterrichts | Konstruktive Nutzung des Feedbacks (G*S) (1 Kriterium) |
| V. Unterstützung des Lernens aller Schüler*innen | |
| Bereitstellung eines Lernumfelds, das produktives Verhalten fördert (z. B. Eigenverantwortung, eigenständiges Lernen, Identität, Durchhaltevermögen) | Autonomie der Schüler/-innen (G*S) (8 Kriterien)<br>Selbstwahrnehmung der Schüler/-innen (G) (2 Kriterien) |
| Differenzierung und Adaptivität | Individualisierung/Differenzierung (G*S) (8 Kriterien)<br>Adaptive Erleichterung (G*S) (2 Kriterien) |
| Förderung der aktiven Mitwirkung von allen Schüler*innen | Förderung der Schülerbeteiligung (G) (3 Kriterien)<br>Lernbegleitende Unterstützung der Schüler/-innen (G*S) (14 Kriterien) |
| VI. Sozio-emotionale Unterstützung | |
| Beziehung zwischen Lehrperson und Schüler*innen | Unterstützende Lehrer-Schüler-Interaktion (G) (9 Kriterien) |
| Beziehung der Schüler*innen untereinander | – |
| *Ergänzungen aus fachspezifischen Untersuchungen* | Beziehung zum Inhalt (Antizipation von Angst, Respekt oder Ekel; theoriebasiert ergänzt aus Fachspezifika) |
| VII. Klassenführung | |
| Verhaltensmanagement | Allgegenwärtigkeit (G) (6 Kriterien)<br>Prävention (G) (15 Kriterien)<br>Intervention (G) (2 Kriterien) |
| Zeitmanagement | Zeitliche Strukturierung der Stunde (G*S) (9 Kriterien)<br>Phasenübergänge (G*S) (2 Kriterien)<br>Pacing (G + S) (8 Kriterien) |
| *Ergänzungen aus fachspezifischen Untersuchungen* | Sicherheit (S) (2 Kriterien)<br>Raum- und Materialmanagement (theoriebasiert ergänzt aus Fachspezifika) |

ell oder in Gruppen gehaltvolle Aufgaben zu bearbeiten.“ in der Subdimension „Unterstützung der kognitiven Aktivität der Schüler/-innen“). In den Videostudien wird das kooperative Arbeiten jedoch anhand von 13 Kriterien als spezifisches Mittel zur kognitiven Aktivierung beschrieben. Hierbei wird es häufig stark ausdifferenziert betrachtet und teilweise naturwissenschaftsspezifisch auf praktische Arbeitsphasen oder daran anschließende Diskussionen ausgelegt. Im Syntheseframework kann es dagegen nur als eine austauschbare Methode erfasst werden und nimmt somit nicht denselben Stellenwert ein, wie in den Videostudien. Der Fokus des Syntheseframeworks liegt hierbei auf der kognitiven Aktivierung, wie diese erreicht wird steht jedoch nicht im Fokus. Aus der Untersuchung naturwissenschaftsdidaktischer Unterrichtsforschung ging dagegen hervor, dass ein kooperatives Arbeiten, je nach Zielstellung des Unterrichts, der Einzelarbeit gegenüber zu bevorzugen ist, was mit dem Syntheseframework nicht erfasst werden kann. Durch die Fokussierung des kooperativen Arbeitens auf bestimmte Aspekte des Unterrichts, beispielsweise auf die Durchführung von Experimenten, kann es zu einer Überschneidung zu bereits bestehenden Perspektivierungen kommen. Da jedoch die besondere Rolle des kooperativen Arbeitens durch diese „teilweisen Übereinstimmungen“ nicht abgedeckt werden kann und die differenzierte Betrachtung des kooperativen Arbeitens sich auf mehr als eine Subdimension des Syntheseframeworks erstrecken würde, wird es als eine Ergänzung aufgeführt.

**VII. Klassenführung**

- *Sicherheit:* Der Sicherheitsaspekt stellt im naturwissenschaftlichen Unterricht besonders beim Experimentieren ein zentrales Thema dar. Bei Schulz (2011) wird die Sicherheit beim Experimentieren durch die Kriterien *personale Sicherheit* und *Umgang mit Geräten und Chemikalien* abgedeckt. Diese Aspekte werden bisher nicht im Syntheseframework von Praetorius und Charalambous (2018) erfasst, sollten aber auf Basis der zentralen Stellung des Experiments im naturwissenschaftlichen Unterricht, berücksichtigt werden.

### 5.2 Einblicke in ausgewählte Spezifika der Ziele, Inhalte und Methoden der naturwissenschaftlichen Unterrichtsfächer

Im Abschn. 4 wurden bereits Bezüge zwischen der Beschreibung von Zielen, Inhalten und Methoden der naturwissenschaftlichen Fächer und der Unterrichtsqualitätsforschung angedeutet. Im Folgenden wird dargestellt, inwiefern sich exemplarische Erweiterungen von Kriterien im Syntheseframework ergeben, was zur Beantwortung der zweiten Forschungsfrage beitragen soll. Dabei wird sich an den Dimensionen des Syntheseframeworks orientiert und auf Ebene der Subdimensionen ergänzt. Die Ergänzungen werden jeweils kursiv dargestellt.

**I. Auswahl und Thematisierung von Inhalten und Fachmethoden**

- *Adäquate didaktische Reduktion unter Berücksichtigung zukünftiger fachlicher Lernschritte:* Lehrkräfte sind bei der Auswahl, bei der Aufbereitung und bei der

didaktischen Reduktion von Inhalten gefordert, ihre Schülerinnen und Schüler nicht bei zukünftigen Lernschritten zu behindern („Prinzip der Ausbaufähigkeit", Risch und Peifer 2018). So existieren für den Chemieunterricht Hinweise, die Teilchen des einfachen Teilchenmodells nicht allesamt kugelförmig zu repräsentieren. Eine „fachliche Sackgasse" für zukünftige Lernschritte ergibt sich dann, wenn der Molekülbegriff erarbeitet wird und dabei deutlich wird, dass nicht alle Teilchen kugelförmig sind, sondern Moleküle über eine spezifische Geometrie verfügen und Atomen die Repräsentation der Kugel zugeordnet wird.

**VI. Sozio-emotionale Unterstützung**

- *Beziehung zum Inhalt (Motivierung, Förderung von Interesse an Inhalten, Antizipation von Angst, Respekt oder Ekel):* Gerade vor dem Hintergrund eines abnehmenden Interesses an den naturwissenschaftlichen Fächern und Fachinhalten mit zunehmenden Alter ist die Motivierung und Unterstützung einer Motivations- und Interessensentwicklung nicht nur lernunterstützend, sondern auch als eigenständiges Unterrichtsziel anerkannt. Aber auch das Empfinden von Gefahr, z. B. beim Anzünden eines Brenners im Anfangsunterricht, oder das Empfinden von Ekel, z. B. beim Sezieren von Schweineherzen im Biologieunterricht, sind emotionale Bezüge zum Inhalt, die im Sinne einer fachspezifischen Unterrichtsqualität antizipiert oder aufgegriffen werden sollten.

**VII. Klassenführung**

- *Raum- und Materialmanagement:* In diesem Kontext ist auch der Raum bedeutsam, der für experimentelle Tätigkeiten genutzt werden muss. Experimentiertische sollten abgeräumt sein, Kittel und Schutzbrillen sollten bereitliegen. Auch ausreichend viele Experimentiermaterialen oder auch funktionierende Gas- oder Wasserhähne oder Mikroskope sind Aspekte, um naturwissenschaftliches Lernen überhaupt zu ermöglichen. Veraltete Lösungen oder unreine Chemikalien, z. B. eingetrübte Calciumhydroxidlösungen, in denen sich schon Kohlenstoffdioxid aus der Luft gelöst hat, verhindern erfolgreiches Arbeiten im Unterricht. Der oben beschriebene Aspekt der Sicherheit wird durch Kriterien aus Videostudien ergänzt. Beim *Raum- und Materialmanagement* geht es nicht lediglich um einen besseren Unterrichtsfluss, vielmehr werden Aspekte beschrieben, die für den naturwissenschaftlichen Unterricht eine notwendige Voraussetzung darstellen.

## 6 Diskussion

### 6.1 Vergleichbarkeit von naturwissenschaftsdidaktischen Forschungen mit dem generischen Syntheseframework

Zur Diskussion der Befundlage dieses Artikels werden drei Perspektiven angelegt: erstens, die Perspektive des Syntheseframeworks auf naturwissenschaftsdidaktische Unterrichtsforschungen, zweitens, die Perspektive der naturwissenschaftsdidaktischen Unterrichtsforschungen auf das Syntheseframework, drittens die Perspektive

der Fachspezifika auf die Unterrichtsforschung und das Syntheseframework. Durch dieses Vorgehen sollen die Forschungsfragen zusammengeführt werden, um somit den Vergleich der naturwissenschaftsdidaktischen Unterrichtsforschungen mit dem Syntheseframework im Rahmen dieser Arbeit abzuschließen.

Aus der ersten Perspektive wird deutlich, dass naturwissenschaftsdidaktische Studien zahlreiche, der in Dimensionen und Subdimensionen abgebildeten, Aspekte des Syntheseframeworks in den Blick genommen haben. Auch wenn sich eine deutliche Fokussierung auf die Dimensionen I (Auswahl und Thematisierung von Inhalten und Fachmethoden) und II (Kognitive Aktivierung) nachweisen lässt, erscheint der Stand der videobasierten Unterrichtsforschung in den Naturwissenschaftsdidaktiken als vergleichsweise breit. Vermutlich liegen gerade in den – in Dimension I und II abgebildeten – Facetten von Unterrichtsqualität die größten Verknüpfungen zum Umgang mit Fachinhalten und fachspezifischen Denk- und Arbeitsweisen, so dass hier das Kerngeschäft naturwissenschaftsdidaktischer Forschung abgebildet wird. Damit stellt sich aber auch die Frage, ob nicht auch andere Facetten von Unterrichtsqualität, z. B. im Falle der Unterstützung des Übens, des formativen Assessments, der sozio-emotionalen Unterstützung oder Differenzierung, in den Blick genommen werden können, so dass sich auf der einen Seite vielversprechende Schnittmengen zur generisch orientierten Unterrichtsforschung zeigen sowie, auf der anderen Seite, Fragen nach fachspezifischen Operationalisierungen dieser Felder aufgeworfen werden. Welche Rolle spielt das Üben für „guten“ naturwissenschaftlichen Unterricht? Gibt es naturwissenschaftsspezifische Formen des formativen Assessments und wie und mit welchen Effekten werden sie in den Unterricht eingebunden? In diesem Sinne lässt sich das Syntheseframework als Reflexionsschablone für die Ausrichtung und Schwerpunktsetzung von fachdidaktischer Forschung verstehen.

In der zweiten Perspektive zeigt sich das Syntheseframework als weitgehend tragfähig für einen Vergleich und eine Systematisierung naturwissenschaftsdidaktischer Unterrichtsforschungen. Dies zeigt sich vor allem darin, dass der Anteil von Kriterien, die mit „teilweiser Übereinstimmung“ oder „starker Abweichung“ in das Syntheseframework eingeordnet wurden, mit circa einem Viertel der Kriterien vergleichsweise gering erscheint. Gründe für Abweichungen liegen vor allem in der stärkeren Ausdifferenzierung fachspezifischer Qualitätskriterien, die dafür sorgen, dass die Beschreibungen der Subdimensionen im Syntheseframework nicht ausreichen um diese vollständig abzubilden. Des Weiteren können einige der fachspezifischen Kriterien lediglich durch mehrere Subdimensionen des Syntheseframeworks gleichzeitig erfasst werden, wodurch die betreffenden Kriterien nicht eindeutig einer Subdimension zugeordnet werden können. Ein letzter Abweichungsgrund liegt in der schülerseitigen Auslegung und somit Fokussierung der Angebotsnutzung einiger, in den Videostudien verwendeter, Kriterien, die bisher im Syntheseframework nur in stark reduzierter Form erfasst werden. Eine Erfassung der Nutzung des Lernangebotes ist somit nicht mit demselben Grad der Ausdifferenzierung möglich wie die Erfassung des Lernangebotes selbst. Das verdeutlicht, dass weniger grundlegendkonzeptuelle Abweichungen bestehen, sondern vielmehr unterschiedliche Logiken in der Systematisierung von Kriterien für Unterrichtsqualität angewendet werden. Die nicht stets gegebene Trennschärfe kann bei der Zusammenführung von Studienlagen Probleme aufwerfen, die im konkreten Fall zu diskutieren und zu lösen

sind. Die Befunde dieses Artikels können einen Hinweis darauf geben, bei welchen Kriterien das insbesondere der Fall sein kann.

Der Befund der hohen Übereinstimmung zwischen Syntheseframework und Videostudien (bei gleichzeitiger naturwissenschaftsdidaktischer Konkretisierung) mag darin begründet sein, dass zahlreiche Arbeiten der naturwissenschaftsdidaktischen Unterrichtsforschung auch generisch ausgerichtete Studien zur Unterrichtsqualität im Rahmen ihrer theoretischen Fundierungen wahrnehmen. Gleichzeitig könnten sich darin aber auch Parallelen zeigen zwischen dem Lehren und Lernen von Mathematik, einem Bereich, in dem zahlreiche Studien der Unterrichtsforschung durchgeführt wurden, und dem Lehren und Lernen in den naturwissenschaftlichen Fächern, so dass eine Verortung von Unterrichtsqualitätskriterien leichter möglich sein könnte, als es zwischen anderen Fächern der Fall ist. Schließlich stellt das Syntheseframework selbst eine Synthese verschiedener Ansätze dar und ist – zumindest auf der Ebene der Dimensionen und Subdimensionen – auf einer vergleichsweise allgemeinen Ebene beschrieben. Auf diesen Ebenen hilft die Verallgemeinerung, Verortungen und Verknüpfungen zwischen den naturwissenschaftsdidaktischen und den generischen Ansätzen herstellen zu können. Betrachtet man die Ebene der Subdimensionen zeigt sich, dass fachspezifische Theoretisierungen und Terminologien zunehmen. Auf der Ebene der Items schließlich, auf der die Kodierungen im Review basierten, herrschen zu großen Teilen fach- und teilweise themenspezifische Operationalisierungen vor. Hier bildet sich in hohem Maße die fachspezifische Perspektivierung und Konkretisierung der Qualitätskriterien ab.

Das erweiterte Syntheseframework lässt sich aus dieser Perspektive als Kommunikationstool zwischen den naturwissenschaftsdidaktischen Disziplinen, der fachübergreifend ausgerichteten allgemeinen Unterrichtsforschung und weiteren fachdidaktischen Disziplinen interpretieren. Mit der Möglichkeit, Merkmale aus unterschiedlichen Disziplinen auf Ebene der Subdimensionen des Syntheseframeworks zu verorten, kann disziplinenübergreifende Verständigung unterstützt werden. Im Falle unterschiedlicher Begrifflichkeiten und Operationalisierungen kann Kommunikation ermöglicht werden, indem die Bezüge zur Ebene der Dimensionen hergestellt werden. So lassen sich beispielsweise Parallelen zwischen dem Unterrichtsqualitätsmerkmal der „metakognitiv-epistemologische Ebene“ (Zülsdorf-Kersting, in diesem Heft) und dem, in diesem Beitrag verorteten, Kriterium der „Explizierung von Denkprozessen“ identifizieren, indem eine Verknüpfung über die Ebene der kognitiven Aktivierung hergestellt wird.

Schließlich zeigt die Perspektive der Fachspezifika auf die Unterrichtsforschung und das Syntheseframework Möglichkeiten zur theoriebasierten Erweiterung von Kriterien für Unterrichtsqualität auf. Es deutet sich gerade mit fachspezifischem Fokus an, dass die Reflexion der oben beschriebenen impliziten Bezüge zwischen den Diskursen über Unterrichtsqualität und naturwissenschaftsdidaktischer Forschung fruchtbar für eine weiterführende Theoriebildung sein könnten. Die Ergänzung der Fragen nach einer adäquaten didaktischen Reduktion unter Berücksichtigung zukünftiger fachlicher Lernschritte könnte darauf hindeuten, dass durchaus fachdidaktisch relevante Aspekte existieren, die bisher eher selten in der naturwissenschaftsdidaktischen Unterrichtsforschung berücksichtigt wurden.

### 6.2 Typisierung von Fachspezifik im Verhältnis zu generischen Unterrichtsqualitätskriterien

Aus der Zuordnung von Kriterien in das Syntheseframework und der Erarbeitung naturwissenschaftsdidaktischer Perspektivierungen wird eine Typisierung des Verhältnisses fachspezifischer und generischer Kriterien deutlich. Diese Typisierung ist hierbei nicht als Ersatz zur Einteilung in generische und fachspezifische Anteile zu verstehen, wie sie in Anlehnung an Praetorius und Charalambous (2018), im Methodenteil beschrieben wurde. Vielmehr ergab sich die Typisierung bei der Identifikation der fachspezifischen Anteile der naturwissenschaftsdidaktischen Perspektivierung und stellt somit eine Ergänzung zur Präzisierung fachspezifischer Aspekte dar. Sie umfasst drei Typen, die insbesondere beschreiben, wie Fachspezifik im Verhältnis zu generischen Kriterien ausgestaltet sein kann:

1. *Fachspezifik durch Fokussierung generisch formulierter Kriterien auf einen spezifischen Ausschnitt des Faches („Fokussierungs-Spezifik"):* Fachspezifische Kriterien können durch eine Fokussierung von generischen Merkmalen auf Spezifika eines Faches etabliert werden, indem ein spezifischer Ausschnitt des Faches beurteilt wird und andere Bereiche des Unterrichts nicht im Fokus sind. Dabei kann das Verständnis des generischen Merkmals vergleichbar bleiben, jedoch auf ein spezifisches Fachmerkmal gerichtet sein. Beispiele hierfür sind die „kognitive Aktivierung beim Experimentieren" oder „Klarheit und Strukturiertheit des Experiments"
2. *Fachspezifik durch Konkretisierung und Operationalisierung generisch formulierter Kriterien mittels fachdidaktischer Theorieelemente („Theoretisierungs-Spezifik")*: Fachspezifische Kriterien können gebildet werden, indem generische Begrifflichkeiten genutzt, aber mittels fachdidaktischer Terminologien und Theorien operationalisiert werden. Ein Beispiel hierfür sind die Kriterien zur „Konstruktive Einbindung von eigenen Ideen und Schülervorstellungen in den Unterricht" (Dimension II „Kognitive Aktivierung"), bei denen eine fachspezifische Theorie den Fokus vorgibt, die Übertragung auf den Unterricht jedoch mit Hilfe verschiedener generischer Aspekte erfolgt. Schülervorstellungen stellen in diesem Fall ein Beispiel dar, das in der naturwissenschaftsdidaktischen Forschung einen zentralen Platz einnimmt und durch zahlreiche naturwissenschaftsdidaktische Theorien beschrieben wird (für einen Überblick siehe z. B. von Aufschnaiter und Rogge 2015). Wenngleich für eine kognitive Aktivierung der Einbezug des aktuellen Lernstandes von Schülerinnen und Schülern in einer generischen Perspektive für eine Vielzahl von Fächern wichtig ist, ist die Nutzung von Schülervorstellungen ein Spezifikum, das stärker auf die naturwissenschaftlichen Fächer zutrifft, da hier ein Konzept genutzt wird, das zur Beschreibung von fachspezifischer Kognition beim naturwissenschaftlichen Lernen genutzt wird. Fachspezifik wird durch die Einbringung dieses Konzeptes in den Diskurs der Unterrichtsqualität generiert.
3. *Fachspezifik durch Ergänzung von Kriterien, die nicht fächerübergreifend sind („genuine Fachspezifik"):* Im fachdidaktischen Diskurs werden Theorieelemente benannt, die in generisch orientierten Diskursen oder Frameworks nicht auftauchen. Beispiele hierfür sind die in diesem Beitrag genannten Ergänzungen, wie

die Berücksichtigung zukünftiger fachlicher Lernschritte bei der Auswahl und der Einbindung von Inhalten in den Unterricht.

### 6.3 Merkmale von Unterrichtsqualität in Abhängigkeit der Zieldimensionen von Unterricht

Aber nicht allein die Ableitung von weiteren Kriterien für Unterrichtsqualität kann von dieser Perspektive profitieren. Vielmehr erscheint die Frage nach Unterschieden in der Bedeutung von Qualitätskriterien für das Erreichen von verschiedenen Zielen, die Unterricht haben kann, als bisher kaum bearbeitet. Eine Systematisierung oder gar Modellierung solcher Abhängigkeiten zwischen Zieldimensionen, wie hier im Artikel beschrieben, und Qualitätsmerkmalen liegt für die naturwissenschaftlichen Fächer kaum vor – obwohl die Definition von Zielen einer Unterrichtsstunde und deren Konsequenzen für die Beurteilung der Qualität des Unterrichts durchaus Gegenstand der Lehramtsausbildung sind und für die unterschiedliche Relevanz drei Basisdimensionen für kognitive oder affektive Ziele beschrieben wurden (Klieme und Rakoczy 2008).

An dieser Stelle sei dieser Aspekt exemplarisch anhand der hier beschriebenen Zieldimensionen und Qualitätsmerkmalen konkretisiert: Für das Erreichen von Zielen aus der Zieldimension „knowing science“ (in der Sprache der Nationalen Bildungsstandards: Kompetenzbereich Fachwissen) kann eine hohe Explizierung von Vorgehensweisen von naturwissenschaftlichen Denk- und Arbeitsweisen als Teil einer kognitiven Aktivierung weniger zielführend sein, als für Ziele in der Zieldimension „doing science“ oder „knowing about science“ (in der Sprache der Nationalen Bildungsstandards: Kompetenzbereich Erkenntnisgewinnung). Wenn z. B. im letzteren Fall die verstehende Anwendung von Strategien zur Umsetzung von Experimenten erlernt werden soll, ist eine Explizierung von Vorgehensweisen insbesondere lernwirksam (Schwichow et al. 2016a), während sie im ersten Fall Unterrichtszeit binden kann, die eher für die Einbindung von Schülervorstellungen über einen Fachinhalt (kognitive Aktivierung) nützlich sein könnte. Ebenso könnte die „Gelungenheit des Experiments“ (Schulz 2011; verortet in der „Auswahl und Thematisierung von Inhalten und Fachmethoden“) für Ziele des „knowing science“ (Kompetenzbereich Fachwissen) insbesondere bedeutsam sein. Hier würde Unterrichtszeit genutzt, um ein Phänomen aus Natur oder Technik in den Unterricht zu integrieren und dann zu theoretisieren, um Fachwissen aufzubauen. Dass ein Experiment dann „klappt“, kann insbesondere für die im Rahmen der Dimension 1 „Auswahl und Thematisierung von Inhalten und Fachmethoden“ beschriebenen Perspektivierung „Strukturierter Ablauf und Sequenzierung der Stunde“ bedeutsam sein. Jedoch für Ziele im Bereich „knowing about science“ (Kompetenzbereich Erkenntnisgewinnung) könnte ein nicht gelungenes Experiment jedoch gerade ein wertvoller und kognitiv aktivierender Anlass sein, über das Wesen von Experimenten zu lernen und herauszuarbeiten, dass Experimente im engeren wissenschaftstheoretischen Sinne nicht „nicht gelingen“ können. So werden konfundierte, „nicht gelungene“ Experimente als Möglichkeit zur Förderung der Variablenkontrollstrategie beschrieben (Schwichow et al. 2016b). Anhand dieses Beispiels mit Bezügen zur Dimension 1

(„Auswahl und Thematisierung von Inhalten und Fachmethoden“) wird exemplarisch deutlich, dass das Syntheseframework einen Mehrwert gegenüber einer reinen Betrachtung von Unterrichtsqualität durch die Basisdimensionen bieten kann. Eine theoriebasierte Ableitung und Systematisierung solcher Zusammenhänge mit dem Ziel empirisch prüfbare Hypothesen abzuleiten, könnte Teil zukünftiger Aktivitäten der empirischen Unterrichtsforschung sein. Mit der Übersicht über Zieldimensionen und Qualitätsmerkmale könnte das vorliegende Paper auch dabei einen Beitrag aus naturwissenschaftsdidaktischer Perspektive leisten.

### 6.4 Limitationen der Untersuchung

Durch das beschriebene Vorgehen konnten die aufgestellten Forschungsfragen grundsätzlich beantwortet werden, dennoch ergeben sich einige Limitationen in Bezug auf das Ergebnis, die berücksichtigt werden müssen. Zunächst beschränkt sich der Abgleich zwischen dem Syntheseframework und den naturwissenschaftsdidaktischen Untersuchungen auf den deutschsprachigen Raum. Eine Erweiterung auf den internationalen Raum steht somit aus, sollte jedoch berücksichtigen, dass Unterschiede zwischen den deutschsprachigen und den internationalen Konzeptionen von Unterrichtsqualität auftreten können und dies im Forschungsansatz mitberücksichtigen. Weiterhin bleibt offen, ob das Syntheseframework den besten Ansatz zur Abbildung naturwissenschaftsdidaktischer Qualitätskriterien darstellt, auch wenn es sich für einen fächerübergreifenden Vergleich als gute Herangehensweise herausgestellt hat. An dieser Stelle würde auch die Frage anschließen, inwiefern das Syntheseframework den Ergänzungsbedarf der naturwissenschaftlichen Unterrichtsfächer, der innerhalb der drei Basisdimensionen besteht, vollständig decken kann. Zusätzlich sollte hierbei beachtet werden, dass die Naturwissenschaftsdidaktik in diesem Fall mehrere Fächer umfasst, die mitunter eine eigene Strukturierung von Unterrichtsqualität bevorzugen könnten. Hier steht die Unterrichtsqualitätsforschung insbesondere vor bisher kaum gelösten Herausforderungen, da auch innerhalb der Fächer unterschiedliche Inhaltsbereiche ausgemacht werden können (z. B. Fachgebiete der Chemie wie anorganische, organische, analytische oder physikalische Chemie, die sich im Schulcurriculum abbilden), in denen prinzipiell andere Dimensionen bedeutsam sein könnten. Abschließend bleibt anzumerken, dass der Vergleich der unterschiedlichen Systematisierungen von Unterrichtsqualität aufgezeigt hat, dass die Hierarchie bei der Darstellung von Unterrichtsqualität je nach Argumentation unterschiedlich ausgelegt werden kann. Das bedeutet für die Anwendung eines Frameworks zur Systematisierung von Unterrichtsqualität, dass die gewählten Ebenen und die Zuordnung von Aspekten zu diesen Ebenen durchaus diskutiert werden können und ein Ansatz zur strukturierten Darstellung von Unterrichtsqualität je nach Anwendungszweck mehr oder weniger hilfreich sein kann. Zielführend für die weitere Forschung wird die Offenlegung der konkreten Ziele der Verwendung eines Frameworks im Verhältnis zu seinen Zwecken sein.

**Danksagung** Die Autoren bedanken sich bei *Prof. Dr. Susanne Weßnigk und Dr. Sarah Dannemann* für ihre Anregungen zu den physik- und biologiedidaktischen Anteilen des Reviews sowie bei *Jasmin Meyer* für die Unterstützung während der Ratings und der Diskussion der Kriterien.

**Funding** Open Access funding provided by Projekt DEAL.



## Literatur

von Aufschnaiter, C., & Rogge, C. (2015). Conceptual change in learning. In R. Gunnstone (Hrsg.), *Encyclopedia of science education* (S. 209–218). Dordrecht, Heidelberg, New York, London: Springer.

Barke, H.-D. (2006). *Chemiedidaktik. Diagnose und Korrektur von Schülervorstellungen*. Berlin: Springer.

Beerenwinkel, A., & von Arx, M. (2016). Constructivism in practice: an exploratory study of teaching patterns and student motivation in physics classrooms in Finland, Germany and Switzerland. *Research in Science Education*. https://doi.org/10.1007/s11165-015-9497-3.

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, *56*(3), 205–213. https://doi.org/10.1177/0022487105275904.

Bernholt, S., Parchmann, I., & Commons, M. L. (2009). Kompetenzmodellierung zwischen Forschung und Unterrichtspraxis. *Zeitschrift Für Didaktik Der Naturwissenschaften*, *15*, 219–245.

Börlin, J. (2012). Das Experiment als Lerngelegenheit: Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* Bd. 132. Berlin: Logos.

Börlin, J., & Labudde, P. (2014). Practical work in physics instruction: an opportunity to learn? In H. E. Fischer, P. Labudde, K. Neumann & J. Viiri (Hrsg.), *Quality of instruction in physics. Comparing Finland, Switzerland and Germany* (S. 111–127). Münster: Waxmann.

Brovelli, D. (2018). *Wirksamer Physikunterricht*. Baltmannsweiler: Schneider Verlag Hohengehren.

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal Für Mathematik-Didaktik*, *39*(2), 257–284. https://doi.org/10.1007/s13138-017-0122-z.

Bybee, R. W. (1997). *Achieving scientific literacy: from purposes to practices*. Portsmouth: Heinemann.

Cauet, E. (2015). *Testen wir relevantes Wissen? Zusammenhänge zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten*. Dissertation: Universität Duisburg-Essen.

Chinn, C. A., & Malhotra, B. (2002). Epistemologically authentic inquiry in schools: a theoretical framework for evaluating inquiry tasks. *Science Education*, *86*(2), 175–218. https://doi.org/10.1002/sce.10001.

Dorfner, T., Förtsch, C., & Neuhaus, B. J. (2017). Die methodische und inhaltliche Ausrichtung quantitativer Videostudien zur Unterrichtsqualität im mathematisch-naturwissenschaftlichen Unterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, *23*(1), 261–285. https://doi.org/10.1007/s40573-017-0058-3.

Duit, R., Gropengießer, H., Kattmann, U., Komorek, M., & Parchmann, I. (2012). The model of educational reconstruction—a framework for improving teaching and learning science. In D. Jorde & J. Dillon (Hrsg.), *Science education research and practice in Europe* (S. 13–37). Rotterdam: SensePublishers. https://doi.org/10.1007/978-94-6091-900-8_2.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9.

Förtsch, C., Heidenfelder, K., Spangler, M., & Neuhaus, B.J. (2018). How does the use of core ideas in biology lessons influence students' knowledge development? *Zeitschrift Für Didaktik Der Naturwissenschaften*, *24*(1), 35–50. https://doi.org/10.1007/s40573-018-0071-1.
Givvin, K.B., Hiebert, J., Jacobs, J.K., Hollingsworth, H., & Gallimore, R. (2005). Are there national patterns of teaching? Evidence from the TIMSS 1999 video study. *Comparative Education Review*, *49*(3), 311–343. https://doi.org/10.1086/430260.
Glemnitz, I.B. (2007). Vertikale Vernetzung im Chemieunterricht. Ein Vergleich von traditionellem Unterricht mit Unterricht nach Chemie im Kontext. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* Bd. 62. Berlin: Logos.
Gräber, W. (1992). Interesse am Unterrichtsfach Chemie, an Inhalten und Tätigkeiten. *Chemie in Der Schule*, *39*, 354–358.
Gräber, W., Nentwig, P., Koballa, T., & Evans, R. (Hrsg.). (2002). *Scientific Literacy: Der Beitrag der Naturwissenschaften zur Allgemeinen Bildung*. Opladen: Leske + Budrich.
Gropengießer, H. (2009). Beobachten. In H. Gropengießer, U. Harms & U. Kattmann (Hrsg.), *Fachdidaktik Biologie* (S. 273–277). Hallbergmoos: Auslis.
Gropengießer, H., & Kattmann, U. (2013). Didaktische Rekonstruktion. In H. Gropengießer, U. Kattmann & U. Harms (Hrsg.), *Fachdidaktik Biologie* (S. 16–23). Hallbergmoos: Aulis.
Gyllenpalm, J., & Wickman, P.-O. (2011). "Experiments" and the inquiry emphasis conflation in science teacher education. *Science Education*, *95*(5), 908–926. https://doi.org/10.1002/sce.20446.
Hammann, M. (2019). Organisationsebenen biologischer Systeme unterscheiden und vernetzen: Empirische Befunde und Empfehlungen für die Praxis. In J. Groß, M. Hammann, P. Schmiemann & J. Zabel (Hrsg.), *Biologiedidaktische Forschung: Erträge für die Praxis* (S. 77–91). Berlin, Heidelberg: Springer.
Herweg, C. (2008). *Zielorientierung im deutschen und schweizerischen Physikunterricht – eine Videostudie*. Dissertation: Christian Albrechts Universität zu Kiel.
Hodson, D. (2014). Learning science, learning about science, doing science: different goals demand different learning methods. *International Journal of Science Education*, *36*(15), 2534–2553. https://doi.org/10.1080/09500693.2014.899722.
Höft, L., Bernholt, S., Blankenburg, J.S., & Winberg, M. (2019). Knowing more about things you care less about: cross-sectional analysis of the opposing trend and interplay between conceptual understanding and interest in secondary school chemistry. *Journal of Research in Science Teaching*, *56*(2), 184–210. https://doi.org/10.1002/tea.21475.
Höttecke, D., & Rieß, F. (2015). Naturwissenschaftliches Experimentieren im Lichte der jüngeren Wissenschaftsforschung – Auf der Suche nach einem authentischen Experimentbegriff der Fachdidaktik. *Zeitschrift Für Didaktik Der Naturwissenschaften*. https://doi.org/10.1007/s40573-015-0030-z.
Johnstone, A.H. (1991). Thinking about thinking. *International Newsletter on Chemical Education*, *36*, 7–10.
Kattmann, U. (2015). *Schüler besser verstehen. Alltagsvorstellungen im Biologieunterricht*. Hallbergmoos: Aulis.
Kauertz, A., Fischer, H., Mayer, J., Sumfleth, E., & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift Für Didaktik Der Naturwissenschaften*, *16*, 135–154.
Kind, P.M. (2013). Conceptualizing the science curriculum: 40 years of developing assessment frameworks in three large-scale assessments. *Science Education*, *97*(5), 671–694. https://doi.org/10.1002/sce.21070.
Klein, P., Kuhn, J., & Müller, A. (2018). Förderung von Repräsentationskompetenz und Experimentbezug in den vorlesungsbegleitenden Übungen zur Experimentalphysik. *Zeitschrift Für Didaktik Der Naturwissenschaften*. https://doi.org/10.1007/s40573-018-0070-2.
Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, *54*, 222–237.
KMK (2005a). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss*. München: Luchterhand.
KMK (2005b). *Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss*. München: Luchterhand.
KMK (2005c). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss*. München: Luchterhand.
Kunter, M., & Ewald, S. (2016). Bedingungen und Effekte von Unterricht: Aktuelle Forschungsperspektiven aus der pädagogischen Psychologie. In N. McElvany, W. Bos, H.G. Holtappels, M.M. Gebauer & F. Schwabe (Hrsg.), *Bedingungen und Effekte guten Unterrichts* (S. 9–31). Münster: Waxmann.

Lau, A. (2011). *Passung und vertikale Vernetzung im Chemie- und Physikunterricht*. Berlin: Waxmann.
Mikelskis-Seifert, S., & Fischler, H. (2003). Die Bedeutung des Denkens in Modellen ei der Entwicklung von Teilchenvorstellungen – Stand der Forschung und Entwurf einer Unterrichtskonzeption. *Zeitschrift Für Didaktik Der Naturwissenschaften*, *9*, 75–88.
Moher, D., Liberati, A., Tetzlaff, J., & Altman, D.G. (2009). Preferred reporting items for systematic reviews and Meta-analyses: the PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–270.
Muckenfuss, H. (2001). Retten uns die Phänomene? Anmerkungen zum Verhältnis von Wahrnehmung und Theorie. *Naturwissenschaften im Unterricht. Physik*, *12*(63/64), 74–77.
Nehring, A., Stiller, J., Nowak, K. H., Upmeier zu Belzen, A., & Tiemann, R. (2016). Naturwissenschaftliche Denk- und Arbeitsweisen im Chemieunterricht – eine modellbasierte Videostudie zu Lerngelegenheiten für den Kompetenzbereich der Erkenntnisgewinnung. *Zeitschrift Für Didaktik Der Naturwissenschaften*. https://doi.org/10.1007/s40573-016-0043-2.
Osborne, J. (2007). Science education for the twenty first century. *Eurasia Journal of Mathematics, Science and Technology Education*, *3*(3), 173–184. https://doi.org/10.1080/13586840903194748.
Podschuweit, S., Bernholt, S., & Brückmann, M. (2016). Classroom learning and achievement: how the complexity of classroom interaction impacts students' learning. *Research in Science & Technological Education*, *34*(2), 142–163. https://doi.org/10.1080/02635143.2015.1092955.
Polte, S., & Wilde, M. (2018). Wirkt Ekel vor lebenden Tieren bei Schülerinnen und Schülern als Prädiktor für ihr Flow-Erleben? *Zeitschrift Für Didaktik Der Naturwissenschaften*, *24*(1), 287–292. https://doi.org/10.1007/s40573-018-0075-x.
Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0.
Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002.
Praetorius, A., Herrmann, C., Gerlach, E., Zülsdorf-Kersting, M., Heinitz, B., Nehring, A. (eingereicht). Unterrichtsqualität in den Fachdidaktiken – zwischen generischen Dimensionen und fachspezifischer Ausdifferenzierung.
Rehm, M. (2018). *Wirksamer Chemieunterricht*. Unterrichtsqualität: Perspektiven von Expertinnen und Experten, Bd. 2. Baltmannsweiler: Schneider Verlag Hohengehren.
(2016). Eine Studie zwischen Kontinuität und Innovation. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme & O. Köller (Hrsg.), *PISA 2015. Eine Studie zwischen Kontinuität und Innovation*. Münster: Waxmann.
Risch, B., & Peifer, P. (2018). Didaktische Reduktion – Elementarisierung. In K. Sommer, J. Wambach-Laicher & P. Peifer (Hrsg.), *Konkrete Fachdidaktik Chemie* (S. 45–63). Seelze: Aulis.
Schecker, H., Wilhelm, T., Hopf, M., & Duit, R. (2018). *Schülervorstellungen und Physikunterricht. Ein Lehrbuch für Studium, Referendariat und Unterrichtspraxis*. Berlin, Heidelberg: Springer Spektrum. https://doi.org/10.1007/978-3-662-57270-2.
Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM – Mathematics Education*, *48*(1/2), 29–40. https://doi.org/10.1007/s11858-016-0765-0.
Schulz, A. (2011). Experimentierspezifische Qualitätsmerkmale im Chemieunterricht: Eine Videostudie. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* Bd. 113. Berlin: Logos.
Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016a). Teaching the control-of-variables strategy: a meta-analysis. *Developmental Review*, *39*, 37–63. https://doi.org/10.1016/j.dr.2015.12.001.
Schwichow, M., Zimmerman, C., Croker, S., & Härtig, H. (2016b). What students learn from hands-on activities. *Journal of Research in Science Teaching*, *53*(7), 980–1002.
Seidel, T., Prenzel, M., Rimmele, R., Dalehefte, I.M., Herweg, C., Kobarg, M., & Schwindt, K. (2006a). Blicke auf den Physikunterricht. Ergebnisse der IPN Videostudie. *Zeitschrift für Pädagogik*, *52*(6), 799–821.
Sjöström, J., & Talanquer, V. (2014). Humanizing chemistry education: from simple contextualization to multifaceted problematization. *Journal of Chemical Education*, *91*(8), 1125–1131. https://doi.org/10.1021/ed5000718.
Sommer, K., & Pfeifer, P. (2018). Experiment und Erkenntnis. In K. Sommer, J. Wambach-Laicher & P. Pfeifer (Hrsg.), *Konkrete Fachdidaktik Chemie. Grundlagen für das Lernen und Lehren im Chemieunterricht* (S. 70–88). Hannover: Friedrich.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study: methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. Washington DC: U.S. Government Printing Office.

Sumfleth, E., & Fischer, H. E. (Hrsg.). (2013). *nwu-essen 10 Jahre Forschung zum naturwissenschaftlichen Unterricht*. Berlin: Logos.

Szogs, M., Krüger, M., & Korneck, F. (2017). Erhebung von Unterrichtsqualität mittels hoch-inferenter Videorating – das Ratingmanual der Φactio-Studie. In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung in Zürich 2016* (S. 256–259).

Talanquer, V. (2011). Macro, Submicro, and symbolic: the many faces of the chemistry "triplet.". *International Journal of Science Education*, *33*(2), 179–195. https://doi.org/10.1080/09500690903386435.

Talanquer, V. (2015). Central ideas in chemistry: an alternative perspective. *Journal of Chemical Education*, *93*(1), 3–8. https://doi.org/10.1021/acs.jchemed.5b00434.

Treagust, D., Chittleborough, G., & Mamiala, T. (2003). The role of submicroscopic and symbolic representations in chemical explanations. *International Journal of Science Education*, *25*(11), 1353–1368. https://doi.org/10.1080/0950069032000070306.

Unfallkasse NRW (2018). *Gemeinsames Lernen im Chemieunterricht der Sekundarstufe I*

Vorholzer, A., von Aufschnaiter, C., & Boone, W. J. (2018). Fostering upper secondary students' ability to engage in practices of scientific investigation: a comparative analysis of an explicit and an implicit instructional approach. *Research in Science Education*. https://doi.org/10.1007/s11165-018-9691-1.

Wagenschein, M. (1976). *Die Pädagogische Dimension der Physik*. Braunschweig: Westermann.

Wellnitz, N., & Mayer, J. (2013). Erkenntnismethoden in der Biologie – Entwicklung und Evaluation eines Kompetenzmodells. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 315–346.

Widodo, A., & Duit, R. (2004). Konstruktivistische Sichtweisen vom Lehren und Lernen und die Praxis des Physikunterrichts. *Zeitschrift für Didaktik der Naturwissenschaften*, *10*, 233–255.

Wilhelm, M. (2018). *Wirksamer Biologieunterricht*. Baltmannsweiler: Schneider Verlag Hohengehren.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriesystemen und Ratingskalen*. Göttingen: Hogrefe.

Wüsten, S. (2010). *Allgemeine und fachspezifische Merkmale der Unterrichtsqualität im Fach Biologie. Eine Video- und Interventionsstudie*. Berlin: Logos.

Ziegler, A., Antes, G., & König, I. (2011) Bevorzugte Report Items für systematische Übersichten und Meta-Analysen: Das PRISMA-Statement. *DMW – Deutsche Medizinische Wochenschrift*, *136*(08):e9–e15.

### *11.2. Beitrag 2: Unterrichtsqualität in den Naturwissenschaften. Eine vergleichende Gegenüberstellung von Ansätzen zwischen Fachspezifik und Generik.*

Heinitz, B., Szogs, M., Förtsch, C., Korneck, F., Neuhaus, B. J., & Nehring, A. (2022). Unterrichtsqualität in den Naturwissenschaften. Eine vergleichende Gegenüberstellung von Ansätzen zwischen Fachspezifik und Generik. Zeitschrift Für Didaktik Der Naturwissenschaften, 28(1), 10. https://doi.org/10.1007/s40573-022-00146-5

Zusammenfassung:
Die Frage danach, was einen guten naturwissenschaftlichen Unterricht ausmacht und die Frage, inwiefern Unterrichtsqualitätsmerkmale fachspezifisch oder generisch betrachtet werden müssen, sind grundlegende Fragestellungen mit denen sich die Unterrichtsqualitätsforschung beschäftigt. Inzwischen zeichnet sich in den Naturwissenschaftsdidaktiken ein breiter Konsens ab, dass die drei Basisdimensionen der Unterrichtsqualität, *Klassenführung, konstruktive Unterstützung* und kognitive Aktivierung, fachspezifisch ausdifferenziert und ergänzt werden müssen. Zur konkreten fachspezifischen Ausdifferenzierung und Ergänzung existieren in den Naturwissenschaftsdidaktiken jedoch unterschiedliche Ansätze. Im Rahmen dieses Beitrages wurden exemplarisch drei Ansätze zur fachspezifischen Ausdifferenzierung von Unterrichtsqualitätsmerkmalen herausgegriffen und vergleichend betrachtet, um so zu einem umfassenden Bild aus der Perspektive der Naturwissenschaften zu gelangen. Dazu wurden die drei Ansätze aus dem naturwissenschaftlichen Fachbereich hinsichtlich des Verwendungszwecks, der theoretischen Fundierung und der Operationalisierung einzelner Qualitätsmerkmale verglichen. Anschließend wurden die in einem Ansatz genutzten Qualitätsmerkmale jeweils in den beiden anderen Ansätzen verortet. Hierbei konnten fünf Kategorien herausgearbeitet werden, die für einen zukünftigen systematischen Vergleich mit weiteren Ansätzen genutzt werden können. Der Beitrag stellt somit eine Möglichkeit vor, unterschiedliche Forschungsansätze zur Unterrichtsqualität systematisch aufeinander zu beziehen, um so ein umfassendes Bild der Unterrichtsqualität zu erhalten.

CRediT Author Statement zur Eigenleistung:
Benjamin Heinitz: Conceptualization**,** Writing — original draft, review & editing, Investigation, Data Curation, Formal analysis, Visualization

ZfDN (2022)28:10
https://doi.org/10.1007/s40573-022-00146-5

ORIGINAL PAPER

# Unterrichtsqualität in den Naturwissenschaften. Eine vergleichende Gegenüberstellung von Ansätzen zwischen Fachspezifik und Generik

**Benjamin Heinitz[1] · Michael Szogs[2] · Christian Förtsch[3] · Friederike Korneck[2] · Birgit J. Neuhaus[3] · Andreas Nehring[1]**

Eingegangen: 19. November 2021 / Angenommen: 5. September 2022


**Zusammenfassung**

Die Frage danach, was einen guten naturwissenschaftlichen Unterricht ausmacht und die Frage, inwiefern Unterrichtsqualitätsmerkmale fachspezifisch oder generisch betrachtet werden müssen, sind grundlegende Fragestellungen mit denen sich die Unterrichtsqualitätsforschung beschäftigt. Inzwischen zeichnet sich in den Naturwissenschaftsdidaktiken ein breiter Konsens ab, dass die drei Basisdimensionen der Unterrichtsqualität, *Klassenführung, konstruktive Unterstützung* und *kognitive Aktivierung,* fachspezifisch ausdifferenziert und ergänzt werden müssen. Zur konkreten fachspezifischen Ausdifferenzierung und Ergänzung existieren in den Naturwissenschaftsdidaktiken jedoch unterschiedliche Ansätze. Im Rahmen dieses Beitrages wurden exemplarisch drei Ansätze zur fachspezifischen Ausdifferenzierung von Unterrichtsqualitätsmerkmalen herausgegriffen und vergleichend betrachtet, um so zu einem umfassenden Bild aus der Perspektive der Naturwissenschaften zu gelangen. Dazu wurden die drei Ansätze aus dem naturwissenschaftlichen Fachbereich hinsichtlich des Verwendungszwecks, der theoretischen Fundierung und der Operationalisierung einzelner Qualitätsmerkmale verglichen. Anschließend wurden die in einem Ansatz genutzten Qualitätsmerkmale jeweils in den beiden anderen Ansätzen verortet. Hierbei konnten fünf Kategorien herausgearbeitet werden, die für einen zukünftigen systematischen Vergleich mit weiteren Ansätzen genutzt werden können. Der Beitrag stellt somit eine Möglichkeit vor, unterschiedliche Forschungsansätze zur Unterrichtsqualität systematisch aufeinander zu beziehen, um so ein umfassendes Bild der Unterrichtsqualität zu erhalten.

**Schlüsselwörter** Unterrichtsqualität · Systematisierung · Fachspezifik · Generik · Messinstrumente · Vergleichsanalyse

✉ Benjamin Heinitz
heinitz@idn.uni-hannover.de

[1] Institut für Didaktik der Naturwissenschaften, Leibniz Universität Hannover, Am Kleinen Felde 30, 30167 Hannover, Deutschland

[2] Institut für Didaktik der Physik, Goethe-Universität Frankfurt am Main, Max-von-Laue-Str. 1, 60438 Frankfurt am Main, Deutschland

[3] Didaktik der Biologie, Ludwig-Maximilians-Universität München, Winzererstr. 45/II, 80797 München, Deutschland

Published online: 12 October 2022

## Instructional Quality in Natural Sciences. A Comparison of Subject-Specific Approaches

**Abstract**
*What constitutes to good science teaching* and *to which extent should criteria of instructional quality be considered subject-specific or generic* are fundamental questions within educational research. Currently, a consensus about the necessity to differentiate and add domain-specific criteria to the broadly used basic dimensions of teaching quality, *classroom management, constructive support* and *cognitive activation* has emerged in science education. However, there are many approaches to differentiate and add to the three basic dimensions. In the context of this contribution, three approaches of instructional quality were selected and compared in order to arrive at a comprehensive picture from the perspective of the natural sciences. For this purpose, the three science specific approaches were compared with regard to the purpose of use, the theoretical foundation and the operationalization of individual criteria of instructional quality. Subsequently, the criteria used in one approach were transferred into the other two approaches to compare their compatibility. Five categories were identified, which can be used for a future systematic comparison with other approaches. The article thus presents a possibility to systematically relate different research approaches of instructional quality to each other in order to obtain a comprehensive picture of instructional quality.

**Keywords** Instructional quality · Systematisation · Subject specific · Generic · Measurement tools · Comparative analysis

## Einleitung

In der Unterrichtsforschung nimmt die Erfassung von Unterrichtsqualität eine zentrale Rolle ein. Häufig kommen dabei Merkmalslisten zum Einsatz, die generisch, also fächerübergreifend, oder auch fachspezifisch ausgelegt sind. Eine bedeutsame Zusammenstellung wesentlicher generisch konzeptualisierter Unterrichtsqualitätsmerkmale im deutschsprachigen Raum stellen die drei Basisdimensionen *Klassenführung, konstruktive Unterstützung* und *kognitive Aktivierung* dar (Klieme et al. 2001). Die drei Basisdimensionen sind fächerübergreifend ausgelegt und haben eine breite Anwendung in der Lehr-Lernforschung und fachdidaktischen Forschung gefunden z. B. in den Untersuchungen von COACTIV (Kunter und Voss 2011) und PISA (Klieme und Rakoczy 2003). In jüngeren Arbeiten wird jedoch ein generischer und auch fachspezifischer Ergänzungsbedarf der drei Basisdimensionen formuliert. Vor allem wird argumentiert, dass es nicht ausreiche, die Basisdimensionen fachspezifisch auszudifferenzieren. Stattdessen werde neben der generischen auch eine fachspezifische Perspektive auf den Unterricht benötigt (Praetorius et al. 2020a).

Eine solche Ausdifferenzierung und Ergänzung fachspezifischer Perspektiven ist bereits Gegenstand naturwissenschaftsdidaktischer Arbeiten (z. B. Korneck et al. 2017; Neuhaus 2021; Heinitz und Nehring 2020). Offen ist hierbei, inwiefern diese Ansätze miteinander vergleichbar sind und ob sie sich zu einem gemeinsamen Bild der Unterrichtsqualität in den Naturwissenschaften ergänzen könnten.

In diesem Beitrag werden exemplarisch Ansätze aus den Fächern Physik (Φactio-Ansatz), Biologie (Professionswissensansatz) und ein fachbereichsübergreifender Ansatz der Naturwissenschaften (Naturwissenschaftsdidaktische Perspektivierungen) gegenübergestellt. Dabei wird die Frage verfolgt, inwiefern innerhalb der Naturwissenschaftsdidaktiken die generischen Grundlagen vergleichbar ausdifferenziert und ergänzt werden. Damit wird exemplarisch gezeigt, wie Ansätze aus verschiedenen Fächern miteinander verglichen werden können. Weiterhin können für zukünftige Studien bestehende Ansätze zur Erfassung von Unterrichtsqualität zweckbezogen angepasst oder neue Ansätze abgeleitet werden.

## Theoretischer Hintergrund

Um den Vergleich der drei Ansätze ausreichend präzise und transparent führen zu können, wird im Folgenden die von uns verwendete Terminologie erläutert, verschiedene Interpretationen von Fachspezifik und Generik beleuchtet und eine systematische Herangehensweise von Praetorius und Charalambous (2018) dargestellt.

### Verwendete Terminologien zur Gegenüberstellung der Ansätze

Ein Vergleich unterschiedlicher Merkmalslisten der Unterrichtsqualitätsforschung macht deutlich, dass Terminologien abhängig von der zugrundeliegenden Theorie verwendet werden. So kann es durchaus vorkommen, dass dieselben Begriffe unterschiedlich ausgelegt oder für denselben Inhalt unterschiedliche Begriffe genutzt werden. Deshalb werden im Folgenden die zentralen Terminologien, die in diesem Artikel genutzt werden, in Tab. 1 definiert.

**Tab. 1** Zentrale Terminologien des Artikels zur Gegenüberstellung der Ansätze

| Terminologie | Definition |
|---|---|
| Ansatz (zur Erfassung der Unterrichtsqualität) | Eine systematische Auflistung von Qualitätsmerkmalen, die auf Basis einer gemeinsamen theoretischen Grundlage und eines gemeinsamen Verwendungszweckes erstellt wurde |
| Instrument | Eine Möglichkeit zur Messung von Unterrichtsqualität auf Grundlage unterschiedlicher Qualitätsmerkmale, Kriterien und Indikatoren. Instrumente können zweckbezogen aus einem Ansatz abgeleitet werden |
| Dimension | Die höchste Ebene einer systematischen Auflistung von Qualitätsmerkmalen, die durch Subdimensionen weiter unterteilt werden kann |
| Merkmal | Facette zur Erfassung der Unterrichtsqualität, die sich einer Dimension unterordnen lässt |
| Kriterium | Eine konkretere Operationalisierung eines Merkmals. Kriterien weisen einen Bezug zum beobachtbaren, unterrichtlichen Handeln von Lehrenden und Lernenden auf |
| Indikator/Item | Fragen oder Aufforderungen eines Instruments, die ein Urteil in Bezug auf die Ausprägung eines Kriteriums oder eines Merkmals einfordern |
| Hierarchie | Die zugrundeliegende Struktur einer Systematisierung von Unterrichtsqualität. Dimension; Merkmal etc. werden darin auf verschiedenen Ebenen abgebildet. Eine Ebene weist jeweils eine vergleichbare Abstraktion auf. Prinzipiell können zwischen der höchsten Ebene (Dimension) und den direkt beobachtbaren Indikatoren/Items beliebig viele Ebenen liegen, sofern sie sich theoretisch begründen lassen |
| Abstraktionsgrad | Bildet die Generalisierung innerhalb einer Hierarchie ab. Hohe Ebenen sind abstrakt beschrieben, erfassen inhaltlich mehrere Aspekte der Unterrichtsqualität und benötigen Schritte der Operationalisierung zur Beurteilung von konkretem Unterricht. Niedrigere Ebenen werden typischerweise konkreter im Hinblick auf die Beurteilung von unterrichtlichen Aktivitäten beschrieben |

## Fachspezifik und Generik – ein Begriffspaar mit vielschichtigen Bedeutungen

Bei einem Vergleich unterschiedlicher Ansätze der Unterrichtsqualität, stellt sich die Frage, inwiefern sich fachspezifische und generische Merkmale unterscheiden. Dafür spielt vor allem der Abstraktionsgrad eines Merkmals eine entscheidende Rolle. Je stärker die Formulierung eines Merkmals von konkret beobachtbaren Lehr-Lern-Aktivitäten abstrahiert betrachtet wird (z. B. „Auswahl und Thematisierung von Inhalten“ gegenüber „Auswahl nach der Eignung und dem Effekt des Experiments“), desto mehr Interpretationsspielraum ergibt sich bei der Anwendung auf unterschiedliche Fächer. Aber auch bei einem geringeren Abstraktionsgrad sind unterschiedliche Interpretationen möglich. Im Folgenden werden Möglichkeiten der Unterscheidung vorgestellt, in denen sich auch die drei Ansätze für den Vergleich in diesem Artikel verorten.

1. *Unterscheidung von Fachspezifik und Generik durch Interpretation des Qualitätsmerkmals*

Brunner (2018) verdeutlicht die Interpretationsmöglichkeiten am Merkmal der „Klarheit“, welches je nach generischer (Verständlichkeit des Lernziels) oder fachspezifischer Sichtweise (inhaltliche Kohärenz) interpretiert und auf unterschiedliche Aspekte des Unterrichts bezogen werden kann. Je konkreter die Formulierungen werden, desto geringer wird der Raum für verschiedene Interpretationen. Dementsprechend wird auch eine Übertragung von konkret formulierten Merkmalen zwischen unterschiedlichen Fächern erschwert. Wird ein Merkmal mit Bezug zu einer konkreten Unterrichtssituation formuliert, ist es naheliegend, dass es dadurch fachspezifischer wird.

Lindmeier und Heinze (2020) betonen jedoch, dass dieselbe Unterrichtssituation durch eine generische oder fachspezifische Interpretation abweichend beurteilt werden kann. Weiterhin ist es möglich, dass vergleichbare Unterrichtssituationen in unterschiedlichen Fächern stattfinden können, z. B. wenn der Umgang mit einer Störung im Unterricht beurteilt werden soll. Damit wäre selbst ein Merkmal mit einem eher geringen Abstraktionsgrad übertragbar. Insgesamt lässt sich also festhalten, dass ein höherer Abstraktionsgrad die Übertragbarkeit erleichtert, ein geringerer Abstraktionsgrad eine Übertragung auf andere Fächer aber nicht ausschließt. Wiprächtiger-Geppert et al. (2021) nutzen in diesem Kontext den Begriff des „fachbezogenen“ Merkmals. Damit wird ein generisches Merkmal beschrieben, das auf eine konkrete Situation im Fachunterricht angewandt wird, ohne dabei seine Übertragbarkeit zu verlieren. Hiermit wird deutlich, dass ein Merkmal zwar für den Fachunterricht relevant und entsprechend ausformuliert sein kann, jedoch grundlegend in anderen Fächern angewandt werden kann.

2. *Gemeinsames Auftreten von Fachspezifik und Generik*

Wüsten (2010) nutzt zur Unterscheidung von Fachspezifik und Generik die professionelle Wissensbasis von Lehrkräften. Sie geht davon aus, dass zur Umsetzung fachspezifischer Merkmale das fachliche oder fachdidaktische Wissen einer Lehrkraft bedeutsam ist (vgl. Shulman 1986, 1987). Somit kann auch ein eher abstrakt formuliertes und damit fächerübergreifend nutzbares Qualitätsmerkmal als

fachspezifisch angesehen werden, wenn für die Umsetzung im Unterricht entsprechend fachspezifisches Wissen notwendig ist. Dies bedeutet jedoch auch, dass Merkmale zwischen Fachspezifik und Generik verortet sein könnten, wenn sie gleichzeitig fachspezifisches und auch pädagogisch-psychologisches Wissen erfordern. Neuhaus (2021) stellt diese Verortung von Merkmalen zwischen allgemein (generisch) und fachspezifisch auf einer kontinuierlichen Skala dar. Somit würde jedes Merkmal Aspekte aus beiden Bereichen enthalten, allerdings der Anteil entsprechend variieren.

3. *Kategorische Betrachtung von Fachspezifik und Generik*

Praetorius und Charalambous (2018) verwenden den Begriff des „Hybrids“ für Fälle in denen Fachspezifik und Generik nebeneinander auftreten. So liegen beim Merkmal „Presenting the content in a structured way“ sowohl generische Anteile („Lesson objectives are clear“), als auch fachspezifische Anteile („Lesson ideas are connected to prior and next lessons and/or other ideas“) vor und eine Umsetzung würde sowohl fachspezifisches als auch pädagogisch-psychologisches Wissen voraussetzen. Neben diesen „hybriden“ Merkmalen unterscheiden Praetorius und Charalambous (2018) noch genuin fachspezifische und generische Merkmale. Diese Kategorisierung wird ebenfalls hinsichtlich der Diskussion um einen fachspezifischen oder generischen Ergänzungsbedarf der drei Basisdimensionen diskutiert (Praetorius et al. 2020a). Eine solche Kategorisierung wird auch bei Ansätzen zur Aufteilung oder Ergänzung der drei Basisdimensionen unter fachspezifischer Perspektive zugrunde gelegt (z. B. Jentsch et al. 2021; Szogs et al. 2021). Auch wenn bei diesen Systematisierungen zunächst generische Formulierungen genutzt werden, wird doch bereits bei einem hohen Abstraktionsgrad von einer Fachspezifik gesprochen.

4. *Fachspezifik durch stärkere Ausdifferenzierung*

Dreher und Leuders (2021) argumentieren, dass grundsätzlich zwischen Fächern übertragbare Merkmale generisch sind und zunächst fach- oder inhaltsspezifisch ausdifferenziert werden müssen. So würde prinzipiell jedes Merkmal zunächst generisch sein, bis es so präzise ausdifferenziert wurde, dass es nicht mehr auf andere Fächer übertragen werden kann. Sobald es auf einen bestimmten Inhalt bezogen ist, wäre ein Merkmal fachspezifisch. Die Abgrenzung von Generik und Fachspezifik wäre somit hauptsächlich eine Frage des Abstraktionsgrades und nicht von vornherein für ein Merkmal festgelegt.

5. *Fachspezifik durch Ausdifferenzierung und konzeptuelle Auslegung*

Die beiden Annahmen, dass Fachspezifik und Generik nebeneinander vorliegen, aber auch, dass Fachspezifik bei einem geringeren Abstraktionsgrad stärker hervortritt, werden in den Typisierungen von Fachspezifik aufgegriffen (Heinitz und Nehring 2020). Hierbei werden drei Formen unterschieden: Die Fokussierungs-Spezifik, die Theoretisierungs-Spezifik und die genuine Fachspezifik. Die Fokussierungs-Spezifik, tritt dann auf, wenn ein generisches Merkmal auf eine fachspezifische Anwendung zugeschnitten wird. Die generische Grundlage des Merkmals bleibt grundsätzlich erhalten, allerdings wird auf einen fachspezifischen Inhalt fokussiert. Um an das vorangegangene Beispiel der „Klarheit“ anzuknüpfen, wäre die „Klarheit und Strukturiertheit beim Experimentieren“ zu nennen. Die Theoretisierungs-Spezifik umfasst Qualitätsmerkmale, welche aus fachspezifischen Theorien abgeleitet, allerdings mit generischen Merkmalen verknüpft werden. Ein Beispiel hierfür ist die „Konstruktive Einbindung von eigenen Ideen und Schüler*innenvorstellungen in den Unterricht“, bei welcher die Fachspezifik den Fokus auf die Schüler*innenvorstellungen vorgibt, diese jedoch generisch als Teil des aktuellen Lernstandes interpretiert werden können. Bei der genuinen Fachspezifik wird direkt deutlich, dass fachspezifisches Wissen notwendig ist, um das Merkmale zu beurteilen, auch wenn es prinzipiell in anderen Fächern ebenfalls angewandt werden kann. Hier werden Qualitätsmerkmale, wie die „adäquate didaktische Reduktion unter Berücksichtigung zukünftiger fachlicher Lernschritte“ aus dem fachspezifischen Diskurs heraus abgeleitet. Die Typisierung stellt heraus, ab wann und woher eine Fachspezifität innerhalb eines Merkmals deutlich wird. Neben fachspezifischen Merkmalen gibt es aber nach wie vor generische Merkmale, die auch bei einem geringen Abstraktionsgrad auf andere Fächer übertragen werden können. So kann das „Zeitmanagement“ im Sinne der Einhaltung der vorgegebenen Unterrichtszeit von jeder Person beurteilt werden und somit auch bei einer spezifischen Ausdifferenzierung generisch bleiben. Es besteht jedoch ebenso die Möglichkeit dasselbe Merkmal fachspezifisch auszulegen, indem z. B. die zeitliche Einteilung einzelner Phasen beim Experimentieren beurteilt wird. Beide Auslegungen bewegen sich auf einem vergleichbaren Abstraktionsgrad, setzen jedoch unterschiedliches Wissen zur Beurteilung voraus.

Insgesamt wird deutlich, dass es unterschiedliche Verständnisse von Fachspezifik und Generik gibt, die sich nicht in jedem Aspekt präzise trennen lassen. Bei jeder Verwendung der beiden Begriffe scheint es notwendig, das zugrundeliegende Verständnis der beiden Begriffe zu verdeutlichen. Beim Vergleich der drei hier beschriebenen Ansätze werden fachspezifische Merkmale benannt und miteinander verglichen. Dabei wird sich an dem zugrundenliegenden Verständnis von Fachspezifik des jeweiligen Ansatzes orientiert. Somit können auch Merkmale als fachspezifisch bezeichnet werden, die mit gleicher oder ähnlicher Terminologie in allen drei Ansätzen auftauchen.

## Systematische Gegenüberstellung unterschiedlicher Ansätze zur Erfassung der Unterrichtsqualität

Praetorius und Charalambous (2018) nutzten drei Leitfragen für die Beschreibung und den strukturierten Vergleich unterschiedlicher mathematikspezifischer Instrumente der Unterrichtsforschung. Diese Leitfragen bieten auch einen strukturieren Zugang für eine Gegenüberstellung von unterschiedlichen Ansätzen der Unterrichtsqualitätsforschung, indem sie den Verwendungszweck („Warum"), die theoretische Grundlage („Was") und die Operationalisierung („Wie") der jeweiligen Ansätze aufgreifen.

1. *Der Verwendungszweck eines Ansatzes („Warum")*

Ansätze der Unterrichtsqualitätsforschung werden mit unterschiedlichen Zielen erstellt. Diese beeinflussen sowohl den Umfang, als auch die Struktur und die Operationalisierung des Ansatzes. Ein Ansatz kann zwar später auch für andere Zwecke eingesetzt werden, dennoch muss die ursprüngliche Intention bei der Interpretation des Ansatzes berücksichtigt und ggf. Anpassungen an den neuen Einsatzbereich vorgenommen werden.

2. *Die theoretische Grundlage eines Ansatzes („Was")*

Die theoretische Grundlage eines Ansatzes beeinflusst welche Begriffe in dem Ansatz genutzt werden und wie dieser strukturiert wird. Häufig wird bei einem Ansatz auf mehrere Theorien zurückgegriffen, die im Rahmen des Ansatzes miteinander verbunden werden. Für einen Vergleich verschiedener Ansätze ist es daher wichtig, die theoretische Grundlage mit im Auge zu behalten.

3. *Die Operationalisierung eines Ansatzes („Wie")*

Verwendungszweck und theoretische Grundlage werden abschließend in der Operationalisierung zusammengeführt. Hier wird die Hierarchie zwischen den verwendeten Qualitätsmerkmalen konkretisiert und damit die Struktur festgelegt. Untersuchen zwei Ansätze grundlegend unterschiedliche Qualitätsmerkmale, ist eine direkte Gegenüberstellung schwieriger, als wenn sie ähnliche Merkmale erfassen. Bilden zwei Ansätze einen ähnlichen Fokus ab, unterscheiden sich aber im Abstraktionsgrad, wird ein direkter Vergleich ebenfalls erschwert.

## Beschreibung der drei Ansätze zur Erfassung der Unterrichtsqualität

Im Folgenden werden die Ansätze aus den Fächern Physik (Φactio-Ansatz), Biologie (Professionswissensansatz) und der fächerübergreifende Ansatz der Naturwissenschaften (Naturwissenschaftsdidaktische Perspektivierungen) entlang der drei Leitfragen beschrieben. Diese Beschreibungen bieten die Grundlage, um die Ansätze im weiteren Verlauf miteinander zu vergleichen.

### Φactio – Unterrichtshandeln von Physiklehrkräften (Physik)

***Verwendungszweck („Warum?")*** Im Φactio-Ansatz wird Unterrichtsqualität erhoben, um die Zusammenhänge der Qualität des unterrichtlichen Handelns von (angehenden) Physiklehrkräften mit ihrer professionellen Kompetenz sowie ihrer Bereitschaft und Fähigkeit zur Reflexion zu analysieren. Ziel ist zu ergründen, welche Kompetenzen in der Professionalisierung der Lehrkräfte gefördert werden müssen, um einen möglichst qualitativ hochwertigen Physikunterricht zu erreichen.

Das hierfür entwickelte Instrument wird für die Qualitätseinschätzung von circa zwölfminütigen Unterrichtsminiaturen im Rahmen eines physikdidaktischen Microteaching-Seminars genutzt, in dem Lehramtsstudierende und Lehrkräfte im Vorbereitungsdienst jeweils zwei Mal halbe Schulklassen zu Freihandexperimenten der Mechanik unterrichten. Zusätzlich zu den Zusammenhangsanalysen, werden die Analyseergebnisse zur gezeigten Unterrichtsqualität als Feedback an die Lehrkräfte genutzt.

***Theoretische Grundlage („Was?")*** Die drei Basisdimensionen der Unterrichtsqualität wurden als Grundlage für das Instrument des Φactio-Ansatzes ausgewählt, um die Operationalisierung möglichst sparsam und leicht nachvollziehbar auszugestalten. Dadurch werden die Unterrichtsqualitätsmerkmale für Rater*innen klar unterscheidbar und das Feedback für die Lehrkräfte gut umsetzbar. Das Instrument des Φactio-Ansatzes differenziert die Basisdimensionen des generischen Ansatzes jedoch weiter aus. Zur Untersuchung der fachlichen Qualität und Angemessenheit wird die Dimension „Fachliche Qualität" ergänzt, welche aus den Subdimensionen „Fachliche Korrektheit" sowie „Sachgerechtigkeit" besteht. Weiterhin wird die konstruktive Unterstützung ähnlich wie in verwandten Erhebungen (Kunter und Voss 2011) in zwei Bereiche aufgeteilt: Mit der „affektiven konstruktiven Unterstützung" werden die sozio-emotionalen Merkmale getrennt von den eher didaktisch-methodischen Aspekten erfasst, die hier mit der „strukturellen konstruktiven Unterstützung" beschrieben werden. Diese Differenzierung wurde in der COACTIV-Studie angelegt und findet sich in ähnlicher Form auch bei Kleickmann et al. (2020) mit der „kognitiven Unterstützung" und „emotionale Unterstützung" sowie bei TEDS-Unterricht (Jentsch et al. 2021) mit der „kognitiven Unterstützung" und „motivationale Unterstützung".

***Operationalisierung („Wie?")*** Die Operationalisierung des Ratingmanuals orientiert sich zunächst an der PERLE-Vi-

deostudie (Lotz et al. 2013). Der Itempool wurde darüber hinaus durch die IPN-Videostudie (Seidel et al. 2006), die COACTIV-Studie (Kunter und Voss 2011), die Pythagoras-Studie (Rakoczy und Pauli 2006) und den Beobachtungsbogen der hessischen Lehrkräfte im Vorbereitungsdienst-Ausbildung ergänzt. Zusätzlich wurden für einige Subdimensionen eigene Items entwickelt, die zum Teil aus aufgezeichneten Unterrichtsreflexionen in der Physikdidaktik abgeleitet wurden.

Grundlegend besteht das Instrument aus generischen Unterrichtsqualitätsmerkmalen, die in ihrer Auswahl jedoch durch physikdidaktische Inhalte und Methoden beeinflusst sind. Die beiden Subdimensionen „fachliche Korrektheit“ und „Sachgerechtigkeit“ beschreiben die Überführung von Fachinhalten in den Unterricht und können nur unter einer physikdidaktischen Perspektive bewertet werden, sodass das Instrument insgesamt einen hybriden Charakter aufweist.

Bei der Gruppierung und Auswahl der Items wurden zunächst inhaltlich begründete Zuordnungen vorgenommen und die Items in zum Teil neu generierten Subdimensionen restrukturiert. Das Ergebnis eines iterativen Prozesses aus inhaltlichen Strukturierungen und empirischen Prüfungen sind, wie in Abb. 1 dargestellt, die Dimensionen „Fachliche Qualität“ mit zwei Subdimensionen und 14 Items, die „Kognitive Aktivierung“ mit vier Subdimensionen und 28 Items, die „Strukturelle konstruktive Unterstützung“ mit fünf Subdimensionen und 33 Items, die „Affektive konstruktive Unterstützung“ mit sechs Subdimensionen und 31 Items sowie die „Klassenführung“ mit vier Subdimensionen und 23 Items.

## Professionswissensansatz (Biologie)

***Verwendungszweck („Warum?“)*** Ziel des Ansatzes war es, analog zur COACTIV-Studie (Kunter et al. 2011), Zusammenhänge zwischen unterschiedlichen Dimensionen des Professionswissens einer Lehrkraft, deren Unterrichtsqualität und letztendlich auch der Schülerleistung zu analysieren. Die Dimensionen, die auf Ebene der Lehrkraft Berücksichtigung fanden, waren das fachdidaktische Wissen (PCK), das Fachwissen (CK) und das pädagogisch-psychologische Wissen (PK). Dimensionen der Unterrichtsqualität wurde über Lehrbücher der Biologiedidaktik gesammelt und als generische und fachspezifische Merkmale definiert (vgl. Wüsten et al. 2010). Mittels quantitativer Videostudien wurden einerseits Facetten des Professionswissens identifiziert, die Einfluss auf die Unterrichtsqualität nehmen und andererseits Unterrichtsqualitätsmerkmale, die die Schülerleistung beeinflussen. Die Erkenntnisse wurden genutzt, um Lehrkräfte durch Reflexion eigener Unterrichtsvideos zu schulen. Schließlich wurden die empirischen Ergebnisse in ein Planungsmodell für angehende Biologielehrkräfte zur Planung eines qualitativ hochwertigen Unterrichts überführt (Dorfner et al. 2019).

***Theoretische Grundlagen („Was?“)*** Grundlage des Professionswissensansatzes bildet einerseits die Lehrerprofessionalitätsforschung, andererseits die Unterrichtsqualitätsforschung. Im Bereich der Lehrerprofessionalitätsforschung wurde aufbauend auf den Arbeiten von Shulman (1986, 1987), Bromme (1992, 1997), Baumert und Kunter (2006) und Kunter et al. (2011) auf die kognitive Facette der professionellen Handlungskompetenz von Lehrkräften mit den Wissensdimensionen PCK, CK und PK fokussiert. Für die-

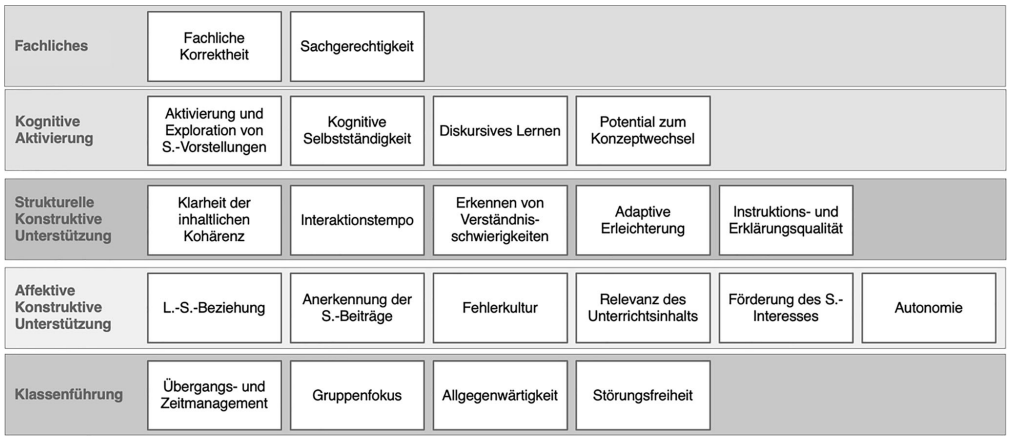



**Abb. 1** Übersicht der Dimensionen und Subdimensionen des Φactio-Ansatzes (Szogs et al. 2021)

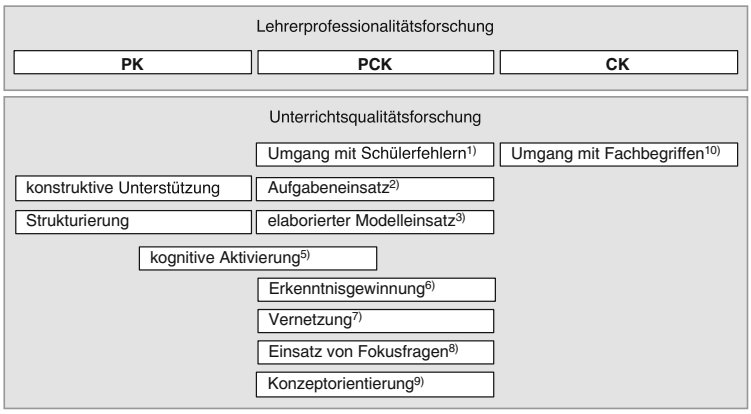



**Abb. 2** Theoretisches Modell des Professionswissensansatzes. Die verschiedenen Qualitätsmerkmale wurden den Dimensionen des Professionswissens zugeordnet. Merkmale, die mehreren Professionswissensdimensionen zugeordnet werden können, wurden mittig angeordnet (z. B. die kognitive Aktivierung). Merkmale, zu denen aus der Arbeitsgruppe heraus bereits Artikel entstanden sind, wurden mit einer Zahl versehen: (1) von Kotzebue et al. (2021), (2) Förtsch et al. (2018) & Nawani et al. (2016), (3) Werner et al. (2019), (4) Förtsch et al. (2018), (5) Förtsch et al. (2017), (6) Förtsch et al. (2016) & Dorfner et al. (2018), (7) Wadouh et al. (2014), (8) Nawani et al. (2018), (9) Förtsch et al. (2020), (10) Dorfner et al. (2020)

se drei Wissensdimensionen wurden Testinstrumente entwickelt (vgl. u. a. Jüttner et al. 2013).

Im Bereich der Unterrichtsqualitätsforschung wurde das Modell der Basisdimensionen von Klieme et al. (2001) aufgegriffen, das aber zum Teil fachspezifisch interpretiert und zusätzlich um fachspezifische Merkmale aus gängigen Lehrbüchern der Fachdidaktik ergänzt wurde (vgl. Abb. 2). Zu jedem analysierten Qualitätsmerkmal wurde theoriebasiert ein Kategoriensystem bzw. ein Ratingmanual entwickelt.

Beide Theorieansätze, die zum Professionswissen und die zur Unterrichtsqualität, wurden über ein eigens in der Arbeitsgruppe entwickeltes Modell miteinander verknüpft (vgl. auch Abb. 2). Hierzu wurden den Dimensionen PCK, CK und PK auf Basis theoretischer Überlegungen fachspezifische oder generische Merkmale der Unterrichtsqualität zugeordnet (vgl. Wüsten et al. 2010). Unterrichtsqualitätsmerkmale, für deren Durchführung Fachwissen benötigt wird, wurden als fachspezifisch definiert, solche für deren Durchführung kein Fachwissen benötigt wurde als generisch. Dieser Zusammenhang wurde später empirisch überprüft. Unterrichtsqualitätsmerkmale, die in den Datensätzen mit dem CK oder PCK der Lehrkraft korrelieren wurden als fachspezifisch interpretiert, solche, die mit dem PK korrelieren als generisch.

***Operationalisierung („Wie?")*** Die Erkenntnisse, die zur Entwicklung des Ansatzes führten, stammen aus drei quantitativen, korrelativen Videostudien, die im Rahmen verschiedener Drittmittelprojekte durchgeführt wurden: der DFG-Forschergruppe *nwu Essen*, dem DFG-Projekt *LerNT* und dem BMBF-Projekt *ProwiN*. Aus messmethodischer Perspektive wurde der Inhaltsbereich in allen drei Studien sehr klein gefasst und alle Messinstrumente auf ein eng umrissenes Inhaltsgebiet bezogen (vgl. Kotzebue et al. 2015). In der *nwu*-Videostudie Biologie war dies das Thema „Blut und Blutkreislauf" (9. Jahrgangsstufe), in der *LerNT*-Videostudie das Thema „Botanik" (6. Jahrgangsstufe) und in der *ProwiN*-Videostudie Biologie das Thema „Neurobiologie" (9. Jahrgangsstufe). Bei der Erhebung der Schülerleistung lag ein Schwerpunkt auf dem konzeptuellen Wissen der Schüler*innen.

## Naturwissenschaftsdidaktische Perspektivierung (fachbereichsübergreifend)

***Verwendungszweck („Warum?")*** Dem Ansatz der naturwissenschaftsdidaktischen Perspektivierungen liegen zwei übergreifende Ziele zugrunde. Zum einen sollen Unterrichtsqualitätsmerkmale möglichst systematisch und umfangreich abgebildet werden. Hierbei werden sowohl generische, als auch fachdidaktische Aspekte aus dem naturwissenschaftlichen Fachbereich (Biologie, Chemie, Physik, Sachunterricht) berücksichtigt (Heinitz und Nehring 2020). Grundsätzlich ist es damit möglich aus den naturwissenschaftsdidaktischen Perspektivierungen ein oder auch mehrere Instrumente zur Erfassung der Unterrichtsqualität abzuleiten. Durch die Bereitstellung einer möglichst umfangreichen Systematisierung ist es möglich Instrumente zu erstellen, die auf spezifische Aspekte fokussieren und

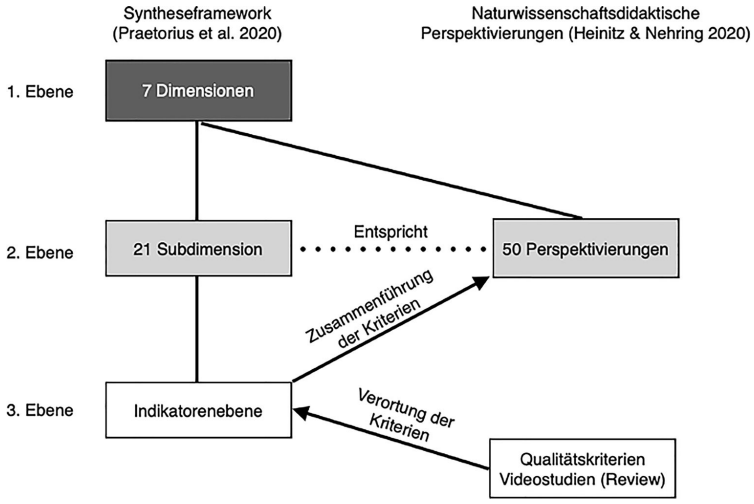



**Abb. 3** Systematisierung der naturwissenschaftsdidaktische Perspektivierungen

diese trotzdem aufeinander zu beziehen. Zum anderen kann dieser Ansatz als Kommunikationstool genutzt werden, da er direkt mit dem generischen Syntheseframework (Praetorius et al. 2020a) verknüpft ist. Durch den gemeinsamen generischen Bezugspunkt wird eine fächerübergreifende Kommunikation ermöglicht (Praetorius et al. 2020b).

***Theoretischer Hintergrund („Was?")*** Die Systematisierung der naturwissenschaftsdidaktischen Perspektivierungen orientiert sich am generischen Syntheseframework nach Praetorius und Charalambous (2018). Das Syntheseframework bietet dabei die generische Basis und die Perspektivierungen ergänzen diese um eine naturwissenschaftsdidaktische Sichtweise. Die Perspektivierungen selbst stellen inhaltliche Bündelungen von Qualitätskriterien dar, die in naturwissenschaftsdidaktischen Videostudien eingesetzt und in einem Review herausgearbeitet wurden (Heinitz und Nehring 2020). Insgesamt gibt das Syntheseframework sieben Qualitätsdimensionen vor, welchen insgesamt 50 Perspektivierungen zugeordnet sind (Abb. 3). Der Begriff der Perspektivierung wurde gewählt, um zu verdeutlichen, dass nicht lediglich eine fachspezifische oder eine generische Perspektive auf den Unterricht geworfen wird. Vielmehr dient die generische Perspektive als fächerübergreifend anwendbarer Ausgangspunkt, der durch eine fachspezifische Perspektive ausdifferenziert und ergänzt wird. Damit können sowohl generische als auch fachspezifische Aspekte vergleichbar angewendet und kommuniziert werden. Eine Perspektivierung kann als ein Qualitätsmerkmal in der Systematisierung der naturwissenschaftsdidaktischen Perspektivierungen verstanden werden.

***Operationalisierung („Wie?")*** Die Perspektivierungen liegen als kurze Beschreibungen oder Stichworte zum jeweils abgebildeten Qualitätsmerkmal vor (Tab. 2). Diese wurden ausgehend von der Indikatorenebene des Syntheseframeworks (Praetorius et al. 2020b) gebildet, wobei die Qualitätskriterien der Videostudien genutzt wurden, um fachspezifische Ausrichtungen gezielt herauszuarbeiten. Die Perspektivierungen wurden den Dimensionen des Syntheseframeworks zugeordnet und werden als parallele Ebene der Subdimensionen dargestellt. Die Dimensionen als nächsthöhere Ebene bilden somit das Bindeglied zwischen der generischen und der naturwissenschaftsdidaktischen Perspektive auf die Unterrichtsqualität. Diese generischen Konzepte sind grundsätzlich so abstrakt formuliert, dass sie problemlos auf andere Fächer übertragen werden können.

## Methodisches Vorgehen

Die Ansätze aus Physik, Biologie und der fächerübergreifende Ansatz der Naturwissenschaften konnten entlang der Leitfragen („Warum?", „Was?" und „Wie?") übersichtlich abgebildet werden. Um Gemeinsamkeiten und Unterschiede zwischen den Ansätzen herauszuarbeiten, wird im Folgenden eine Analyse in zwei Schritten vorgestellt. Der erste Schritt fokussiert dabei auf den Ansatz in seiner Gesamtheit und der zweite Schritt geht ins Detail der Operationalisierung. Durch das Vorgehen in zwei Schritten soll, entsprechend der Beschreibung von Lindmeier und Heinze (2020), neben den Unterschieden in der Operationalisierung auch der Ursprung derselben herausgearbeitet werden. Perspektivisch wird damit ein Vorgehen geboten, mit welchem

**Tab. 2** Die 50 Naturwissenschaftsdidaktische Perspektivierungen, eingeteilt in die 7 Dimensionen des Syntheseframeworks

| Dimensionen der Unterrichtsqualität | | | | | | | |
|---|---|---|---|---|---|---|---|
| Nr. | I. Auswahl und Thematisierung von Inhalten | II. Kognitive Aktivierung | III. Unterstützung des Übens | IV. Formatives Assessment | V. Unterstützung des Lernens aller Schüler*innen | VI. Sozio-emotionale Unterstützung | VII. Klassenführung |
| 1 | Auswahl und Einbindung von Fachinhalten | Auswahl herausfordernder Lerngelegenheiten | Wiederholende Anwendung von Fachinhalten und Methoden | Unterrichtsbezogene Rückmeldung | Autonomie der Schüler*innen | Unterstützende Lehrer-Schüler*in-Interaktion | Allgegenwärtigkeit |
| 2 | Auswahl und Einbindung naturwissenschaftlicher Denk- und Arbeitsweisen | Problemlösendes Lernen | Konstruktiver Umgang mit Fehlern und Schwierigkeiten von Schüler*innen beim Üben | Regelmäßige Überprüfung des schüler*innenseitigen Verständnisses | Selbstwahrnehmung der Schüler*innen | Beziehung der Schüler*innen untereinander | Prävention |
| 3 | Lernen im Kontext und Verknüpfung zu socio-scientific issues | Nutzung multipler Repräsentationen und Lösungswege | | Konstruktives Feedback | Individualisierung/ Differenzierung | Beziehung zum Inhalt (Antizipation von Angst, Respekt oder Ekel) | Intervention |
| 4 | Motivierende Einbettung der Inhalte | Einsatz komplexer und vernetzender Aufgaben | | Konstruktive Nutzung des Feedbacks | Adaptive Erleichterung | | Zeitliche Strukturierung der Stunde |
| 5 | Zielklarheit | Kognitiv aktivierender Einsatz naturwissenschaftlicher Denk- und Arbeitsweisen | | | Förderung der Schüler*innenbeteiligung | | Phasenübergänge |
| 6 | Strukturieren und Fokussieren von Schlüsselaspekten | Aktivierung von Vorwissen | | | Lernbegleitende Unterstützung der Schüler*innen | | Pacing |
| 7 | Horizontale und vertikale Vernetzung | Kognitiv-konstruktive Fehlerkultur | | | | | Sicherheit |
| 8 | Strukturierter Ablauf und Sequenzierung der Stunde | Explizierung von Denkprozessen | | | | | Vorbereitete Umgebung |
| 9 | Progression innerhalb der Stunde | Konstruktive Einbindung von eigenen Ideen und Schüler*innenvorstellungen in den Unterricht | | | | | |
| 10 | Angemessene Repräsentation der Inhalte | Unterstützung kognitiv aktivierender Prozesse | | | | | |
| 11 | Nutzung präziser Fachsprache | Unterstützung und Einbindung metakognitiver Prozesse | | | | | |
| 12 | Vermeidung inhaltlicher Fehler | Explizierung von naturwissenschaftlichen Denkweisen | | | | | |
| 13 | Fachlich adäquate Einbindung von Inhalten und Denk- und Arbeitsweisen | Kooperatives Arbeiten zur zielführenden Aktivierung der Schüler*innen | | | | | |
| 14 | Adäquate didaktische Reduktion unter Berücksichtigung zukünftiger fachlicher Lernschritte | | | | | | |

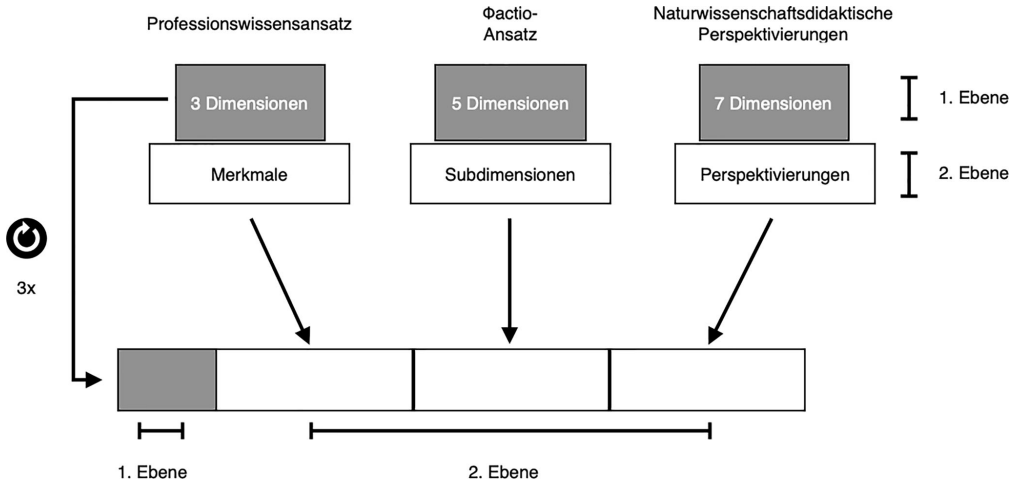



**Abb. 4** Verortung der Qualitätsmerkmale in den Systematisierungen der drei Ansätze: Die 1. Ebene des Professionswissensansatzes gibt beispielhaft die Struktur für die Gegenüberstellung vor

auch weitere Ansätze in den Vergleich aufgenommen werden können.

**Erster Schritt: Konzeptueller Vergleich der Ansätze entlang der zentralen Leitfragen** Im ersten Schritt wurden die Ansätze entlang der Leitfragen von Praetorius und Charlambous (2018) in den Tab. 3, 4 und 5 gegenübergestellt. Für einen präzisen Vergleich wurden die Leitfragen in mehrere Unterpunkte unterteilt und entsprechend für jeden Ansatz beantwortet.

**Zweiter Schritt: Gegenseitige Verortung der Operationalisierungen** Der zweite Schritt schließt an den konzeptuellen Vergleich an, fokussiert allerdings stärker auf die verwendeten Qualitätsmerkmale. Grundsätzlich kann für jeden Ansatz eine erste Ebene (Dimensionen) und eine zweite Ebene (Merkmale, Subdimensionen, Perspektivierungen) herausgestellt werden, wobei die zweite Ebene die Qualitätsmerkmale enthält, die im Vergleich gegenübergestellt wurden. Die Autor*innen der drei Ansätze haben die Operationalisierungen ihrer Qualitätsmerkmale verglichen, indem jeweils einer der Ansätze die Struktur vorgegeben hat und die anderen Ansätze ihre Qualitätsmerkmale (zweite Ebene) darin verortet haben. Somit wurden insgesamt drei getrennte Übersichtstabellen erstellt, wobei jeweils ein Ansatz die erste Ebene vorgab und die anderen ihre jeweils zweite Ebene darin verortet haben (Abb. 4).

Dieses Vorgehen setzte einen inhaltlichen Vergleich der Operationalisierungen voraus und geht über eine einfache Betrachtung der Terminologien hinaus. Merkmale die nicht verortet werden konnten, wurden separat aufgeführt. Weiterhin gab es die Möglichkeit, ein Merkmal mehrfach in der vorangestellten Systematisierung zu verorten, wenn es an unterschiedlichen Stellen passend wäre. Die hieraus resultierenden Tabellen sind entsprechend der drei Ansätze zu umfangreich, um sie an dieser Stelle vollständig abzubilden, weshalb sie als Onlineanhang zur Verfügung gestellt werden.

Durch die Berücksichtigung aller drei Perspektiven konnten die Gemeinsamkeiten und Unterschiede der drei Ansätze im weiteren Verlauf systematisch betrachtet werden. Hierzu wurden zunächst die Abweichungen zwischen den Ansätzen von den Autor*innen besprochen und Gründe für diese herausgestellt. Durch die Erläuterung der jeweiligen Perspektiven konnten dann induktive Kategorien abgeleitet werden, in denen jeweils mehrere Fälle von Abweichungen zwischen den Ansätzen gesammelt werden konnten. Diese Kategorien wurden zunächst offen formuliert und dann bei der Besprechung weiterer Kategorien präzisiert und voneinander abgegrenzt. Gemeinsamkeiten konnten anschließend auf dieselbe Weise kategorisiert werden. Zentral für die Betrachtung der Gemeinsamkeiten war hierbei, ob sie aus denselben theoretischen Grundlagen stammen und ob sie mit demselben Verwendungszweck in den jeweiligen Ansatz integriert wurden.

## Vergleichende Gegenüberstellung des Verwendungszwecks, der theoretischen Grundlage und der Operationalisierung

Im Folgenden werden die drei Ansätze tabellarisch gegenübergestellt, wobei Gemeinsamkeiten und Unterschiede mit Verweis auf die jeweiligen Tabellen zusammengefasst werden. Diese direkte Gegenüberstellung bietet einen einfachen Zugang, um die Operationalisierungen miteinander zu vergleichen.

Beim Vergleich des Verwendungszwecks (Tab. 3) wird deutlich, dass alle drei Ansätze die Unterrichtsqualität in unterschiedlichem Umfang erfassen und mit weiteren Aspekten verknüpfen. Φactio und der Professionswissensansatz verknüpfen die Unterrichtsqualität direkt mit professionellen Kompetenzen bzw. Professionswissen und sind damit bereits mit diesem konkreteren Verwendungszweck erstellt worden. Die naturwissenschaftsdidaktischen Perspektivierungen bieten eine möglichst breite Systematisierung unterschiedlicher Merkmale.

Beim Vergleich des theoretischen Hintergrunds (Tab. 4) wird deutlich, dass sowohl der Φactio-Ansatz, als auch die naturwissenschaftsdidaktischen Perspektivierungen aus den drei Basisdimensionen hervorgehen, diese jedoch in einer erweiterten Form nutzen. Der Professionswissensansatz nutzt dagegen das Professionswissen einer Lehrkraft

**Tab. 3** Gegenüberstellung des Verwendungszwecks der drei Ansätze

| *Warum?* Verwendungszweck | Φactio | Professionswissensansatz | Naturwissenschaftsdidaktische Perspektivierungen |
|---|---|---|---|
| *Fokus der Untersuchung/ Ursprüngliche Zielsetzung* | Zusammenhänge zwischen professioneller Kompetenz (Wissen, Überzeugungen, Reflexivität) und Unterrichtsqualität angehender Physiklehrkräfte | Zusammenhänge zwischen Dimensionen des Professionswissens (PCK, CK und PK) und einzelnen Merkmalen der Unterrichtsqualität | Systematisierung von Kriterien naturwissenschaftsdidaktischer Videostudien und fächerübergreifende Kommunikation der Unterrichtsqualität |
| *Untersuchung von Zusammenhängen zwischen Unterrichtsqualität und S*Outcome* | Nicht geplant | S*Outcome als Wissen und situationales Interesse gemessen | Nicht geplant |
| *Feedback an Lehrpersonen* | Einschätzungen der Unterrichtsqualität von zwei Unterrichtsversuchen durch Peers und Schüler*innen | Lehrkräfte nutzen die Merkmale der Unterrichtsqualität zur Reflexion ihrer Unterrichtsvideos | Rahmung zur Systematisierung von Feedback |
| *Evaluation von Unterricht* | Untersuchung der relativen Unterschiede zwischen den Lehrkräften | Ursprünglich nicht geplant, aber möglich und in Folgeprojekten bereits teilweise umgesetzt | Bisher nicht umgesetzt, es wird jedoch ein Rahmen geboten, in welchem Unterrichtsevaluationen vergleichend abgebildet werden können |

**Tab. 4** Gegenüberstellung des theoretischen Hintergrunds der drei Ansätze

| *Was?* Theoretischer Hintergrund | Φactio | Professionswissensansatz | Naturwissenschaftsdidaktische Perspektivierungen |
|---|---|---|---|
| *Theoretische Grundlage* | Orientiert an den drei Basisdimensionen der Unterrichtsqualität (Klieme et al. 2001) | Modell zum Professionswissen von Shulman (1986, 1987), zur professionellen Kompetenz (Baumert und Kunter 2011) und erweitertes PCK-Modell (Carlson und Daehler 2019) | Syntheseframework (Praetorius et al. 2020a) und Review zu Kriterien aus Videostudien (Heinitz und Nehring 2020) |
| *Theoretische und methodische Begründung für die Auswahl der Dimensionen* | Vergleichbarkeit der erhobenen Daten und Ergebnisse mit verwandten Studien (Kunter und Voss 2011) | Einfluss des Professionswissens der Lehrkraft auf die Unterrichtsqualität und auf Schüleroutcomes | Systematische Abbildung von generischen und fachspezifischen Unterrichtsqualitätskriterien |
| *Theoretische und methodische Begründung für die Ausdifferenzierung der Dimensionen/ Merkmale* | Abdeckung der lernrelevanten Aspekte des Unterrichts, sparsame sowie nachvollziehbare Ausgestaltung und Strukturierung der Unterrichtsqualitätsmerkmale. Die Basisdimensionen von Unterrichtsqualität wurden in fünf Dimensionen mit physikdidaktisch relevanten Subdimensionen ausdifferenziert | Strukturgebend sind die Dimensionen des Professionswissens, PCK, CK und PK, darunter liegen Unterrichtsqualitätsmerkmale aus biologiedidaktischer und pädagogisch-psychologischer Literatur | Sieben Dimensionen des Syntheseframeworks (Praetorius und Charalambous 2018). Ausdifferenzierung der Perspektivierungen erfolgte durch Gegenüberstellung und Clusterung empirisch genutzter Qualitätskriterien |

**Tab. 5** Gegenüberstellung der Operationalisierung der drei Ansätze

| *Wie?* Operationalisierung | Φactio | Professionswissensansatz | Naturwissenschaftsdidaktische Perspektivierungen |
|---|---|---|---|
| *Itemquellen* | PERLE-Videostudie (Lotz et al. 2013), IPN-Videostudie (Seidel et al. 2006), COACTIV-Studie (Kunter und Voss 2011), Pythagoras-Studie (Rakoczy und Pauli 2006)<br>der Beobachtungsbogen der hessischen Referendariatsausbildung, sowie Themenschwerpunkte von Unterrichtsreflexionen aus Seminaren | U. a. wurde sich an Studien aus der Mathematik (z. B. Pythagoras und COACTIV Studie) orientiert. Zudem wurden auf Basis von fachdidaktischer Literatur eigene Items entwickelt | Die Perspektivierungen enthalten keine konkreten Items, wurden jedoch unter Berücksichtigung der Kriterien und Items aus dem Review (Heinitz und Nehring 2020) gebildet |
| *Generisch/ fachspezifisch/ hybrid* | Überwiegend generische Merkmale, die durch einzelne Items fachlich gefärbt sind. Ergänzung der Basisdimensionen um zwei Subdimensionen zur Fachlichkeit (fachliche Korrektheit und Sachgerechtigkeit) | Orientiert an generischen Basisdimensionen, die fachspezifisch ergänzt wurden. Zudem wurde das Merkmal der kognitiven Aktivierung fachspezifisch für den Biologieunterricht ausdifferenziert (Förtsch et al. 2017) | Die Perspektivierungen stellen zwar eine fachbereichsspezifische Auslegung von Unterrichtsqualität dar, enthalten jedoch auch generische und hybride Konzepte |
| *Fokus auf spezifischen Inhalt/ Themen* | Inhaltlich sind die Qualitätsmerkmale für alle Themen des Physikunterrichts gültig. Methodisch beziehen sich circa 10 % der Items auf die Einbindung eines Experiments im Unterricht | Fokussierung auf die Inhalte Botanik, Blut und Blutkreislauf, sowie Neurobiologie. In einigen Bereichen konnten die Ergebnisse aus einem Inhaltsgebiet bereits in einem anderen Inhaltsgebiet reproduziert werden | Kein spezifischer Fokus |
| *Beobachtungsumfang* | Unterrichtsminiatur (abgeschlossener Unterricht von 12–15 min Länge) | Ja nach Studie wurde eine bis drei Unterrichtsstunden pro Lehrkraft erhoben | Nicht festgelegt |

als Grundlage der theoretischen Rahmung. Bei den darunter verorteten Qualitätsmerkmale des Ansatzes werden zudem auch die Basisdimensionen aufgegriffen. Alle drei Ansätze nutzen ähnliche generische Grundlagen und weisen Unterschiede in deren Ergänzung auf.

Der Verwendungszweck und theoretische Hintergrund spiegeln sich in der Operationalisierung der Ansätze (Tab. 5) wider. Der Φactio-Ansatz ist auf der ersten Ebene durch die erweiterten Basisdimensionen strukturiert, denen wiederum zwei bis fünf Subdimensionen (zweite Ebene) mit vier bis zehn Items (dritte Ebene) untergeordnet sind. Der Professionswissensansatz setzt die drei Dimensionen des Professionswissens auf die erste Ebene und ordnet diesen Unterrichtsqualitätsmerkmale (zweite Ebene), sowie dazugehörige Indikatoren (dritte Ebene) zu. Die naturwissenschaftsdidaktischen Perspektivierungen besetzen die erste Ebene der Hierarchie mit den sieben Dimensionen des Syntheseframeworks. Die fünfzig Perspektivierungen (zweite Ebene) werden den Dimensionen inhaltlich zugeordnet.

## Wechselseitige Verortung der Qualitätsmerkmale in die Systematisierungen der anderen Ansätze

Bei der wechselseitigen Verortung der Qualitätsmerkmale ließen sich fünf generalisierte Kategorien herausarbeiten. Diese sind aus der vollständigen Gegenüberstellung der Ansätze (Onlineanhang 1) abgeleitet und bilden induktive Zusammenfassungen von Gemeinsamkeiten und Unterschieden. Die Kategorien vereinfachen den Vergleich der Ansätze. Somit wird nicht jede einzelne Operationalisierung separat betrachtet, stattdessen können diese Gruppenweise gegenübergestellt werden.

**Kategorie 1: Vergleichbare Verwendung generischer Qualitätsmerkmale** Die Nutzung gemeinsamer Grundlagen, wie der Bezug zu den drei Basisdimensionen, führt teilweise dazu, dass dieselben generischen Merkmale für den jeweiligen Ansatz übernommen und nicht weiter ausdifferenziert werden. Beim Vergleich der drei Ansätze fiel aber auch auf, dass die zugrundeliegende Theorie einen Einfluss darauf hat, wie diese Merkmale verortet und ausdifferenziert werden.

**Kategorie 2: Vergleichbare Ergänzungen aus der fachspezifischen Anwendung** Alle drei Ansätze enthalten nach ihrem zugrundeliegenden Verständnis sowohl generische, als auch fachspezifische Anteile. Die Ansätze bedienen zwar unterschiedliche Fächer der Naturwissenschaftsdidaktiken, wurden jedoch an vielen Stellen ähnlich ausdifferenziert oder ergänzt. Beim Beispiel der „fachlichen Korrektheit" (Tab. 6) zeigt sich ein ähnliches Bild, wie bei der gemeinsamen Verwendung generischer Merkmale. Alle drei Ansätze beschreiben ein vergleichbares Merkmal, allerdings werden

**Tab. 6** Ergänzung aus der fachspezifischen Anwendung am Beispiel der fachlichen Korrektheit

| 1. Ebene | 2. Ebene | | |
|---|---|---|---|
| **Φactio** | *Φactio* | *Professionswissen* | *Perspektivierungen* |
| Fachliche Aspekte des Unterrichts | Fachliche Korrektheit<br>– Fachliche Kompetenz/ Vermeidung von Fehlern<br>– Erkennen von Fehlern<br>– Richtigstellung von Fehlern bei S* | Fachliche Richtigkeit und Stimmigkeit | Vermeidung inhaltlicher Fehler |
| | | | Nutzung präziser Fachsprache |
| | | | Fachlich adäquate Einbindung von Inhalten und Denk- und Arbeitsweisen |
| **Professionswissen** | *Professionswissen* | *Perspektivierungen* | *Φactio* |
| CK | Fachliche Richtigkeit und Stimmigkeit | Vermeidung inhaltlicher Fehler | Fachliche Korrektheit<br>– Fachliche Kompetenz/Vermeidung von Fehlern<br>– Erkennen von Fehlern<br>– Richtigstellung von Fehlern bei S* |
| | | Nutzung präziser Fachsprache | Fachliche Transparenz<br>– Sachgerechtigkeit von Kontexten und Modellen<br>– Trennung von Fach- und Alltagssprache<br>– Orchestrierung der Fachinhalte |
| | | Fachlich adäquate Einbindung von Inhalten und Denk- und Arbeitsweisen | |
| **Perspektivierungen** | *Perspektivierungen* | *Φactio* | *Professionswissen* |
| Auswahl und Thematisierung von Inhalten | Vermeidung inhaltlicher Fehler | Fachliche Korrektheit<br>– Fachliche Kompetenz/Vermeidung von Fehlern<br>– Erkennen von Fehlern<br>– Richtigstellung von Fehlern bei S* | Fachliche Richtigkeit und Stimmigkeit |

**Tab. 7** Unterschiedliche Verortung innerhalb der Ebenen am Beispiel der „Förderung des S*-Interesses“

| 1. Ebene | 2. Ebene | | |
|---|---|---|---|
| **Φactio** | *Φactio* | *Professionswissen* | *Perspektivierungen* |
| Affektive konstruktive Unterstützung | Förderung des S*-Interesses<br>– Spannende Inszenierung<br>– Interessante Auswahl von Inhalten und Methoden | Aktivierung (Einstieg) | Motivierende Einbettung der Inhalte |
| **Professionswissen** | *Professionswissen* | *Perspektivierungen* | *Φactio* |
| PCK | Aktivierung (Einstieg) | Motivierende Einbettung der Inhalte | Förderung des S*-Interesses<br>– Spannende Inszenierung<br>– Interessante Auswahl von Inhalten und Methoden |
| **Perspektivierungen** | *Perspektivierungen* | *Φactio* | *Professionswissen* |
| Auswahl und Thematisierung von Inhalten | Motivierende Einbettung der Inhalte | Förderung des Schülerinteresses<br>– Spannende Inszenierung<br>– Interessante Auswahl von Inhalten und Methoden | Aktivierung (Einstieg) |

**Tab. 8** Aufteilung eines Merkmals innerhalb anderer Systematisierungen am Beispiel des Modelleinsatzes

| 1. Ebene | 2. Ebene | | |
|---|---|---|---|
| **Professionswissen** | *Professionswissen* | *Perspektivierungen* | *Φactio* |
| PCK | Modelleinsatz<br>– Modellkritik;<br>– Art des Modells (Abstraktion, Komplexität, Illustration, Lernzielpassung);<br>– Scientific Inquiry, Modellkritik, Kritische Reflektion;<br>– Zweck des Modells, Einführung des Modells, Schüleraktivität | Auswahl und Einbindung naturwissenschaftlicher Denk- und Arbeitsweisen | Fachliche Transparenz<br>– Sachgerechtigkeit von Kontexten und Modellen |
| | | Angemessene Repräsentation der Inhalte | |
| | | Kognitiv aktivierender Einsatz naturwissenschaftlicher Denk- und Arbeitsweisen | |
| **Perspektivierungen** | *Perspektivierungen* | *Φactio* | *Professionswissen* |
| Auswahl und Thematisierung von Inhalten | Auswahl und Einbindung naturwissenschaftlicher Denk- und Arbeitsweisen | Fachliche Transparenz<br>– Sachgerechtigkeit von Kontexten und Modellen | Modelleinsatz<br>– Scientific Inquiry, Modellkritik, Kritische Reflektion<br>– Zweck des Modells, Einführung des Modells, Schüleraktivität |
| | | | Erkenntnisweg |
| | | | Beobachtungskompetenz |
| | | | Experimenteinsatz |
| | Angemessene Repräsentation der Inhalte | Fachliche Transparenz<br>– Sachgerechtigkeit von Kontexten und Modellen | Modelleinsatz<br>– Art des Modells (Abstraktion, Komplexität, Illustration, Lernzielpassung) |
| | | | Einsatz realer Objekte |
| Kognitive Aktivierung | Kognitiv aktivierender Einsatz naturwissenschaftlicher Denk- und Arbeitsweisen | Kognitive Selbstständigkeit<br>– (–) rezeptives Lernverständnis<br>– Aufgabenkultur mit S*-Selbstständigkeit | Modelleinsatz<br>– Modellkritik |
| **Φactio** | *Φactio* | *Professionswissen* | *Perspektivierungen* |
| Fachliche Aspekte des Unterrichts | Fachliche Transparenz<br>– Sachgerechtigkeit von Kontexten und Modellen<br>– Trennung von Fach- und Alltagssprache<br>– Orchestrierung der Fachinhalte | Lernen im Kontext | Angemessene Repräsentation der Inhalte |
| | | | Nutzung präziser Fachsprache |
| | | | Angemessene Einführung von fachsprachlichen Konstrukten |

durch die Systematisierungen von Φactio und dem Professionswissensansatz mehrere naturwissenschaftsdidaktische Perspektivierungen erfasst. Dies ist damit begründet, dass das Merkmal in diesen beiden Ansätzen bereits auf der zweiten Ebene stärker ausdifferenziert wird. Die Unterschiede ergeben sich hierbei nicht dadurch, dass dasselbe Merkmal fachspezifisch ausdifferenziert wird, sondern auf einer anderen Ebene verortet ist und dadurch mit einem abweichenden Abstraktionsgrad erfasst wird. Auch wenn die Merkmale damit in ähnlicher Form in den drei Ansätzen auftauchen, werden sie nach den zugrundeliegenden Verständnissen als fachspezifisch bezeichnet.

**Kategorie 3: Verortung eines Merkmals in unterschiedlichen Ebenen** Die grundlegende Hierarchisierung der Ansätze erfolgte jeweils vor einem anderen theoretischen Hintergrund. Diese somit zu Beginn festgelegte übergreifende Struktur sorgt dafür, dass mitunter vergleichbar operationalisierte Merkmale mit einer unterschiedlichen Logik in den Ansätzen verortet sind. So zeigt sich bei der Betrachtung der „Förderung des S*-Interesses“ (Tab. 7), dass dieses Merk-

mal beim Φactio-Ansatz in der Dimension „Affektive konstruktive Unterstützung“ verortet ist und die „Interessante Auswahl von Inhalten und Methoden“ hiervon einen Unterpunkt darstellt. Bei den naturwissenschaftsdidaktischen Perspektivierungen ist die „Motivierende Einbettung der Inhalte“ ein Merkmal in der Dimension „Auswahl und Thematisierung von Inhalten“. Beide Ansätze erfassen somit inhaltlich ein vergleichbares Merkmal, verorten es aber in unterschiedlich ausgelegten Dimensionen.

**Kategorie 4: Unterschiedlich aufgeteilte oder zusammengefasste Merkmale** Wird ein spezifisch fokussiertes Merkmal, wie beispielsweise der „Modelleinsatz“ im Professionswissensansatz für einen Vergleich der Ansätze genutzt (Tab. 8), zeigen sich ebenfalls Unterschiede in der Struktur. Dieses Merkmal fokussiert auf einen zentralen Aspekt naturwissenschaftlicher Denk- und Arbeitsweisen und verbindet damit mehrere Merkmale, die in den anderen Ansätzen in unterschiedlichen Dimensionen verteilt sind. Der Unterschied in der Struktur wird besonders deutlich, wenn diese Sichtweise umgekehrt wird. Werden alle Merkmale die dem „Modelleinsatz“ in den naturwissenschaftsdidaktischen Perspektivierungen zugeordnet wurden, wiederum im Professionswissensansatz gesucht, zeigt sich diese breite Aufteilung deutlich. Zwar beinhalten auch die anderen Ansätze die Aspekte des „Modelleinsatzes“ führen sie jedoch nicht in einem Merkmal zusammen, sondern betrachten sie mit einer anderen Struktur.

**Kategorie 5: Unterschiedlich operationalisierte Merkmale durch abweichenden Verwendungszweck oder theoretische Grundlagen** Eine weitere Abweichung zwischen den Ansätzen beruht auf Unterschieden im Verwendungszweck oder der theoretischen Grundlage. Hierbei handelt es sich um Qualitätsmerkmale, die in einem Ansatz erfasst werden, in einem anderen jedoch ausgelassen oder zumindest nicht so ausführlich aufgegriffen werden. Dies äußert sich dadurch, dass bestimmte Merkmale bei der Gegenüberstellung keine entsprechenden Verortungen in den anderen Systematisierungen erhalten haben, oder Merkmalen gegenüberstellt wurden, die nur teilweise denselben Inhalt erfassen (vgl. Onlineanhang 1).

## Diskussion

Insgesamt lassen sich drei Kernpunkte ableiten, die den Vergleich, die Erweiterung und auch die Erstellung zukünftiger Ansätze erleichtern könnten.

### Zusammenführung von Ansätzen der Unterrichtsqualitätsforschung

Der systematische Vergleich, wie er in der Gesamttabelle (Onlineanhang 1) abgebildet ist, zeigt deutlich, dass sich die Ansätze trotz abweichender Struktur in hohem Maße überschneiden. Es konnte festgestellt werden, dass sich eine Reihe von Qualitätsmerkmalen wie Bausteine in andere Ansätze übertragen lassen. Dabei ist es jedoch auch wichtig, dass die unterschiedlichen Perspektiven der Ansätze berücksichtigt werden, damit sich die Qualitätsmerkmale nicht überschneiden. Vor allem Merkmale mit einem besonderen Verwendungszweck stellen eine größere Herausforderung dar. Wird ein spezifisches Merkmal wie z. B. der „Modelleinsatz“ (Kategorie 4) einfach in einen anderen Ansatz übertragen, ist er nicht mehr trennscharf, wenn bereits Teile dieses Merkmals an anderer Stelle erfasst werden. Dies ist ebenso beim Erstellen eines gänzlich neuen Ansatzes zu berücksichtigen. Mögliche Überschneidungen zwischen Qualitätsmerkmalen sollten gezielt beachtet werden, um einen neuen Ansatz anschlussfähig an bereits bestehende Ansätze und inhaltlich kohärent zu gestalten. Bei spezifisch ausdifferenzierten Merkmalen kann es hilfreich sein, diese deshalb zunächst mit einem höheren Abstraktionsgrad zu betrachten.

### Generische Perspektiven und fachspezifische Anwendungen

Durch die erste Kategorie (gemeinsame generische Merkmale) konnte herausgestellt werden, dass auch fachspezifisch konzipierte Ansätze, generische Merkmale nutzen und teilweise vollständig übernehmen ohne sie weiter auszudifferenzieren. Durch die zweite Kategorie (vergleichbare fachspezifische Anwendung) wird wiederum die vergleichbare fachspezifische Ausdifferenzierung generischer Merkmale und der Ergänzungsbedarf (z. B. „Fachliche Korrektheit“) verdeutlicht. Die Ergänzungen wurden mit unterschiedlichen Verwendungszwecken und theoretischer Grundlagen hinzugefügt und lassen sich dennoch vergleichbar gegenüberstellen. Die jeweiligen Verständnisse der Fachspezifik schließen somit die Übertragbarkeit in ein anderes Fach nicht aus. Da alle drei Ansätze im Fachbereich der Naturwissenschaften verortet sind, ließe dies auch auf eine mögliche fachbereichsspezifische Ausdifferenzierung oder Ergänzung schließen. Alle drei Ansätze nutzen die drei Basisdimensionen und erweitern sie. Erweiterungen der drei Basisdimensionen wurden bereits in vorangegangenen Untersuchungen diskutiert und werden von Praetorius et al. (2020a) zu einem allgemeinen Ergänzungsbedarf zusammengefasst. Abhängig vom Verständnis der Fachspezifik könnten diese Ergänzungen zwar sowohl als generisch, als auch als fachspezifisch interpretiert wer-

den, es wäre jedoch in beiden Fällen eine Ergänzung der als generisch angesehenen Basis. Auch bei einer Faktoranalyse konnten Jentsch et al. (2021) eine mögliche Aufteilung der Dimension „konstruktive Unterstützung“ in eine motivationale und eine kognitiv-strukturierende Konzeptualisierung der Dimension herausstellen. Die kognitiv-strukturierende Konzeptualisierung umfasst die Merkmale der fachdidaktischen Strukturierung und würde somit weiter fachspezifisch ausdifferenziert. Damit entsprechen die drei Ansätze dem aktuellen Stand der Unterrichtsqualitätsforschung, wenn sie ihre generischen Grundlagen fachspezifisch ausdifferenzieren bzw. um fachspezifisch interpretierte Merkmale ergänzen.

Für den Vergleich von Ansätzen mit fachbezogenem Einsatz und die Übertragung von Merkmalen ist es wichtig herauszustellen, welches Verständnis von Fachspezifik vorliegt. So kann es durchaus sein, dass auch als fachspezifisch bezeichnete Merkmale in andere Ansätze übertragen werden können, oder ggf. ein gemeinsamer generischer Kern herausgearbeitet werden muss, wie es auch von Dreher und Leuders (2021) beschrieben wird. Abhängig vom Verständnis der Fachspezifik müssten diese jedoch mehr oder weniger stark abstrahiert werden. Eine Analyse der Fachspezifik hat für den Vergleich der drei Ansätze eine strukturierte Herangehensweise geboten, um Gemeinsamkeiten und Unterschiede herauszustellen.

Es hat sich jedoch in der ersten Kategorie auch gezeigt, dass selbst eine Dimension wie die „Klassenführung“, die häufig als generisch bezeichnet wird, auf tieferen Ebenen fachspezifische Operationalisierungen aufweisen kann (z. B. „Sicherheit“). Damit wäre diese ausdifferenzierte Dimension nicht vollständig auf andere Fächer übertragbar, da „Sicherheit“ nicht für alle Fächer ein notwendiger Unterpunkt der „Klassenführung“ ist. Ein Vergleich und ein Übertragen von Merkmalen aus unterschiedlichen Ansätzen sollte auch bei generisch konzeptualisierten Merkmalen die Ebenen mit einer stärkeren Ausdifferenzierung berücksichtigen.

Für fachspezifische Ausdifferenzierungen oder Ergänzungen der drei Basisdimensionen folgt insgesamt, dass diese über den Bezug zur gemeinsamen generischen Grundlage auch fächerübergreifend miteinander verglichen und übertragen werden können. Auch wenn die generische Grundlage nicht direkt ersichtlich ist, kann sie in den meisten Fällen herausgearbeitet werden.

### Gegenüberstellung unter Berücksichtigung von Hierarchie und Fokus

Durch die dritte Kategorie (unterschiedliche Hierarchie) und vierte Kategorie (unterschiedlich aufgeteilte oder zusammengefasste Merkmale) wird deutlich, dass die Struktur der Ansätze für einen Vergleich wichtig ist. Die drei Ansätze erweisen sich als weniger unterschiedlich, als sie auf den ersten Blick erscheinen. Dies liegt vor allem daran, dass sie zwar an vielen Stellen ähnliche Merkmale verwenden, diese aber anders strukturieren. Hierzu gehört sowohl die Verortung vergleichbarer Merkmale in unterschiedlichen Dimensionen (z. B. „Interesse“, Kategorie 3), als auch unterschiedlich aufgeteilte oder zusammengefasste Merkmale (z. B. „Modelleinsatz“, Kategorie 4).

Die unterschiedliche Hierarchie und der Fokus können durch die theoretische Grundlage und den Verwendungszweck begründet sein, oder durch eine fachbezogene Auslegung von Qualitätsmerkmalen auftreten. Werden gezielt die Operationalisierungen der Ansätze verglichen, lassen sich Gemeinsamkeiten leichter herausarbeiten und auch Merkmale gezielt zwischen den Ansätzen übertragen.

## Fazit

Der vorliegende Vergleich machte deutlich, dass Unterrichtsqualität in den drei aktuellen naturwissenschaftsdidaktischen Ansätzen nicht grundlegend unterschiedlich beschrieben wird. Neben vielen Gemeinsamkeiten in der Verwendung generischer Qualitätsmerkmale und fachspezifischer Ausdifferenzierungen oder Ergänzungen, zeigen sich die Unterschiede vorrangig in der Hierarchie und der unterschiedlichen Fokussierung von Merkmalen.

Ein gezielter Vergleich und ein Überführen einzelner Qualitätsmerkmale in einen anderen Ansatz (Onlineanhang 1) ist zu großen Teilen möglich und wird durch die Leitfragen und die von uns entwickelten fünf Kategorien erleichtert. Die drei Ansätze ergänzen sich folglich auf einer theoretischen Ebene zu einem gesamtheitlichen Bild der Unterrichtsqualität.

Wird der Verwendungszweck eines bestehenden Ansatzes erweitert oder ein neuer Ansatz mit neuem Verwendungszweck erstellt, hilft die von uns aufgezeigte systematische Vorgehensweise, um eine passende Erweiterung vorzunehmen. Besonders beim Zusammenführen unterschiedlicher Qualitätsmerkmale ist es wichtig die theoretischen Grundlagen und die konkreten Operationalisierungen zu berücksichtigen, damit inhaltliche Überschneidungen vermieden werden können und der neue oder erweiterte Ansatz trennscharf bleibt.

Für zukünftige Forschung wäre es interessant, die Qualitätseinschätzung mehrerer Ansätze direkt miteinander zu vergleichen und hierbei besonders die Überschneidungen zwischen den Ansätzen zu berücksichtigen. Lassen sich auch konkrete Qualitätseinschätzungen zu einem umfassenden Bild der Unterrichtsqualität ergänzen, oder gelingt dies nur auf einer theoretischen Ebene durch eine Kombination der Merkmalslisten?

**Funding** Open Access funding enabled and organized by Projekt DEAL.



## Literatur

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, *9*(4), 469–520.

Baumert, J., & Kunter, M. (2011). Das Kompetenzmodell von COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 29–53). Münster: Waxmann.

Bromme, R. (1992). *Der Lehrer als Experte: Zur Psychologie des professionellen Wissens*. Bern: Huber.

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln der Lehrer. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule*. Enzyklopädie der Psychologie, (Bd. 3, S. 177–212). Göttingen: Hogrefe.

Brunner, E. (2018). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *J Math Didakt*, *39*, 257–284. https://doi.org/10.1007/s13138-017-0122-z.

Carlson, J., & Daehler, K. R. (2019). The refined concensus model of pedagogical content knowledge in science education. In A. Hume, R. Cooper & A. Boroswki (Hrsg.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (S. 77–92). Singapore: Springer.

Dorfner, T., Förtsch, C., Germ, M., & Neuhaus, B. J. (2018). Biology instruction using a generic framework of scientific reasoning and argumentation with suggested lessons. *Teaching and Teacher Education*, *75*, 232–243.

Dorfner, T., Förtsch, C., Spangler, M., & Neuhaus, B. J. (2019). Wie plane ich eine konzeptorientierte Biologiestunde?: Ein Planungsmodell für den Biologieunterricht – Das Schalenmodell. *MNU Journal*, *72*(4), 300–306.

Dorfner, T., Förtsch, C., & Neuhaus, B. J. (2020). Use of technical terms in german biology lessons and its effects on students' conceptual learning. *Research in Science & Technological Education*, *38*(2), 227–251. https://doi.org/10.1080/02635143.2019.1609436.

Dreher, A., & Leuders, T. (2021). Subject-specificity of instructional quality—From the perspective of mathematics education. *Unterrichtswissenschaft*, *49*(2), 285–292. https://doi.org/10.1007/s42010-021-00116-9.

Förtsch, C., Werner, S., von Kotzebue, L., & Neuhaus, B. J. (2016). Effects of biology teachers' professional knowledge and cognitive activation on students' achievement. *International Journal of Science Education*, *38*(17), 2642–2666. https://doi.org/10.1080/09500693.2016.1257170.

Förtsch, C., Werner, S., Dorfner, T., von Kotzebue, L., & Neuhaus, B. J. (2017). Effects of cognitive activation in biology lessons on students' situational interest and achievement. *Research in Science Education*, *47*(3), 559–578. https://doi.org/10.1007/s11165-016-9517-y.

Förtsch, C., Werner, S., von Kotzebue, L., & Neuhaus, B. J. (2018). Effects of high-complexity and high-cognitive-level instructional tasks in biology lessons on students' factual and conceptual knowledge. *Research in Science & Technological Education*, *36*(3), 353–374. https://doi.org/10.1080/02635143.2017.1394286.

Förtsch, C., Dorfner, T., Baumgartner, J., Werner, S., von Kotzebue, L., & Neuhaus, B. J. (2020). Fostering students' conceptual knowledge in biology in the context of german national education standards. *Research in Science Education*, *50*(2), 739–771. https://doi.org/10.1007/s11165-018-9709-8.

Förtsch, S., Förtsch, C., von Kotzebue, L., & Neuhaus, B. J. (2018). Effects of teachers' professional knowledge and their use of three-dimensional physical models in Biology lessons on students' achievement. *Education Sciences*, *8*(3), 118. https://doi.org/10.3390/educsci8030118.

Heinitz, B., & Nehring, A. (2020). Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung. *Unterrichtswissenschaft*, *48*, 319–360. https://doi.org/10.1007/s42010-020-00074-8.

Jentsch, A., Schlesinger, L., Heinrichs, H., Kaiser, G., König, J., & Blömeke, S. (2021). Erfassung der fachspezifischen Qualität von Mathematikunterricht: Faktorenstruktur und Zusammenhänge zur professionellen Kompetenz von Mathematiklehrpersonen. *J Math Didakt*, *42*, 97–121. https://doi.org/10.1007/s13138-020-00168-x.

Jüttner, M., Boone, W., Park, S., & Neuhaus, B. J. (2013). Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK). *Educational Assessment, Evaluation and Accountability*, *25*(1), 45–67. https://doi.org/10.1007/s11092-013-9157-y.

Kleickmann, T., Steffensky, M., & Praetorius, A.-K. (2020). Quality of teaching in science education. More than three basic dimensions? *Zeitschrift für Pädagogik*, *66*(1/20), 37–53.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I. „Aufgabenkultur" und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), *TIMSS-Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 43–57). Bonn: Bundesministerium für Bildung und Forschung.

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In J. Baumert, et al. (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*. Wiesbaden: VS. https://doi.org/10.1007/978-3-322-97590-4_12.

Korneck, F., Krüger, M., & Szogs, M. (2017). Professionswissen, Lehrerüberzeugungen und Unterrichtsqualität angehender Physiklehrkräfte unterschiedlicher Schulformen. In E. Sumfleth & H. Fischler (Hrsg.), *Professionelle Kompetenzen von Lehrkräften der Chemie und Physik*. Studien zum Physik- und Chemielernen, Bd. 200. Berlin: Logos.

von Kotzebue, L., Förtsch, C., Reinold, P., Werner, S., Szcudlek, M., & Neuhaus, B. J. (2015). Quantitative Videostudien zum gymnasialen Biologieunterricht in Deutschland – Aktuelle Tendenzen und Entwicklungen. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 231–237. https://doi.org/10.1007/s40573-015-0033-9.

von Kotzebue, L., Förtsch, C., Förtsch, S., & Neuhaus, B. J. (2021). Dealing with student errors in whole-class discussions of biology lessons at German secondary schools. *International Journal*

*of Science and Mathematics Education*. https://doi.org/10.1007/s10763-021-10171-4.

Kunter, M., & Voss, T. (2011). *Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 85–113).

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Lindmeier, A., & Heinze, A. (2020). Die fachdidaktische Perspektive in der Unterrichsqualitätsforschung: (bisher) ignoriert, implizit enthalten oder nicht relevant? *Zeitschrift für Pädagogik*, *66*, 255–267.

Lotz, M., Lipowsky, F., & Faust, G. (Hrsg.). (2013). *Dokumentation der Erhebungsinstrumente des Projekts „Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE). 3. Technischer Bericht zu den PERLE-Videostudien*. Frankfurt a.M.: Gesellschaft zur Förderung Pädagogischer Forschung (GFPF).

Nawani, J., Rixius, J., & Neuhaus, B.J. (2016). Influence of using challenging tasks in biology classrooms on students' cognitive knowledge structure: An empirical video study. *International Journal of Science Education*, *38*(12), 1882–1903. https://doi.org/10.1080/09500693.2016.1213456.

Nawani, J., von Kotzebue, L., Rixius, J., Graml, M., & Neuhaus, B.J. (2018). Teachers' use of focus questions in german biology classrooms: A video-based naturalistic study. *International Journal of Science and Mathematics Education*, *16*(8), 1431–1451. https://doi.org/10.1007/s10763-017-9837-z.

Neuhaus, B.J. (2021). Unterrichtsqualität aus der Perspektive der Biologiedidaktik. *Unterrichtswissenschaft*, *49*, 273–283. https://doi.org/10.1007/s42010-021-00114-x.

Praetorius, A.-K., & Charalambous, C.Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM Mathematics Education*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0.

Praetorius, A.-K., Rogh, W., & Kleickmann, T. (2020a). Blinde Flecken des Modells der drei Basisdimensionen von Unterrichtsqualität? Das Modell im Spiegel einer internationalen Synthese von Merkmalen der Unterrichtsqualität. *Unterrichtswissenschaft*, *48*, 303–318. https://doi.org/10.1007/s42010-020-00072-w.

Praetorius, A.-K., Herrmann, C., Gerlach, E., Zülsdorf-Kersting, M., Heinitz, B., & Nehring, A. (2020b). Unterrichtsqualität in den Fachdidaktiken im deutschsprachigen Raum – zwischen Generik und Fachspezifik. *Unterrichtswissenschaft*, *48*, 409–446. https://doi.org/10.1007/s42010-020-00082-8.

Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 3)* (S. 189–205). Frankfurt a.M.: GFPF.

Seidel, T., Prenzel, M., Rimmele, R., Dalehefte, I.M., Herweg, C., Kobarg, M., & Schwindt, K. (2006). Blicke auf den Physikunterricht. Ergebnisse der IPN Videostudie. *Zeitschrift für Pädagogik*, *52*(6), 799–821.

Shulman, L.S. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Shulman, L.S. (1987). Knowledge and teaching of the new reform. *Harvard Educational Review*, *57*, 1–22.

Szogs, M., Oettinghaus, L., Krüger, M., Große, A., & Korneck, F. (2021). *Ratingmanual zur Einschätzung der Unterrichtsqualität im Physikunterricht [Ratingmanual: Version 1.0]. Erstanwendung 2018*. Frankfurt a.M.: Forschungsdatenzentrum Bildung am DIPF. https://doi.org/10.7477/614:326:1.

Wadouh, J., Liu, N., Sandmann, A., & Neuhaus, B.J. (2014). The effect of knowledge linking levels in biology lessons upon students' knowledge structure. *International Journal of Science and Mathematics Education*, *12*(1), 25–47.

Werner, S., Förtsch, C., Boone, W., von Kotzebue, L., & Neuhaus, B.J. (2019). Investigating how german biology teachers use three-dimensional physical models in classroom instruction: A video study. *Research in Science Education*, *49*, 437–463. https://doi.org/10.1007/s11165-017-9624-4.

Wiprächtiger-Geppert, M., Stahns, R., & Riegler, S. (2021). Fachspezifität von Unterrichtsqualität in der Deutschdidaktik. *Unterrichtswiss*, *49*, 203–209. https://doi.org/10.1007/s42010-021-00109-8.

Wüsten, S. (2010). *Allgemeine und fachspezifische Merkmale der Unterrichtsqualität im Fach Biologie. Eine Video- und Interventionsstudie*. Berlin: Logos.

Wüsten, S., Schmelzing, S., Sandmann, A., & Neuhaus, B.J. (2010). Fachspezifische Qualitätsmerkmale von Biologieunterricht. In U. Harms & I. Mackensen-Friedrichs (Hrsg.), *Lehrund Lernforschung in der Biologiedidaktik. „Heterogenität erfassen – individuell fördern im Biologieunterricht": Internationale Tagung der Fachsektion Didaktik der Biologie im VBIO, Kiel 2009* (S. 119–134). Innsbruck: Studienverlag.

### *11.3. Beitrag 3: Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors*

Heinitz, B., & Nehring, A. (2023). Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors. International Journal of Science Education. https://doi.org/10.1080/09500693.2023.2213382

Abstract:
Evaluating and improving instructional quality is important in pre-service teacher education, given it is a crucial factor for students' learning gains. This process is complex and involves multiple classroom events with various interpretations. Criteria for instructional quality are rarely applied systematically in teacher education, leading to divergent evaluations and hindering comparable development. Comparability has rarely been researched in science pre-service teacher education. Therefore, we compared evaluations from 17 chemistry-specific advisors and 17 science pre-service teachers with regard to their choice of criteria, their respective rating and grading of a videotaped chemistry lesson, in Germany. Their evaluations were compared using the Science Education Perspectives (SEP) framework for instructional quality. Although advisors can be considered evaluation experts, our findings show differences in the choice of criteria, ratings, and how the lesson was graded within and between both groups. Pre-service teachers focused more on aspects of classroom and time management, and gave higher average ratings, whereas their advisors focused more on cognitive activation. Overall, 16 different criteria were used by the majority of participants. These criteria show a strong science-specific focus. Our findings have implications for science pre-service teacher education, showing a need for a common approach in evaluations, with extended observation periods.

CRediT Author Statement zur Eigenleistung:
Benjamin Heinitz: Conceptualization**,** Writing — original draft, review & editing, Investigation, Data Curation, Formal analysis, Visualization

International Journal of Science Education

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsed20

# Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors

**Benjamin Heinitz & Andreas Nehring**

**To cite this article:** Benjamin Heinitz & Andreas Nehring (2023): Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors, International Journal of Science Education, DOI: 10.1080/09500693.2023.2213382

**To link to this article:** https://doi.org/10.1080/09500693.2023.2213382

INTERNATIONAL JOURNAL OF SCIENCE EDUCATION
https://doi.org/10.1080/09500693.2023.2213382

Routledge
Taylor & Francis Group

RESEARCH ARTICLE



# Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors

Benjamin Heinitz and Andreas Nehring

Institute of Science Education, Leibniz University Hannover, Hannover, Germany

**ABSTRACT**

Evaluating and improving instructional quality is important in pre-service teacher education, given it is a crucial factor for students' learning gains. This process is complex and involves multiple classroom events with various interpretations. Criteria for instructional quality are rarely applied systematically in teacher education, leading to divergent evaluations and hindering comparable development. Comparability has rarely been researched in science pre-service teacher education. Therefore, we compared evaluations from 17 chemistry-specific advisors and 17 science pre-service teachers with regard to their choice of criteria, their respective rating and grading of a videotaped chemistry lesson, in Germany. Their evaluations were compared using the Science Education Perspectives (SEP) framework for instructional quality. Although advisors can be considered evaluation experts, our findings show differences in the choice of criteria, ratings, and how the lesson was graded within and between both groups. Pre-service teachers focused more on aspects of classroom and time management, and gave higher average ratings, whereas their advisors focused more on cognitive activation. Overall, 16 different criteria were used by the majority of participants. These criteria show a strong science-specific focus. Our findings have implications for science pre-service teacher education, showing a need for a common approach in evaluations, with extended observation periods.

**ARTICLE HISTORY**
Received 21 April 2022
Accepted 9 May 2023

**KEYWORDS**
Pre-service teacher education; instructional quality; chemistry education

## Introduction

Instructional quality is important to successful science teaching and learning. Research evidence demonstrates that high instructional quality leads to higher learning gains (e.g. Kang, 2020; Kyriakides et al., 2013; Labudde et al., 2014). Criteria for instructional quality provide a basis for classroom evaluations by referring to indicators of observable behaviour, which supports pre-service teachers to become good teachers. Science learning and teaching has particular features that distinguish it from other domains (e.g. the choice and use of scientific models or the implementation of experiments) and domain-

**CONTACT** Benjamin Heinitz heinitz@idn.uni-hannover.de Institute of Science Education, Leibniz University Hannover, Am Kleinen Felde 30, Hannover 30167, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/09500693.2023.2213382.

specificity should be taken into account. The importance of domain-specificity is supported by meta-analytical findings (Seidel & Shavelson, 2007) that demonstrate that domain-specific criteria have an especially strong effect on student learning gains.

Criteria-based lesson evaluation is important in pre-service teacher education, but it can be very complex. A multitude of classroom events can be observed, interpretated and evaluated in multiple ways. Pre-service teachers must develop a professional vision for instructional quality and focus on relevant criteria to reflect upon (Sherin, 2001). Kang and van Es (2019) describe a 'shared vision of teaching' as one of the 'worthy goals for preservice education' (p. 243); however, a shared vision of good teaching requires a common set of criteria to ensure consistency across evaluators.

Pre-service teacher education usually consists of individual interactions between pre-service teachers and their advisors centred around lessons and is not often based on standardised criteria. In Germany, where this study was conducted, there are few common instruments or frameworks for guiding lesson evaluations in science pre-service teacher education, which is also the case in many other countries. This raises questions about how comparable standards in science teacher education can be ensured, in the absence of common evaluation frameworks or criteria.

To assess the extent to which science pre-service teachers and their advisors produce consistent evaluations in the absence of specific guidelines, we wanted to find out what criteria they use when no instruments are presented, whether they focus on science-specific criteria, and to what extent the evaluations differ. We compared the evaluations of 17 chemistry-specific advisors and 17 science pre-service teachers referring to the same lesson. All participants were asked to evaluate a chemistry lesson recording in individual interviews. We did not provide them with standardised criteria, but analysed and compared their evaluations with the Science Education Perspectives (SEP) framework of instructional quality (Heinitz & Nehring, 2020). Our aim is to use this analysis to highlight potential problems that might impair the comparative development of instructional quality in science pre-service teacher education and discuss possible reasons. We aim to develop a more consistent approach to the evaluation of instructional quality and highlight domain-specificity for evaluations in science pre-service teacher education by building upon a science-specific framework of instructional quality.

## Theoretical background

Our study is focused on evaluations of instructional quality, which will first be conceptualised in more detail before differentiating between science-specific and generic criteria for instructional quality. Next, we introduce the SEP framework, before discussing possible influences on the evaluations made by pre-service teachers and their advisors.

### *Evaluations of instructional quality*

Instructional quality is comprised of different factors that can be influenced or adjusted by the teacher before or during a lesson, and describes the overall teacher-student interaction. The evaluation of instructional quality in research is often connected to three basic dimensions: 'cognitive activation,' 'classroom management,' and 'constructive support' (Klieme et al., 2001[1]). These dimensions are empirically and theoretically-

based, and offer a generic conceptualisation for instructional quality. They are broadly used and recognised for comparisons of instructional quality; for example, in observational studies like the Programme for International Student Assessment (PISA; Klieme & Rakoczy, 2003) or Cognitive Activation in the Classroom (COACTIV[2]; Kunter & Voss, 2011).

The three basic dimensions are conceptualised as a generic framework, but they are also connected to many science-specific observational instruments (e.g. Korneck et al., 2017; Neuhaus, 2021), although often with science-specific operationalisations. A clear distinction between domain-specific and generic criteria cannot be made universally and can differ between instruments (Heinitz et al., 2022). The SEP framework used in this study makes the distinction by analysing the knowledge necessary to evaluate a criterion. A criterion can be applicable in different domains, if it is described in generic terms, but if the evaluation needs content knowledge or pedagogical content knowledge, it is considered as domain-specific (e.g. Selection and implementation of content). Generic criteria, on the other hand, can be evaluated without domain-specific knowledge (e.g. Supportive teacher-student interaction). Hybrid criteria need domain-specific, as well as pedagogical knowledge (e.g. Individualisation/differentiation).

### *A science-specific conceptualisation of instructional quality – the Science Education Perspectives framework*

The existence of different adaptations of the three basic dimension supports the idea of a further development with regards to operationalisation, domain dependency, and comprehensiveness (Praetorius et al., 2018). This has been highlighted in various studies and is summarised by Praetorius et al. (2020a). The framework for instructional quality by Praetorius and Charalambous (2018) includes seven dimensions and takes a wider approach to instructional quality by building upon the three dimensions.

In previous work, the framework by Praetorius and Charalambous (2018) was applied and adapted for different domains by experts in the respective fields, including physical education and history education, as well as science education (Praetorius et al., 2020b). As a result, we proposed the science-specific SEP framework as an adaption of the generic framework (Heinitz & Nehring, 2020).

The SEP framework is structured along seven generic dimensions which present the first and most general level of the framework: (I) 'Content selection and presentation', (II) 'Cognitive activation', (III) 'Practice', (IV) '(Formative) assessment', (V) 'Cutting-across instructional aspects aiming to maximize student learning', (VI) 'Socio-emotional support' and (VII) 'Classroom and time management'. The second, more detailed level includes 50 criteria further differentiating the dimensions. The criteria are to a large extent science-specific but also include hybrid and generic aspects and are used for all further analysis in this article (see Figure 1).

The SEP framework is also linked to the MAIN-Teach model (Charalambous & Praetorius, 2020) as it contains the same seven generic dimensions of instructional quality. This link can be used to connect science-specific instructional quality to other domains as proposed in recent discussion about initiating more collaborative work in the evaluation of instructional quality (Charalambous et al., 2021), possibly leading towards a synthesised framework (Charalambous & Praetorius, 2022).

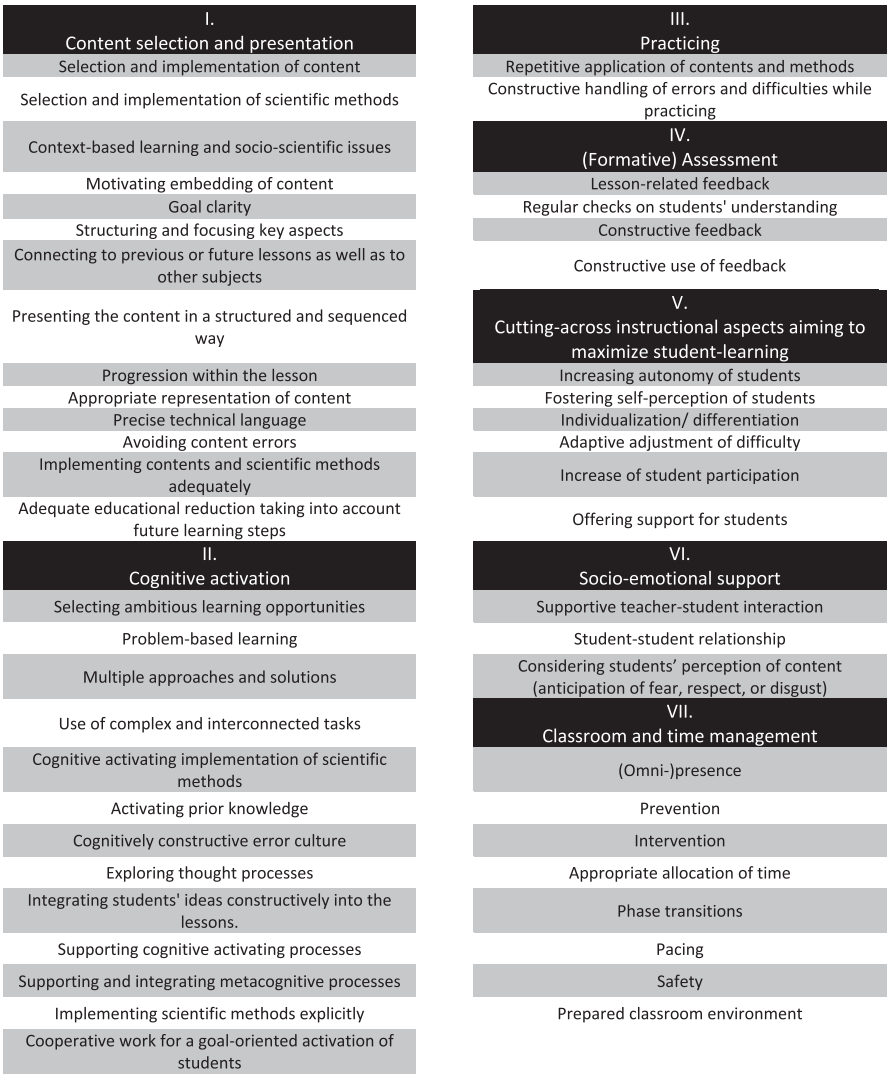



**Figure 1.** Science Education Perspectives framework structured along seven dimensions (I–VII) and including 50 criteria for instructional quality.

### *Different perspectives on instructional quality*

Many studies have shown that individual evaluations of instructional quality vary depending on the perspective of the evaluator. Evaluations of students, teachers, and external observers are often compared to each other, and similarities only occur with regard to some criteria (e.g. Camburn & Barnes, 2004; Clausen, 2002; Desimone et al., 2010; Fauth et al., 2014; Kunter & Baumert, 2006). Individual criteria can be more or less applicable depending on the perspective, which is illustrated by the 'perspective

reference matrix' (Fauth et al., 2020). Furthermore, the rating of a criterion can also be influenced by the rater's position within the evaluation e.g. if their own performance has to be evaluated (Vazire, 2010).

### Challenges when evaluating instructional quality from an outside perspective

Factors other than differing perspectives can lead to divergent evaluations. In a comparative study by Strong et al. (2011), none of the groups involved in teacher education (college students, mentors, teachers, teacher educators, and university professors) were able to unanimously identify the lesson containing 'good teaching'. Criteria for instructional quality can be interpreted differently, as described by Praetorius et al. (2012), who summarised these different ideas as 'implicit theories about good instruction'. In their study, up to ten raters were necessary for a reliable evaluation of instructional quality, depending on the criterion. Taut and Rakoczy (2016) described similar 'cognitive schemas and biases about 'good instruction' that differ across evaluators' (p. 56) and led to different evaluations and opinions about the lesson. Although all of the raters included in that study were full-time evaluators who had experience with school inspections, it was not possible to get generalisable results for the dimension of cognitive activation. This is consistent with Praetorius et al. (2014), who also found that a longer observation period may be necessary for some criteria (e.g. up to nine lessons for cognitive activation).

### The importance of consistent evaluations for pre-service teacher education

In a typical teacher training situation, both pre-service teachers and their advisors evaluate the same lesson and compare their views to inform the pre-service teacher's ongoing professional learning. This process is similar to the diagnosis of classroom instruction, as described by Helmke and Lenske (2013), where multiple perspectives on the same lesson are used for the development of instructional quality. The authors emphasise that some divergence in evaluations of instructional quality can be productive. However, it is unclear whether the divergence has a limit to still be productive.

To the best of our knowledge about the German system, there is no systematic procedure for fostering instructional quality in pre-service teacher education. There are some guidelines available in Germany (e.g. APVO-Lehr[3]), but they are often vague and vary between different states. Other countries (e.g. the USA) have similarly open and decentralised guidelines for teacher education (Tatto, 2021), which might lead to differences in teacher preparation programmes, even within the same state or educational jurisdiction (e.g. Boyd et al., 2009). In many cases, neither pre-service teachers nor their advisors are provided with standardised frameworks or instruments to evaluate instructional quality. This might cause difficulties in evaluations, gradings and in communication about instructional quality.

The grading of pre-service teachers might be dependent on individual advisors, since they are free to focus whichever criteria they prefer in their evaluations. Many studies show generalised problems in evaluations of instructional quality. It might be necessary to consult multiple raters (Praetorius et al., 2012), observe over a longer period of time (Praetorius et al., 2014), consider the perspective of the rater (Fauth et al., 2020), or use anchor ratings (Wind et al., 2021), depending on different criteria.

It seems plausible that science pre-service teacher education would benefit from instructional quality being evaluated in a more standardised way. Domain-specificity adds an additional layer of complexity for science pre-service teachers, noting the necessity of science content knowledge and pedagogical content knowledge. Evaluations should generally be transparent to improve communication about instructional quality and make it more consistent. If domain-specific advisors had a common basis for their evaluations, pre-service teachers could receive consistent training. They could build upon a common basis to plan their lessons, and the lessons would provide equal learning opportunities for all students. The grades pre-service teachers receive from their advisors would be comparable between different locations and less dependent on individual advisors, which has often been criticised in German teacher education (Döbrich & Storch, 2012).

## Research questions

Based on these considerations, our research questions were:

(1) Which criteria are used by science pre-service teachers and their advisors to evaluate a chemistry lesson, if no instrument is provided?
(2) To what extent do science pre-service teachers and their advisors differ in their evaluation of individual criteria and the lesson as a whole?
(3) How do ratings of science pre-service teachers and their advisors compare within their groups?

## Study context

Pre-service teacher education in Germany is divided into two phases. The first phase is based on university education and is completed with a Master of Education degree. During this phase, initial practical experience is gained in school internships, but there is a stronger focus on theoretical knowledge. The second phase is oriented towards teaching practice in a school. Compared to in-service teachers, pre-service teachers have a reduced number of lessons and are supported by experienced teachers outside of lessons. In addition, they attend seminars by domain-specific advisors where they receive general advice on teaching. At regular intervals, the domain-specific advisors observe lessons and give feedback on how to improve instructional quality. At the end of the second phase, all pre-service teachers have to give an examination lesson and are graded by several people, including their domain-specific advisors.

## Study design

Our study was designed to be comparable to the 'real-life'-experience of pre-service teachers and their advisors, with a view to ensure ecological validity. According to Holleman et al. (2020) the term ecological validity is oftentimes too vague to stand on its own, thus it will be specified for our study. We simulated the work environment of domain-specific advisors during classroom observations and pre-service teachers when observing their colleagues' lessons. We presented a videotaped chemistry lesson without the possibility of

stopping or rewinding. The camera was static and set in the back of the classroom for a complete overview. The only movements were occasions where the camera zoomed in on the whiteboard. All participants were provided with a written summary of the teachers aims, lesson plan, and anticipated student outcomes. Additionally, contextual information about the previous lessons and the performance of the class was provided. Comparable context information is usually given to advisors before an examination lesson. The participants received no criteria for the evaluation of instructional quality, but they were allowed to use any lists of criteria they would usually bring to a classroom observation.

### *Videotaped chemistry lesson*

The basis for data collection was a videotaped chemistry lesson (cut down to a 25-minute excerpt) showing a pre-service teacher in a seventh-grade classroom (students aged 12–13). The lesson addressed the reversibility of chemical reactions and included an experiment in which silver was extracted from silver oxide. The teacher provided a context-based introduction and presented a problem. The students had to formulate hypotheses and test them with an experiment. The experiment was predefined by the teacher and included heating silver oxide, followed by a test for oxygen. Afterwards, the results were discussed and all observations were linked to scientific explanations as well as to the context-based problem presented at the beginning.

## Methods

### *Participants*

In total 17 chemistry-specific advisors and 17 science pre-service teachers participated in this study. The advisors were situated in four different federal states in Germany (Berlin, Lower Saxony, North Rhine-Westphalia, Schleswig-Holstein) and all had chemistry as one of their specific domains. All chemistry-specific advisors worked at secondary education schools and had gained several years of professional experience as teachers before becoming advisors. At the time of the interview, they all had at least a year of experience as chemistry-specific advisors, with some working in that position for more than a decade. The participants were selected by opportunity sampling and volunteered to take part in this study. The 17 pre-service teachers were all situated in the same federal state (Lower Saxony). They were included in the study through the voluntary involvement of three advisors, who agreed to include our study in their seminars. The pre-service teachers did not register for the study on their own but could object to participation. Sixteen of the pre-service teachers had a finished Master of Education with chemistry as one of their two subjects. One pre-service teacher changed profession and entered the second phase of pre-service teacher education without a master's degree in education. All of them were in the second phase of pre-service teacher education and had been working in secondary education schools for between 6 and 18 months.

### *Data collection*

After watching the videotaped lesson, interviews were conducted with each participant. The interviews were guided by a central question: 'How would you evaluate the lesson, if

this was a classroom visit as part of the pre-service teacher's education?', with no further restrictions. All participants were familiar with the general procedure of these classroom visits. They were free to focus on whichever aspects of the lesson they deemed to be important, as they would usually do, and could choose whichever criteria they wanted. If participants made statements that were unclear, they were asked to elaborate. At the end of the evaluation, all interviewees were asked to grade the lesson using the six-point scale regularly used for the final exam of pre-service teachers. This scale ranges from 'very good' (1) to 'insufficient' (6). To pass the examination, a minimum grade of 'sufficient' (4) is necessary.

We explicitly decided not to use a standardised set of criteria in order to create an ecologically valid simulation of the typical way in which evaluations of instructional quality take place. Neither pre-service teachers nor their advisors use standardised instruments; the evaluation process is rather free and open. However, the participants were free to use any instruments or frameworks they would usually use.

The advisors each watched the video on their own before being interviewed. The pre-service teachers all watched the video in their seminar (3 different groups) but were not allowed to exchange their ideas and were then interviewed separately in private rooms. The average length of interviews was 60 min (58 min for pre-service teachers and 62 min for advisors), and ranged from 28 to 108 min ($Mdn = 61$ min).

### Data analysis

All interviews were audio-recorded, transcribed, and anonymized. Each statement made by the participants was divided into 'idea units' (Jacobs & Morita, 2002). An idea unit represents a coherent section of the conversation without changing the focus or topic. Each idea unit was assigned to one of the 50 criteria in the SEP framework (Figure 1). In addition, the rating[4] for each criterion (positive/neutral/negative) was also coded. Overall, this included 2735 idea units each assigned to a criterion and a rating. To test for interrater reliability, a total of 14% of the data set was coded by three raters, with rater one coding everything and raters two and three each coding 7% of the data set. The raters reached a relative agreement between 69% and 86%. Considering the number of categories, the agreement by chance is not considered to be very high. The Cohen's Kappa coefficients confirms this ($\kappa = 0.63–0.84$) and they fall within the range of 'substantial' to 'almost perfect'-agreement (Landis & Koch, 1977).

On average there were 65 idea units per interview (56 for pre-service teachers and 74 for advisors). However, this number was highly dependent on the individual interviews and the absolute number of idea units is thus not taken into account for the analysis. In some cases, a single criterion was mentioned multiple times at different points during the interview. In other cases, the interviewees gave an in-depth analysis using a specific criterion but then never mentioned it again. Although the criterion might be explained in the same way across interviewees, the number of idea units could differ. Since we did not want the individual styles of conversation to have an impact on our comparison, the number of times a criterion was mentioned in a single interview is not considered. For all comparisons between different interviews, we focused on the number of different criteria that were used by each participant and how they were rated.

A mean rating of all criteria was calculated using all idea units for each interview (see Table 1). We used a relative ratio of ratings, so the number of idea units did not influence the mean rating. Overall, there is only one rating score for each criterion used in a single interview. For intra- and intergroup comparisons, we only considered criteria with a majority use (at least 50% of participants). This limitation was set to focus on commonly-used criteria and prevent rarely-used criteria from having too much of an impact on the analysis. However, mean ratings for all criteria are reported in the supplementary material. The coding of all idea units was carried out according to the principle of qualitative content analysis (Mayring, 2014), whereby the deductive categories were derived through the SEP framework (Figure 1) and transferred into a coding manual through short descriptions. For example, the criterion 'selection and implementation of content' would be coded every time

> the statement refers to the subject matter and its appropriateness in the observed lesson. Both the basic selection and the actual implementation into the lesson can be considered. Possible aspects are, for example, complexity of the content, comprehensibility from the point of view of students, existence of necessary subject-related prerequisites or suitability of an experiment in terms of its content. (Translated excerpt of the coding manual)

## Results

The results are divided into several sections, starting with the relative frequency with which different criteria were used, followed by between-group comparisons of ratings and the standard deviations of those ratings within groups. We focus on criteria with a majority use in both groups for all further comparisons. For a more extensive comparison, the relative frequencies of use, mean ratings, and standard deviations of all criteria are presented in tabular form in the supplementary material.

### *Criteria for instructional quality in pre-service teacher education*

The criteria that were used most frequently are located in dimension I 'content selection and presentation'. These were used by 77% of the advisors and 67% of the pre-service teachers (see Table 2). Although this result appears to show a common focus, it is also noticeable that there is a great deal of variability, and that some criteria were rarely used. In total, the advisors used 47 of the 50 criteria to evaluate the lesson, with a

**Table 1.** Calculation of mean rating for each criterion.

| Requirements for each level of evaluation within a single interview | 5 Levels of evaluation |
|---|---|
| No negative statements<br>Max. 33% neutral statements | (2) Completely positive |
| Min. 66% positive statements<br>Max. 33% negative or neutral statements | (1) Overwhelmingly positive |
| Max. 66% positive or negative statements OR only neutral statements | (0) Neutral |
| Min. 66% negative statements<br>Max. 33% positive or neutral statements | (−1) Overwhelmingly negative |
| No positive statements<br>Max. 33% neutral statements | (−2) Completely negative |

range of 20–32 different criteria (average = 24). The pre-service teachers used 39 of the 50 criteria, with 14–25 different criteria mentioned per interview (average = 19). Overall, the criteria in dimension II ('cognitive activation') were mainly used by the advisors. The majority of advisors used 7 of the 13 criteria from dimension II, whereas only 2 of the 13 criteria were used by the majority of pre-service teachers (supplementary material 1). Just under half (47%) of the advisors, 24% of the pre-service teachers, used criteria from dimension II. In contrast, pre-service teachers used criteria from dimension VII ('classroom and time management') more frequently than the advisors, although the latter show a higher overall use of different criteria. There are 16 criteria in total that are used by the majority of both groups, which will be focused in the comparison of ratings (see Table 2). Most of these criteria focus on science-specific or hybrid aspects of instructional quality.

### *Rating of instructional quality for each criterion*

The mean ratings for each criterion, across all participants, are shown in Table 2. The first thing to notice is that pre-service teachers ($M = -0.03$, $SD = 0.72$) rated the lesson more positively than the advisors ($M = -0.62$, $SD = 0.56$). An independent-samples t-test showed that there was a significant difference between the two groups ($t(32) = 2.70$, $p = .006$) with a large effect size (*Cohen's* $d = 0.93$).

**Table 2.** Summarised frequency of use, mean rating and standard deviation of all criteria used by the majority of domain specific advisors and pre-service teachers.

| | Frequency of use | | Mean Rating | | Standard deviation | |
|---|---|---|---|---|---|---|
| Criterion | Advisors | Pre-service teachers | Advisors | Pre-service teachers | Advisors | Pre-service teachers |
| **Dimension I total** | 77% | 67% | −0.70 | 0.00 | 1.1 | 1.3 |
| Selection and implementation of content | 100% | 100% | −0.90 | −0.20 | 0.9 | 0.4 |
| Selection and implementation of scientific methods | 100% | 100% | −0.40 | 0.00 | 1.2 | 1.0 |
| Motivating embedding of content | 100% | 71% | 0.10 | 1.00 | 1.7 | 1.3 |
| Structuring and focusing key aspects | 94% | 94% | −0.80 | −0.40 | 1.5 | 1.7 |
| Presenting the content in a structured and sequenced way | 94% | 53% | −0.60 | −0.90 | 1.2 | 1.4 |
| Precise technical language | 88% | 76% | −0.50 | 0.30 | 1.6 | 1.5 |
| Progression within the lesson | 82% | 82% | −1.00 | 0.30 | 1.5 | 1.5 |
| Appropriate representation of content | 82% | 100% | −0.60 | 0.60 | 1.3 | 1.5 |
| Context-based learning and socio-scientific issues | 76% | 88% | −0.30 | 0.70 | 1.1 | 1.4 |
| Goal clarity | 65% | 71% | −1.80 | −0.70 | 0.4 | 1.7 |
| **Dimension II total** | 47% | 24% | −0.30 | −0.20 | 0.8 | 1.1 |
| Integrating students' ideas constructively into the lessons. | 100% | 100% | −1.20 | −0.40 | 0.7 | 1.2 |
| Problem-based learning | 53% | 65% | 0.60 | 0.00 | 1.6 | 1.5 |
| **Dimension V total** | 49% | 37% | −0.90 | −0.40 | 1.2 | 1.1 |
| Increase of student participation | 94% | 82% | −0.90 | −0.40 | 1.2 | 1.3 |
| **Dimension VI total** | 27% | 18% | 1.60 | 1.70 | 1.1 | 0.7 |
| Supportive teacher-student interaction | 71% | 53% | 1.60 | 1.70 | 1.1 | 0.7 |
| **Dimension VII total** | 33% | 37% | −0.50 | −0.70 | 1.2 | 0.8 |
| Appropriate allocation of time | 82% | 59% | −1.60 | −1.50 | 0.9 | 0.8 |
| Safety | 53% | 59% | 0.60 | 0.10 | 1.6 | 1.1 |

Note: Additional information for all 50 criteria of the framework is provided in the supplementary material.

The largest difference in mean ratings can be found for the criterion 'progression within the lesson.' (advisors $M = -1.0$; teachers $M = 0.3$, $t(26) = 2.22$, $p = .035$). Other large differences were found for 'appropriate representation of content' (advisors $M = -0.6$; teachers $M = 0.6$, $t(29) = 2.29$, $p = .030$) and 'goal clarity' (advisors $M = -1.8$; teachers $M = -0.7$, $t(12.24) = 2.19$, $p = .049$). If the two groups are further compared, commonalities can also be identified. For example, the criterion 'supportive teacher-student interaction' is rated positively by both groups. The criterion 'appropriate allocation of time' is clearly rated as negative by both groups. Overall, there is also a general tendency towards 'neutral' for almost all criteria.

For the final grading of the lesson, using the six-point scale used in pre-service teacher education, the advisors' ratings range from 'good' (2) to 'insufficient' (6) ($Mdn$[5] = 4) whereas the pre-service teachers' ratings ranged from 'very good' (1) to 'sufficient' (4) ($Mdn = 3$). The pre-service teacher shown in the recorded lesson was essentially failed by 7 of the 17 advisors, with a grade worse than 'sufficient' (4), but would have been passed by all the pre-service teachers, albeit with differences in grades. A Mann–Whitney U Test showed a significant difference in grading between the two groups with a large effect size ($Z = -3.42$, $p < .001$, $r = -.59$). A Spearman's rank correlation between the mean rating of individual interviewees and their final grade awarded was also calculated. Overall, there was a strong correlation between ratings and grades ($r(32) = -.82$, $p < .001$).

### *Standard deviation in mean rating*

To get a better impression of differences within the groups, the standard deviations of mean ratings (Table 2) were also analysed. Since the rating only includes five levels, a standard deviation of one level already represents a large difference. Larger standard deviations are an indication of strong intra-group differences when the mean rating is within the 'neutral' range (−0.5 to 0.4). Particularly noteworthy are criteria with the largest standard deviation of 1.7: 'motivating embedding of content,' 'goal clarity,' and 'structuring and focusing key aspects'. In the case of 'motivating embedding of content', a large difference becomes apparent on the side of the advisors, since the criterion was rated on average as 'neutral' (0.1) and the standard deviation thus covers the scale from 'completely negative' (−1.6) to 'completely positive' (1.8). The pre-service teachers do not show such a big difference for this criterion. The criterion 'goal clarity' also represents an interesting case, since the advisors show a low standard deviation (0.4) and give an average rating of 'completely negative' (−1.8). All ratings within one standard deviation remain 'completely negative' for this group. The pre-service teachers, on the other hand, are divided between the ratings 'completely negative' (−2) and 'predominantly positive' (1).

Both groups reached a standard deviation of at least 1.5 for 'structuring and focusing key aspects', 'progression within the lesson', 'precise technical language' and 'problem-based learning'.

## Discussion

We present summary answers to the research questions and discuss our results concerning science pre-service teacher education against the background of the existing

literature. We then derive implications for the evaluation of instructional quality in pre-service teacher education in general.

### *Advisors tended to go into a deeper application of science-specific criteria while pre-service teachers tended to stay on the surface structure of instructional quality*

The analysis of our first research question ('Which criteria are used by science pre-service teachers and their advisors to evaluate a chemistry lesson, if no instrument is provided?') shows that a total of 16 criteria were used by the majority of both groups to evaluate the instructional quality of the presented chemistry lesson. Most of these criteria belong to just two dimensions ('content selection and presentation' and 'cognitive activation') and show a strong science-specific focus. This finding underlines the importance of domain-specific criteria. Both groups made frequent reference to the 'selection and implementation of scientific methods', 'precise technical language' and 'integrating students' ideas constructively into the lessons'. The majority of chemistry-specific advisors used significantly more criteria from dimension II ('cognitive activation') than the science pre-service teachers. The latter showed a stronger focus on dimension VII ('classroom and time management'), which has more generic components and might be more easily accessible on a surface level. This difference can have a strong influence on the evaluation of classroom practice. Even without pre-set criteria, common criteria could be established. These criteria were, to a large extend, science-specific and include references to both content knowledge and pedagogical content knowledge. This necessary knowledge must be considered when science pre-service teachers receive feedback on their lessons, since it might be a limiting factor for successful communication. Aside from this common basis, however, there are many individual differences in the choice of criteria with a tendency for advisors to use a larger number of criteria.

### *Expertise impacts the choice of criteria*

The difference in selections of criteria between the groups might be explained by their positions within the educational system. It is quite understandable that a science pre-service teacher is initially concerned that the lesson runs smoothly and prioritises criteria of 'classroom and time management'. For an experienced teacher, these aspects may be self-evident and other aspects, such as 'cognitive activation,' are considered to a greater extent when evaluating a lesson. Since the choice of criteria was completely open, it will have been influenced by their prior knowledge, experiences, and understanding of instructional quality. Differences in the selection of criteria might also indicate that science pre-service teachers have less knowledge about certain (science-related) aspects of instructional quality like 'cognitive activation'. This generalised group comparison shows parallels with Berliner's (1989) expert-novice description. Experts tended to focus their attention on key aspects that directly affect their lesson goals, whereas novices tended to attend to more superficial aspects of the lesson without specific focus. Therefore, evaluations by advisors might not be transparent for science pre-service teachers and should always be further elaborated, even when referring to specific criteria.

The advisors showed a broader view of instructional quality which allows them to work with a wide variety of pre-service teachers and individually focus their education. However, since the participants all watched the same lesson and received the same background information, one might expect the advisors to focus on the same key aspects of the lesson: This missing shared focus seems to contradict the expert-novice definition according to Berliner (1989). However, Berliner also pointed out that expertise in the evaluation of instructional quality cannot necessarily be transferred to unknown classes.

### *Science pre-service teachers rated the lesson more positively than their advisors overall, and in relation to specific criteria*

The analysis of our second research question ('To what extent do science pre-service teachers and their advisors differ in their evaluation of individual criteria and the lesson as a whole?') showed a significant difference with a large effect size between the two groups when comparing the evaluations. Despite the individual evaluations seen in our study, there seem to be commonalities within the groups that cause them to diverge in a generalised comparison.

#### *Group differences in rating might be based on protective bias towards the teacher*

Even though both groups in our study looked at the lesson from the perspective of an external observer, they may interpret the information according to their position in teacher education. Fauth et al. (2020) illustrate these different approaches with their 'perspective reference matrix,' in which one of the factors is that evaluations can be judgmental. In our study, all raters were external observers, but the recording shows a pre-service teacher. The pre-service teachers gave more positive evaluations and better grades than the advisors. This may be due to a positive or protective bias towards the teacher being evaluated. Vazire (2010) discussed similar biases for self-evaluations. Since all pre-service teachers share a position within the teacher education system such biases might be present in our study and might also impact other evaluations.

#### *Expert evaluations of instructional quality might include implicit evaluations and 'encapsulation'*

Implicit evaluations by advisors might pose an additional challenge for pre-service teacher education. It is not uncommon for experts to use implicit knowledge without verbalising it (Berliner, 1989). Evaluations might be based on encapsulated knowledge comparable to the diagnostic expertise of medical practitioners (Schmidt & Rikers, 2007). However, unlike medical diagnosis, evaluations of instructional quality are not standardised to the extent that everyone has the same understanding of technical terms. This makes it necessary to further elaborate technical terms in evaluations of instructional quality.

Implicit theories about good teaching, as described by Praetorius et al. (2012), might have influenced the selection of criteria used to evaluate the target lesson. This raises questions about the validity of evaluations in pre-service teacher education when no criteria are given, which is often the case. All evaluations should be comparable if there is to be equality in pre-service teacher education, but our results show that there are differences in which criteria were chosen. Resuming the medical diagnoses metaphor, we

found that the advisors made a variety of different diagnoses for a single patient. Their implicit theories might connect different criteria even when they are not explicitly mentioned, but this is not accessible from an outside perspective (e.g. to the pre-service teachers).

### *Even though a general distinction between the ratings of pre-service teachers and their advisors is possible, there are many individual differences within the groups*

The analysis of our third research question ('How do ratings of science pre-service teachers and their advisors compare within their groups?') identified 16 criteria that were shared by all participants, but the ratings differed for almost all of them. This becomes particularly clear when looking at the standard deviations within the groups. Nevertheless, there was a strong correlation between mean rating and grading. This result is comparable to the findings of Wind et al. (2021) where the raters also showed individual variance but an overall internal consistency in their ratings of different teachers.

### *Implication 1: grades based on classroom observations might not necessarily be comparable, even when they are based on the same grading system*

According to the results of Praetorius et al. (2014), a more uniform picture would emerge if several lessons were observed. Generally, this would call for a change in the current system of pre-service teacher education if all criteria are to be rated reliably. Alternatively, the evaluation of several raters could be combined to obtain a more reliable evaluation of a lesson. According to Praetorius et al. (2012), this would require at least four raters for evaluations of 'classroom management' and at least ten for 'learner support'. Wind et al. (2021) analysed differences between raters and were able to adjust their ratings so that only one rater would be necessary. However, this was only possible because all raters conducted an anchor rating of the same lesson and could therefore be compared. Rater training for standardised instruments (e.g. the Classroom Assessment Scoring System (CLASS)) include an expert rating that new raters have to match in order to reliably use the instrument, but the training is limited to a specific instrument.

The evaluation of a single lesson conducted by one or even a few raters does not seem appropriate for creating a comprehensive picture of a pre-service teacher's performance, even when pre-set criteria are used. To again compare our findings to a medical diagnosis, even when a group agrees on a patient diagnosis, they may interpret it differently: Some might focus on the same symptoms but rate them differently, while others may focus on different symptoms, altogether. The whole concept of an examination lesson does not seem to be suitable to evaluate a pre-service teacher. Too many factors lead to too many differences to create a valid impression of the teacher's abilities. The wide range of grades given to the same lesson seen in our study confirms this. Standardised training for raters may be necessary, but a common basis for evaluating pre-service teachers would first need to be established.

### *Implication 2: pre-service teacher education should acknowledge aspects of individual development*

Looking into individual evaluations, it can always be debated whether a single person perfectly fits into a certain group or not. This may be due to a similar perspective, similar ideas about 'good teaching', similar expertise or just a case of a margin value within a normal distribution. However, this opens the question on how close evaluations are on an individual level whenever two individuals talk about instructional quality. Pre-service teacher education can be oriented on group-specific differences for group activities (e.g. seminars and theoretical groundwork) but has to acknowledge individual differences for situations of individual development (e.g. classroom observations, evaluations and feedback). This also includes content knowledge or pedagogical content knowledge that is necessary for understanding and using domain-specific criteria.

### *Implication 3: a wide range of different feedback could be beneficial for individual development*

Differences in choice and rating of criteria might reduce the comparability of evaluations. However, classroom observations offer a good basis for general feedback and self-improvement, since different perceptions of the same criteria might offer new insights. This would support the idea of diagnosis of classroom instructions for pre-service teacher education (Helmke & Lenske, 2013). Instructional quality is not standardised to the extent of medical diagnosis. Comparable to the analysis of a new treatment, it can be helpful to focus on different aspects of the same case or hear different interpretations of occurring symptoms.

### *Implication 4: a common science-specific framework for evaluation is necessary for science pre-service teacher education*

Based on our results, we conclude that general frameworks for instructional quality that do not include science-specific criteria may not be suitable for evaluating science teaching. The divergent focus between the two groups in our study could be brought together by using the SEP framework as a common basis. Using a single framework, rather than a collection of different instruments, offers the opportunity to connect different evaluations of instructional quality. Connections between criteria can also be easily explained to science pre-service teachers and reflections about their own instructional quality can be focused on specific criteria without losing the big picture. Good and Lavigne (2015) clearly emphasise that oversimplifying the scope of an evaluation does not provide an optimal opportunity of development for the teacher being evaluated. In their study of written feedback from pre-service teacher education, Puttick and Wynn (2020) found that feedback was generally superficial, not specific to the subject, and rarely based on concrete actions. Overall, the argumentation was more practice-oriented and less theory-driven. A connection to a standardised framework would be one way of improving the quality of such feedback. A unified basis for a more theory-driven evaluation could lead to a transparent process that is less influenced by individual ideas about good teaching. A common basis and a constant exchange would also help to reduce

the perceived differences in the training and grading of pre-service teachers that were found in previous studies (Döbrich & Storch, 2012). Pre-service teacher education in general should not be set to arbitrary standards depending on the current mentors, advisors, or the different frameworks used by raters, but rather on a common basis that is used to advance the field in unison.

### *Implication 5: specific difficulties in evaluations should be addressed in pre-service teacher education*

Even when comparing commonly used criteria, the evaluations made by our participants showed both intra- and inter-group differences. Some of these differences can be explained by taking a deeper look into criteria with the highest standard deviations in our study. We summarised them into three aspects, that might have implications for other evaluations of instructional quality in pre-service teacher education, as well as classroom observations in general:

1. *Deviations through multifaceted interpretations of a criterion:* Regarding the criterion of 'structuring and focusing key aspects,' we found that statements fell into two categories: 'structuring of key aspects to summarise important aspects of the lesson' and 'focusing key aspects without deviation to other topics'. This example shows that raters might focus on different aspects within a criterion, if no further indicators are presented. Thus, it seems to be important that all raters are asked about their understanding of the criteria. Their explanations could then be used as an anchor to compare different evaluations. A common understanding of criteria and technical language is needed to enable communication and develop instructional quality. This also includes considering necessary content knowledge or pedagogical content knowledge when evaluating science-specific criteria.
2. *Deviations through limited access for external observations:* The criterion 'progression within the lesson' was rated mainly with respect to students achieving the lesson goals. This criterion is only accessible to a limited degree, but most raters gave an evaluation without mentioning this limitation. They either used the answers given by individual students as evidence that the goal was achieved, or used the fact that only a few students answered the questions at the end of the lesson as an indicator that most of them did not progress within the lesson. If there is only a limited access to a criterion, it should be rated carefully. The use of predefined indicators, or the request for raters to give indicators for their evaluations, might lessen the impact of individual interpretations. This should also create more transparency for pre-service teachers and help them in their reflection process. Indicators of criteria for instructional quality should be introduced and discussed in pre-service teacher education seminars. The pre-service teachers in this study focused on surface-level criteria. This focus could be broadened if pre-service teachers received examples and indicators for more complex criteria. Furthermore, criteria that cannot be fully rated in a single lesson should not be considered for an examination lesson of pre-service teachers.
3. *Deviations through biases caused by specific criteria:* The use of the criterion 'precise technical language' presented an interesting example, since it is easily accessible but still showed one of the largest standard deviations in mean ratings. Comparing the rating of this criterion to the mean rating for all individual participants, however, shows that

positive evaluations generally also include a more positive rating for this criterion. It may be possible that certain criteria create an individual bias towards the lesson that impacts the rating of other criteria. 'Precise technical language' should be accessible on a surface level and thus we would expect a similar rating from all participants. Biases should generally be considered in pre-service teacher education. They might be present even when rating criteria that seem to be easily accessible.

### Limitations

Although the results of our study are in accordance to current findings, some limitations have to be taken into account. First, the choice of criteria to evaluate the lesson may have been influenced by the lesson itself. We used a chemistry lesson, but other scientific domains (e.g. biology and physics) should also be examined to create a comprehensive picture. It may also be possible that another lesson encourages the use of different criteria or a weaker focus on science-specific criteria. The recording chosen for this study shows a chemistry lesson structured around an experiment. This is a common practice among many chemistry teachers and was therefore chosen to represent a typical lesson. Some criteria might have not been used since they are only accessible over a longer period of time and cannot be rated on the basis of a single lesson. Attempts to rate these criteria might have led to some of the differences seen in the evaluations.

Second, all comparisons are based on interpretations of statements about instructional quality. The interviewees described their evaluations and used criteria and terminology of their choice. Whenever statements were unclear, the interviewees were asked to elaborate. A certain degree of interpretation was still necessary and might have caused some of the divergence within and between the groups. However, the variance shown for certain criteria can also be seen in other studies with set criteria (e.g. Praetorius et al., 2012, 2014; Strong et al., 2011; Wind et al., 2021).

Third, our study design offered an insight into the evaluation process of pre-service teachers and their advisors but did so in a rather open way for reasons of ecological validity. All findings should still be confirmed with an additional study providing all participants with pre-set criteria or instruments. This should also include the provision of additional context information about the lesson. Although information about the class and previous lessons was provided to all participants, some criteria might need even more information to be evaluated. It might also be interesting to take a look into the process of finding a common grade for multiple raters in an additional study. Pre-service teachers final grading is usually determined by several raters, who might influence each other and offer additional information about their evaluation process.

### Conclusion

The findings of this study suggest that science pre-service teacher education would benefit from a more standardised approach to evaluating instructional quality. Science teacher educators and pre-service teachers alike could benefit from shared criteria to establish a common understanding on which to evaluate and reflect upon instructional quality. In particular, science-specific criteria that capture the nuances of quality science teaching are required, with a view to ensure comparability among science pre-service teacher education programmes.

## Notes

1. For an English translation and explanation of the three basic dimensions see Praetorius et al. (2018).
2. Full title: **Co**gnitive **Activ**ation in the Classroom: The Orchestration of Learning Opportunities for the Enhancement of Insightful Learning in Mathematics.
3. Verordnung über die Ausbildung und Prüfung von Lehrkräften im Vorbereitungsdienst (Regulation for the training and examination of pre-service teachers).
4. 'Rating' always refers to the judgement for a criterion (positive/neutral/negative) and the calculated scores. Higher scores equal a more positive rating (see Table 1).
5. Median is reported due to ordinal scaling of grading.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Ethics statement

All participants were informed about the use of their data and were told they could terminate their participation at any point and gave their oral consent before the study. All participants were informed when the audio recording started and stopped. All data were anonymized after the transcription. There were no further ethics requirements from our institution.

## ORCID

*Benjamin Heinitz* http://orcid.org/0000-0001-8626-6587
*Andreas Nehring* http://orcid.org/0000-0002-8723-5552

## References

Berliner, D. C. (1989). Implications of studies of expertise in pedagogy for teacher education and evaluation. In *New directions for teacher assessment: Proceedings of the 1988 ETC invitational conference.* Educational Testing Service.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440. https://doi.org/10.3102/0162373709353129

Camburn, E., & Barnes, C. A. (2004). Assessing the validity of a language arts instruction log through triangulation. *Elementary School Journal*, *105*(1), 49–73. https://doi.org/10.1086/428802

Charalambous, C. Y., & Praetorius, A. K. (2020). Creating a forum for researching teaching and its quality more synergistically. *Studies in Educational Evaluation*, *67*(1), 100894. https://doi.org/10.1016/j.stueduc.2020.100894

Charalambous, C. Y., & Praetorius, A. K. (2022). Synthesizing collaborative reflections on classroom observation frameworks and reflecting on the necessity of synthesized frameworks. *Studies in Educational Evaluation*, *75*(1), 101202. https://doi.org/10.1016/j.stueduc.2022.101202

Charalambous, C. Y., Praetorius, A.-K., Sammons, P., Walkowiak, T., Jentsch, A., & Kyriakides, L. (2021). Working more collaboratively to better understand teaching and its quality: Challenges faced and possible solutions. *Studies in Educational Evaluation*, *71*(1), 101092. https://doi.org/10.1016/j.stueduc.2021.101092

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstukt - und Merkmalsvalidität. [Teaching quality: A question of perspective? Empirical analyses of agreement, construct and characteristic validity]*. Waxmann.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy*, *24*(2), 267–329. https://doi.org/10.1177/0895904808330173

Döbrich, P., & Storch, H. (2012). Pädagogische EntwicklungsBilanzen mit Studien-SEMinaren oder: Lehrerausbildung ohne Bilanzierung? [Pedagogical DevelopmentBalances with Study SEMinars or: Teacher Education without Balancing?]. Frankfurt, Main: GFPF; DIPF 2012, 75, [87]. Materialien zur Bildungsforschung 31/1. https://doi.org/10.25656/01:5987

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift für Pädagogik. Beiheft*, *66*(1), 138–155. https://doi.org/10.25656/01:25870

Good, T. L., & Lavigne, A. L. (2015). Rating Teachers Cheaper, Faster, and Better: Not So Fast. *Journal of Teacher Education*, *66*(3), 288–293. https://doi.org/10.1177/0022487115574292

Heinitz, B., & Nehring, A. (2020). Quality of instruction in science education—a systematic review including goals, contents, and methods of science education. *Unterrichtswissenschaft*, *48*(3), 319–360. Springer VS. https://doi.org/10.1007/s42010-020-00074-8

Heinitz, B., Szogs, M., Förtsch, C., Korneck, F., Neuhaus, B. J., & Nehring, A. (2022). Unterrichtsqualität in den Naturwissenschaften. Eine vergleichende Gegenüberstellung von Ansätzen zwischen Fachspezifik und Generik. *Zeitschrift Für Didaktik Der Naturwissenschaften*, *28*(1), 10. https://doi.org/10.1007/s40573-022-00146-5

Helmke, A., & Lenske, G. (2013). Unterrichtsdiagnostik als Grundlage für Unterrichtsentwicklung. [Diagnosis of classroom instruction from different perspectives as a prerequisite for improving teaching and learning]. *Beiträge zur Lehrerbildung*, *31*(2), 214–233. https://doi.org/10.36950/bzl.31.2013.9653

Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity. *Frontiers in Psychology*, *11*(Article 721), 1–12. https://doi.org/10.3389/fpsyg.2020.00721

Jacobs, J. K., & Morita, E. (2002). Japanese and American Teachers' Evaluations of Videotaped Mathematics Lessons. *Journal for Research in Mathematics Education*, *33*(3), 154. https://doi.org/10.2307/749723

Kang, H., & van Es, E. A. (2019). Articulating Design Principles for Productive Use of Video in Preservice Education. *Journal of Teacher Education*, *70*(3), 237–250. https://doi.org/10.1177/0022487118778549

Kang, J. (2020). Interrelationship Between Inquiry-Based Learning and Instructional Quality in Predicting Science Literacy. *Research in Science Education*, *52*(1), 339–355. https://doi.org/10.1007/s11165-020-09946-6

Klieme, E., Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht [Instructional quality from a student perspective: Culture-specific profiles, regional differences, and associations with effects of instruction]. In J. Baumert (Ed.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 333–359). Wiesbaden: VS. https://doi.org/10.1007/978-3-322-97590-4_12.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I. 'Aufgabenkultur' und Unterrichtsgestaltung. [Teaching mathematics in lower secondary schools. 'Task culture' and lesson design.]. In E. Klieme, & J. Baumert (Eds.), *TIMSS-Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte* (pp. 43–57). Bundesministeriumfür Bildung und Forschung (BMBF).

Korneck, F., Krüger, M., & Szogs, M. (2017). Professionswissen, Lehrerüberzeugungen und Unterrichtsqualität angehender Physiklehrkräfte unterschiedlicher Schulformen. [Professional

knowledge, teacher beliefs, and teaching quality of prospective physics teachers from different types of schools]. In E. Sumfleth, & H. Fischler (Eds.), *Professionelle Kompetenzen von Lehrkräften der Chemie und Physik. Studien zum Physik- und Chemielernen, Bd. 200* (pp. 1–21). Logos.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse [The model of instructional quality in COACTIV: A multicriteria analysis]. *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV*, 85–113.

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143–152. https://doi.org/10.1016/j.tate.2013.07.010

Labudde, P., Viiri, J., Fischer, H. E., & Neumann, K. (2014). Summary and Discussion. In H. E. Fischer, P. Labudde, K. Neumann, & J. Viiri (Eds.), *Quality of instruction in physics. Comparing Finland, Switzerland and Germany* (pp. 111–127). Waxmann.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Mayring, P. (2014). *Qualitative Content Analysis Theoretical Foundation, Basic Procedures and Software Solution.* https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173

Neuhaus, B. J. (2021). Unterrichtsqualität aus der Perspektive der Biologiedidaktik [Teaching quality from the perspective of biology education]. *Unterrichtswissenschaft*, *49*(2), 273–283. https://doi.org/10.1007/s42010-021-00114-x

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM - Mathematics Education*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A. K., Herrmann, C., Gerlach, E., Zülsdorf-Kersting, M., Heinitz, B., & Nehring, A. (2020b). Teaching quality in different subject matters in German-speaking countries—inbetween genericness and subject-specificity. *Unterrichtswissenschaft*, *48*(3), 409–446. https://doi.org/10.1007/s42010-020-00082-8

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM - Mathematics Education*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A. K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, *22*(6), 387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Praetorius, A.-K., Rogh, W., & Kleickmann, T. (2020a). Blind spots of the three basic dimensions model? Reconsidering the model based on an international synthesis on teaching quality. *Unterrichtswissenschaft*, *48*(3), 303–318. https://doi.org/10.1007/s42010-020-00072-w

Puttick, S., & Wynn, J. (2020). Constructing 'good teaching' through written lesson observation feedback. *Oxford Review of Education*, *47*(2), 152–169. https://doi.org/10.1080/03054985.2020.1846289

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

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499. https://doi.org/10.3102/0034654307310317

Sherin, M. G. (2001). Developing a professional vision of classroom events: Teaching elementary school mathematics. In T. Wood, B. Nelson, & S. Warfield (Eds.), *Beyond classical pedagogy: Teaching elementary school mathematics* (pp. 75–93). Erlbaum.

Strong, M., Gargani, J., & Hacifazlioğlu, Ö. (2011). Do We Know a Successful Teacher When We See One? Experiments in the Identification of Effective Teachers. *Journal of Teacher Education*, *62*(4), 367–382. https://doi.org/10.1177/0022487110390221

Tatto, M. T. (2021). Teacher education in the United States of America: An overview of the policies, pathways, issues and relevant research. In Mayer D. (Ed.), *Teacher education policy and research* (pp. 177–194). Springer. https://doi.org/10.1007/978-981-16-3775-9_13

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction*, *46*, 45–60. https://doi.org/10.1016/j.learninstruc.2016.08.003

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300. https://doi.org/10.1037/a0017908

Wind, S. A., Jones, E., & Bergin, C. (2021). Principals' severity affects teacher evaluation: Statistical adjustments mitigate effects. *School Effectiveness and School Improvement*, *32*(3), 413–429. https://doi.org/10.1080/09243453.2021.1892773

## *11.4. Beitrag 4: Perception and Interpretation of Cognitive Activation in Science Pre-Service Teacher Education – The Necessity of Establishing a Common Approach*

Heinitz & Nehring (submitted). Perception and Interpretation of Cognitive Activation in Science Pre-Service Teacher Education – The Necessity of Establishing a Common Approach

Abstract:
Cognitive activation is a central construct for conceptualizing instructional quality. Science pre-service teachers (SPSTs) should be trained to identify cognitive activation and maximize student learning gains. Our study aims to get insights into the training of professional vision in SPST education seminars with high ecological validity. We compare four seminars in their training of professional vision focusing on cognitive activation using video vignettes in a pre-post-assessment (nSPST = 21). The seminars showed a varying degree of success in their training. Our findings have implications for SPST education hinting towards criteria for a successful implementation of cognitive activation.

CRediT Author Statement zur Eigenleistung:
Benjamin Heinitz: Conceptualization**,** Writing — original draft, review & editing, Investigation, Data Curation, Formal analysis, Visualization

Title Page (with Author Details)

# Perception and Interpretation of Cognitive Activation in Science Pre-Service Teacher Education – The Necessity of Establishing a Common Approach

Benjamin Heinitz[a] and Andreas Nehring[b]

[a]*Institute for Psychology in Education, University of Münster, Germany*

[b]*Institute for Science Education, Leibniz University Hannover, Germany*

ORCID (Benjamin Heinitz): https://orcid.org/0000-0001-8626-6587
ORCID (Andreas Nehring): https://orcid.org/0000-0002-8723-5552

Corresponding Author: Benjamin Heinitz, b.heinitz@uni-muenster.de Institute for Psychology in Education, University of Münster, Fliednerstr. 21, 48149, Münster, Germany

Prof. Dr. Andreas Nehring, nehring@idn.uni-hannover.de Institute for Science Education, Leibniz University Hannover, Am Kleinen Felde 30, 30167, Hannover, Germany

# Perception and Interpretation of Cognitive Activation in Science Pre-Service Teacher Education – The Necessity of Establishing a Common Approach


## Abstract

Cognitive activation is a central construct for conceptualizing instructional quality. Science pre-service teachers (SPSTs) should be trained to identify cognitive activation and maximize student learning gains. Our study aims to get insights into the training of professional vision in SPST education seminars with high ecological validity. We compare four seminars in their training of professional vision focusing on cognitive activation using video vignettes in a pre-post-assessment ($n_{SPST} = 21$). The seminars showed a varying degree of success in their training. Our findings have implications for SPST education hinting towards criteria for a successful implementation of cognitive activation.


Keywords: Instructional quality, professional vision, pre-service teacher education, science education

## Introduction

Cognitive activation is a central dimension of instructional quality and in many cases an important predictor for students' learning gains (Praetorius et al. 2018). Evaluating cognitive activation poses a great challenge to science pre-service teachers (SPSTs), as it is often not directly accessible in a science classroom. Cognitive activation requires a more complex interpretation of classroom interactions and may only be observable to a limited degree, even over the course of multiple lessons and is thus harder to rate reliably (Praetorius et al. 2012, 2014). Therefore, it is unsurprising that especially SPSTs focus less often on criteria of cognitive activation compared to experienced teachers when assessing the instructional quality of a lesson (Heinitz & Nehring, 2023).

At the same time, video recordings of classroom interactions present a possible solution for a specific training of professional vision and connecting theoretical knowledge to classroom interactions (Sherin & van Es, 2009). Video vignettes have already proven to be an important tool and are widely used in teacher education (Blomberg et al. 2013, Junker et al. 2022). They have also been used for the targeted training of professional vision of generic, as well as domain-specific criteria of instructional quality (e.g., Stürmer et al. 2013; Hellermann et al. 2015; Sunder et al. 2016; Kramer et al. 2017). The use of video vignettes to specifically implement cognitive activation in SPST education seminars, on the other hand, is rarely the subject of research.

In Germany, where this study was conducted, SPSTs from different stages of education and different previous experiences are trained together by domain specific advisors in SPST education seminars. These seminars are individually designed by the advisors and they usually do not have a common basis of criteria for cognitive activation or instructional quality in general. Cognitive activation is described in different terms and with different criteria in many studies (Christ et al. 2022). Thus, SPSTs might bring divergent theoretical knowledge which influences their training of professional vision. The advisors have to anticipate these differences and adapt their seminars accordingly. Advisors on the other hand might also have an individual understanding of criteria of cognitive activation, which might impact their seminars. The challenges that are connected with these differences are rarely systematically described for SPST education.

To add to this research gap and gain deeper insights into the learning and communication processes about cognitive activation in SPST education seminars, we conducted an exploratory field study. The study compares the changes in professional vision of 21 SPSTs coming from four different seminars before and after attending a teacher education seminar based on a video vignette and focused on four specific criteria of cognitive activation.

**Theoretical Background**

**Conceptualizing cognitive activation according to the Science Education Perspectives framework**

Many frameworks of instructional quality include cognitive activation. Depending on the study it can be directly connected to student learning gains (Praetorius et al., 2018), but differing operationalizations of cognitive activation led to inconsistent results regarding learning gains and difficulties when comparing studies focusing on cognitive activation (Christ et al. 2022). In many cases cognitive activation is described as a subject-specific dimension of instructional quality but it can be operationalized from generic point of view (Praetorius & Charalambous, 2018). The science-specific point of view on cognitive activation for this study is based on a review of criteria used in science education video studies (Heinitz & Nehring, 2020). This review presented the basis for the **S**cience **E**ducation **P**erspectives (SEP) framework (Heinitz & Nehring, 2023). Cognitive activation is one of seven dimensions in the framework and defined by 13 criteria (tab. 1).

| Dimension II Cognitive activation | Description |
|---|---|
| Selecting ambitious learning opportunities | Selection of learning opportunities in relation to cognitive activation. For example, the extent to which a task over- or underwhelming for the students. Only the selection and not the implementation is evaluated. |
| Problem-based learning | Including (independent) problem solving tasks in the lesson. |
| Multiple approaches and solutions | Integrating different forms of representation and providing or accepting different solutions. |
| Use of complex and interconnected tasks | Integrating links or interconnections between tasks. The tasks can be linked within a lesson or in a larger context over multiple lessons or subjects. |
| Cognitive activating implementation of scientific methods | Integrating scientific ways of thinking and working with the goal of cognitive activation. Specific scientific ways of thinking and working can be integrated into the lesson with a varying degree of cognitive activation. For example, planning an experiment is more cognitively demanding than providing instructions for the experiment. |
| Activating prior knowledge | Activating and integrating students' prior knowledge to actively use it during the lesson. This also includes the exploration of existing prior knowledge. It is important that the students are actively encouraged to contribute their prior knowledge and that it is not dictated by the teacher. |
| Cognitively constructive error culture | Constructive handling of mistakes made by the students in a goal-oriented manner. Errors should be used as learning opportunities with the active involvement of students. |

| | |
|---|---|
| Exploring thought processes | Exploring students' thought processes by encouraging them to transparently describe their ideas and reasoning. This can be done during the process (think aloud) or afterwards when explaining solutions. |
| Integrating students' ideas constructively into the lessons | Integrating students' ideas and conceptions into the ongoing teaching process. These ideas can be presented as mental construct regarding specific content or alternative solutions or hypotheses about experiments. The do not have to be scientifically adequate to be integrated productively. |
| Supporting cognitive activating processes | Supporting cognitive activating processes of the students. The students should be confronted with tasks that support active thinking. This can be done by scaffolding certain steps, providing supportive impulses or using open tasks. |
| Supporting and integrating metacognitive processes | Integrating reflective thought processes on a meta-perspective. Reflecting work and thought processes, as well as to the results. |
| Implementing scientific methods explicitly | Explaining the thought process behind scientific methods. This can be done by explicitly reflecting the use of a model or the design of an experiment. |
| Cooperative work for a goal-oriented activation of students | Integrating cooperative processes for students. The teacher should not take a central role in these phases and at most act as a mediator. |

*Table 1: The thirteen criteria of cognitive activation included in the SEP framework (Heinitz & Nehring, 2023).*

The SEP framework is linked to the MAIN-Teach model (Charalambous & Praetorius, 2020) and can be connected to frameworks of different subjects to compare interpretations of cognitive activation (Praetorius et al. 2020). This link presents an important aspect in an attempt for a collaborative approach that seems helpful to connect frameworks of instructional quality and create a more cohesive picture of cognitive activation and instructional quality in general (Charalambous et al. 2021; Charalambous & Praetorius, 2022; Christ et al. 2022).

**Training professional vision**

Cognitive activation presents an important dimension of instructional quality but is also connected to many difficulties in evaluations mainly connected to the limited accessibility during a lesson (Praetorius et al. 2012; 2014). SPSTs are confronted with these difficulties and generally struggle with connecting their theoretical knowledge to classroom interactions (Korthagen & Kessels, 1999). Additionally, SPSTs are early on often not yet able to focus on relevant situations (Star & Strickland, 2008). The connection between theoretical knowledge and classroom interactions is part of a professional vision and can be trained using video vignettes (Sherin, 2001; Sherin, 2007; Sherin & van Es, 2009; Kulgemeyer et al. 2021).

Professional vision has subject-specific aspects (Blomberg et al. 2011; Steffensky et al. 2015) that might be especially relevant, when evaluating cognitive activation which is often described as a subject-specific dimension of instructional quality (Praetorius & Charalambous, 2018; Heinitz & Nehring, 2020). The subject-specific but also generic theoretical knowledge of SPSTs is strongly related to the training of professional vision which highlights the importance of SPSTs' prerequisites (Stürmer et al. 2014). SPSTs' prerequisites might differ depending on their university courses but also on their own experiences as students (Lortie, 1975). Training of professional vision should thus be combined with a pretraining on relevant aspects for a stronger effect (Martin et al. 2023). A common basis to build upon would benefit a shared vision of teaching (Kang & van Es, 2019). At the same time, teachers' professional vision is an important mediator between their knowledge and their students' learning progression (Blömeke et al. 2022) and should thus be considered in SPST education.

**Video vignettes in teacher education**

Video vignettes are an important tool for the training of professional vision and are widely used in teacher education (Blomberg et al. 2013). This includes generic as well as subject-specific applications (e.g., Stürmer et al. 2013; Hellermann et al. 2015; Sunder et al. 2016; Kramer et al. 2017). In Germany, where this study was conducted, video vignettes find a broad application in teacher education often through different web-based offers (Junker et al. 2022). Many SPSTs are therefore used to working with video vignettes in their university courses but less in the practical phases of teacher education, which is the focus of this study.

**Professional vision along seven dimensions of instructional quality**

Professional vision can be described as a situation specific skill that is based on the teacher's knowledge (Blömeke et al. 2015). Teachers have to perceive a situation as relevant, interpret it and base their following decision on their current judgement. Studies show an impact of a teacher's professional vision on their instructional quality with both aspects having an influence on students learning outcomes (Blömeke et al. 2022). Therefore, a direct link between

professional vision and instructional quality might offer an easy and transparent access for a combined analysis. The **S**ystematic and **T**ransferable **A**pproach for **R**atings of Instructional Quality (STAR) Model was developed as a transparent approach for SPST education (Heinitz & Nehring, 2024) and is used for this study. It presents an adaptation of the PID model (Blömeke et al. 2015) and connects the knowledge of an evaluator to specific criteria of instructional quality. This is used as a basis to analyze professional vision using the steps of perception, interpretation and decision making (figure 2). A situation within a lesson can be perceived as relevant but evaluated using different criteria of instructional quality. Depending on the criterion, evaluations might require subject-specific knowledge. Each situation can be evaluated with different criteria of instructional quality, depending on the knowledge and individual perception. The STAR model is connected to the SEP framework and allows an analysis using a broad basis of instructional quality. Evaluations of instructional quality become easier accessible by connecting criteria to observable behavior within a lesson, using the SEP coding manual. Each perception might offer multiple interpretations coming from different evaluators that can be contrasted and explained along the seven dimensions represented as beams of the STAR model. This simple visualization illustrates the different approaches taken to interpret the same perception, which might be based in different theoretical knowledge. The model represents the reaction to a situation going from the inside to the outside of the circle.

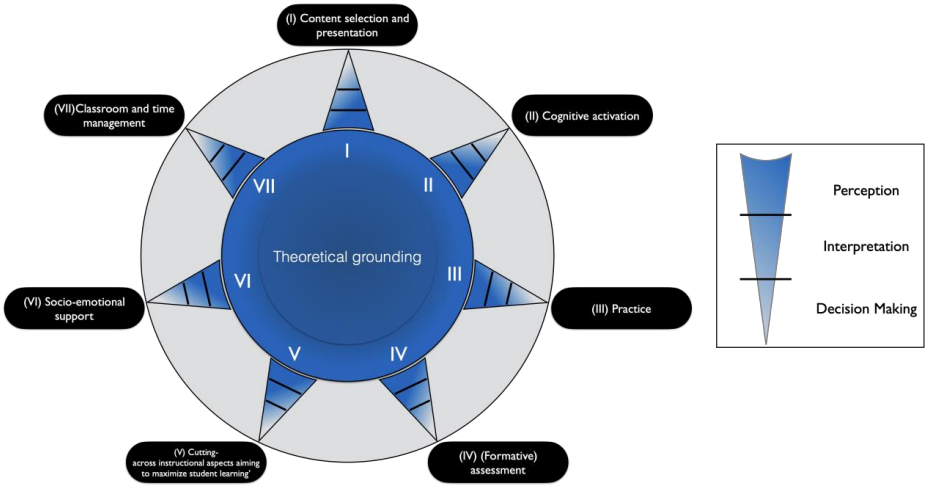



*Figure 1: The **S**ystematic and **T**ransferable **A**pproach for **R**atings of Instructional Quality (STAR) model (adapted, Heinitz & Nehring, 2024)*

**Communication in SPST education**

A successful training of professional vision requires successful communication and thus a common ground between all participants (Clark & Schaefer, 1989). Implicit theories of instructional quality might not only lead to differences in evaluations (Praetorius et al. 2012; Taut & Rakoczy, 2016) but also impact communication about instructional quality. SPSTs and their advisors need a shared understanding of instructional quality to create a shared vision of teaching but this might be impaired by encapsulation of knowledge similar to medical experts (Schmidt & Rikers, 2007). Technical terms can be connected to different knowledge depending on the perspective and experience of the individual. Especially SPST education seminars might be impacted by different understandings due to the variety of backgrounds of the participants.

**Research question**

The impact of cognitive activation for student learning gains combined with the difficulties reported for evaluating criteria of cognitive activation opens the questions for its implementation in SPST education. The broad application of video vignettes for the training of professional vision presents a possible approach for SPST education seminars, but individual differences between the seminars might have a strong impact on the training. Our research

questions are therefore based on an exploratory implementation of cognitive activation using video vignettes:

1. To what extent are perceptions and interpretations of cognitive activation in video vignettes comparable in SPST education seminars?
2. To what extent is the change in perception and interpretation regarding cognitive activation comparable within and between seminars?
3. To what extent is the development of pre-service teachers' professional vision specific to their seminar?

**Study design**

Our study is designed as a semi-structured exploratory field study aiming at ecological validity by creating a situation that is close to the usual work environment of SPSTs. The assessment is based on open questions about cognitive activation without pre-set criteria. Closed questions might improve the comparability of results but might also conceal some of the variability usually present within SPST education seminars. Both open and closed approaches would capture different constructs (Müller & Gold, 2022), but a broad mapping of variance is beneficial for our exploratory approach.

Before the seminar session, the advisors received instructions to use four specified criteria from the SEP framework to create a comparable context in a multiple case design (Yin, 2018). Due to vague or missing guidelines for SPST education in Germany (Kunz & Uhl, 2021), the seminars have an individual approach and our study is designed to capture a realistic picture while still allowing comparisons between the different seminars.

**SPST education seminars**

SPST education in Germany is divided into two phases. The first phase starts with a university education which is completed with a Master of Education degree. In the second phase SPSTs gain practical experience in schools. They work as in-service teachers but with a reduced number of lessons. Additionally, they visit regular teacher education seminars, where they

receive feedback and instructions by domain-specific advisors. These seminars are usually held every two weeks and deal with a specific topic each session. The topics might differ between seminars and are often aligned with requests of SPSTs. The seminars might vary in size and are visited by around four to twelve SPSTs. Each semester new SPSTs are admitted to the second phase and others finish it. Thus, the experience differs between SPSTs within a seminar.

**Video vignettes used in the study**

Two video vignettes were used in the pre- and post-assessment. The first video vignette showed an excerpt of a discussion following an experiment conducted by students (5:53 minutes). During the lesson the students received several tasks guiding them along an experiment including the heating of copper and sulfur in a test tube. Afterwards all observations were collected and discussed. The students were supposed to explain their observations using the Dalton atomic model and their prior knowledge but struggled with this task.

Our second video vignette showed an excerpt from the beginning of a lesson focusing on the use of models as tools of scientific inquiry (3:14 minutes). Using the simple particle model, the lesson started with a picture representing different states of matter on the particle level. The teacher used it as a silent impulse and the students used their prior knowledge describing the schema. Afterwards, a central question regarding the use of models as a scientific method was derived and presented as the main goal of the lesson.

Although the vignettes differed in their length and theme, the content referred to similar criteria of cognitive activation. We used the SEP coding manual to find comparable vignettes and confirm their similarity. The final subdivisions into relevant situations (RS) for both vignettes were carried out inductively, based on the data of SPSTs and their advisors. The learning vignette was split into twelve and the transfer vignette into eight relevant situations (tab. 2).

| | **Relevant situations for the learning vignette** |
|---|---|
| RS1 | Students interpreting their observations |
| RS2 | Student presenting her interpretation including an inadequate idea about the reaction |
| RS3 | Teacher reshapes the answer to include an adequate idea, ignoring the students mistake |

| | |
|---|---|
| RS4 | Teacher presents his interpretation of the students' observations |
| RS5 | Teacher asks a follow-up question |
| RS6 | Student tries to answer the question but is missing the knowledge necessary to fully answer it |
| RS7 | Teacher reshapes the answer, ignoring the students missing knowledge |
| RS8 | Student presents an adequate summary of the interpretations |
| RS9 | Teacher brings up a model experiment from the previous chemistry lesson |
| RS10 | Teacher repeats all answers by a student |
| RS11 | Teacher tries to guide the students to the connections between both lessons |
| RS12 | Student presents an alternative interpretation of the experiment including an inadequate hybrid model |
| **Relevant situations for the transfer vignette** | |
| RS1 | Teacher starts the lesson with a silent impulse including a scheme of the three states of matter using a particle model |
| RS2 | Students explain their interpretations of the scheme |
| RS3 | Teacher challenges the ideas of the students |
| RS4 | Student explains that the scheme is a representation of water |
| RS5 | Teacher again challenges the ideas of the students |
| RS6 | Student explains that particles cannot be seen with the eye and another adds that it is just a form of visualization |
| RS7 | Teacher presents the question of the lesson focusing on the scientific ways of thinking and working |
| RS8 | Teacher gives students their first task |

*Table 2: Inductive subdivision of both vignettes into relevant situations (RS) based on timestamps and descriptions by participants.*

Both vignettes were used in the pre- and post-assessment to explore the individual perception and interpretation of cognitive activation. The learning vignette was later used by the advisors to introduce four criteria of cognitive activation during their seminars. The transfer vignette was only used for the assessments and unknown to the advisors.

**Methods**

**Participants**

In total 26 SPSTs participated in our study with 21 SPSTs (52 % female) being present for both pre- and post-assessment and thus considered as full data sets in the analysis. The four domain-specific advisors (50 % female) were tested separately during the post-assessment, since they were the ones designing the seminar and were not supposed to know the transfer vignette. The participants were selected as an opportunity sample through the voluntary involvement of the domain-specific advisors, but could object to the participation. Each domain-specific advisor has their own seminar with six to ten SPSTs. All participants were situated in the federal state

of Lower-Saxony in Germany. The SPSTs have been working in schools between one and seventeen months (average = seven month). 90 % have a Master of Education and 10 % changed profession and entered the second stage without a specific educational degree. All participants have chemistry as one of their subjects with differing secondary subjects. The most popular being biology (48 %) followed by mathematics (14 %).

**Data collection**

Data collection was split into a pre- and post-assessment and conducted in two consecutive seminar sessions. The pre-assessment started with a short introduction of our study, followed by the first written assignment. All participants were tasked to describe their understanding of cognitive activation and then evaluate the cognitive activation in two video vignettes. The evaluation was pre-structured and asked the participants to provide a timestamp, a short description of the situation and its impact on cognitive activation. The order of the two vignettes was changed for half of the participants to take into account any form of fatigue or learning between the vignettes. The tasks had a soft limit of 60 minutes, which was used by almost all participants. After finishing their written task, the participants were able to take a short break before the session on cognitive activation started. The advisors did not participate in the pre-assessment, since they were only supposed to know one of the video vignettes. They started their session, as soon as the pre-assessment was finished. They were free to structure the session in a way they deemed to be appropriate, since we wanted to keep the usual flow of the seminar. The only instructions they had to follow was an inclusion of the learning vignette and the four criteria: “Integrating students’ ideas constructively into the lessons”, “Cognitive activating implementation of scientific methods”, “Activating prior knowledge”, “Cooperative work for a goal-oriented activation of students”. Definitions for all four criteria were provided a week prior to the seminar session. The session was audio recorded, transcribed and anonymized.

The transcript was used to rank the seminars in their implementation of the four criteria. The post-assessment started with the same written task and evaluation of cognitive activation,

followed by a short break after 60 minutes. The second half of the post-assessment included a written stimulated recall of the previous seminar session based on the transcript. All participants were asked to describe their understanding of six terms that were used during the seminar but were not clearly defined. These terms were selected by using the SEP framework. Whenever a term was used during the discussion that did resemble a technical term but deviated from the original term it was considered for the stimulated recall (e.g., “latent thought”). These pseudo technical terms were not clearly codable with the framework and were used in the stimulated recall to check for a common understanding within the seminar. We chose six terms for each seminar that were the most prominent during the discussion.

**Comparing the implementation of criteria**

Based on our analysis, the seminars received a score for their implementation of the four criteria. This score is split into four categories that are aligned with the instructions the advisors received before their seminars. These categories include a mentioning of the criteria by the advisors during the seminar, an introduction that matched the definition the advisors received, a connection to practical examples and the relative frequency of use during the seminar (compared to the other seminars). The categories were checked for each of the four specific criteria of cognitive activation individually allowing a total of four points per category. Thus, a total of 16 points were achievable and the seminars ranked based on this score ranging from “A” for the highest to “D” for the lowest score. However, this score is solely based on our instructions and does not represent the quality of the seminars in general.

**Data analysis**

All assessments of cognitive activation were connected to timestamps and descriptions of the situation and were clustered into relevant situations inductively (tab. 2). Each relevant situation contains multiple perceptions from different participants. Using the STAR model and the connected SEP framework coding manual each perception was connected to at least one criterion of instructional quality. This coding was based on the interpretation provided by the

participants for each situation. Figure 2 represents an example showing the coding process. A relevant situation is linked to an individual perception by a participant (e.g., “silent impulse at the beginning of the lessons” (ID:2_220512_6_8)). The interpretation provided by the participant (“A silent impulse is a [cognitively] activating method because the students […] have to activate their prior knowledge.” (ID:2_220512_6_8)) could be coded using the SEP framework (“Activating prior knowledge”) and is part of their evaluation of the video vignette.

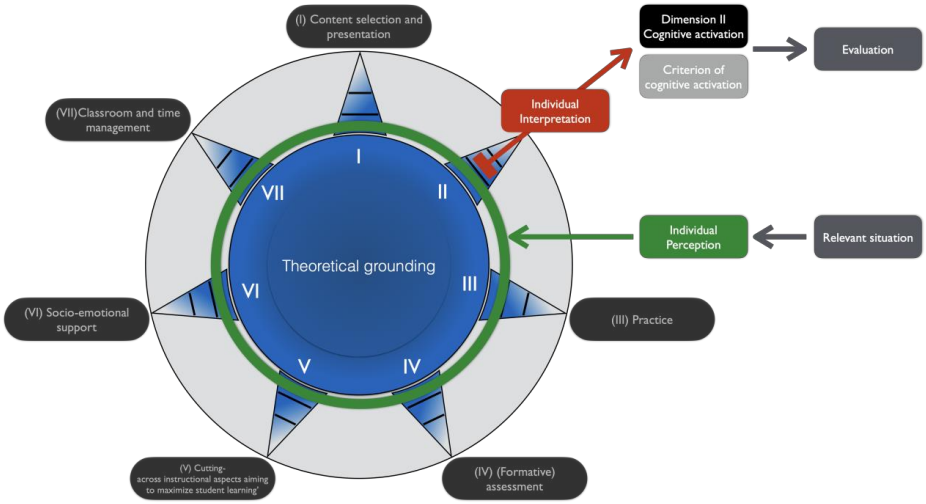



*Figure 2: Analyzing written evaluations of cognitive activation using the STAR-model.*

The use of the STAR model allowed a comparison of different perceptions for the same relevant situation. Additionally, it also provides insights into the use of criteria that do not refer to the same relevant situation and also allows a comparison of different vignettes.

Some descriptions are more complex and refer to multiple criteria. The participants chose relevant situations on their own during data collection and the length of perceptions could differ. Thus, each perception could be linked to multiple criteria of the SEP framework and could be coded multiple times. On average there were seven statements for the learning vignette by SPSTs during the pre-assessment and six statements for the transfer vignette. The post-assessment change to nine statements for the learning vignette and remained six statements for the transfer vignette. The advisors on average gave 19 statements for the learning vignette and

eight for the transfer vignette. For a comparison between the participants but also between pre- and post-assessment, we analyzed which segments were identified as relevant situations, which criteria were mentioned and the number of times a criterion was mentioned for both vignettes. All evaluations by SPSTs were also compared to the evaluation of their respective advisors. To test for interrater reliability, 15 % of the dataset was coded by a second rater. The raters reached a relative agreement of 65 %. Considering the high number of categories the SEP framework offers, we expected the agreement by chance to be relatively low. The Cohens Kappa coefficient confirms this ($\kappa = 0.61$) and is in a range of substantial agreement (Landis & Koch, 1977). The raters only used 30 of 50 possible categories for their rating. Since unused categories are not included in the calculation of Cohens Kappa, the agreement by chance is even lower and we consider the agreement to be acceptable. To further test the reliability, 35% of the dataset was rated by two raters individually and then checked in a consensus rating. 90 % of all disagreements could be resolved with the rating of one of the two raters and the distribution was almost identical between the two. This result further supports the substantial agreement in our interrater reliability test.

**Results**

Our results are presented in five sections starting with a ranking of the seminars based on their implementation of the four criteria. This is followed by SPSTs' perception and their interpretation during the pre- and post-assessment and the advisors' perception and interpretation, as well as the stimulated recall.

**Implementation of criteria**

The seminars show differences in their implementation of the four criteria based on our instructions (Tab. 3). Two seminars achieved an (almost) perfect score showing a preferable implementation of the four criteria. The other two seminars only got a mediocre score and did not include all four criteria in their seminar session.

| Categories for the ranking | Seminar A | Seminar B | Seminar C | Seminar D |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Criterion was mentioned during the seminar | 4 | 4 | 3 | 2 |
| Definition matched the instructions | 4 | 4 | 2 | 1 |
| Connection to examples | 4 | 3 | 2 | 2 |
| Relative frequency | 4 | 4 | 2 | 2 |
| Total score | 16 | 15 | 9 | 7 |

*Table 3: Scores for the ranking of the four seminars. The individual scores for each category range from 0-4.*

**Perception**

**Pre-Assessment**

The learning vignette contains a total of twelve relevant situations and three are perceived as relevant by the majority (>70 %) of all participants in the pre-assessment (tab. 4). The seminars show their strongest overlap in perception for RS 1 (students interpret their observations), RS 4 (teacher presents his interpretation of the students' observations) and RS9 (teacher reminds the students of a previous model experiment). There are also some situations that are only perceived by a few participants and do not show any form of systematic comparability within or between the seminars.

| Relevant situation | Seminar A | | Seminar B | | Seminar C | | Seminar D | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Pre | Post | Pre | Post | Pre | Post | Pre | Post | Pre | Post |
| Learning vignette | | | | | | | | | | |
| 1 | 100% | 100% | 75% | 100% | 71% | 86% | 100% | 75% | 86% | 90% |
| 5 | 67% | 50% | 50% | 100% | 29% | 86% | 0% | 100% | 38% | 81% |
| 4 | 67% | 100% | 50% | 25% | 86% | 86% | 75% | 75% | 71% | 76% |
| 9 | 83% | 100% | 50% | 50% | 86% | 71% | 75% | 50% | 76% | 71% |
| 11 | 67% | 100% | 75% | 25% | 57% | 86% | 25% | 50% | 57% | 71% |
| Average perception | 53% | 63% | 35% | 44% | 41% | 58% | 40% | 48% | 43% | 55% |
| Transfer vignette | | | | | | | | | | |
| 1 | 100% | 100% | 100% | 100% | 86% | 86% | 100% | 100% | 95% | 95% |
| 3 | 83% | 100% | 75% | 75% | 86% | 71% | 100% | 100% | 86% | 86% |
| 7 | 67% | 83% | 50% | 100% | 71% | 71% | 75% | 100% | 67% | 86% |
| 5 | 83% | 67% | 50% | 50% | 71% | 71% | 75% | 100% | 71% | 71% |
| 8 | 67% | 67% | 75% | 75% | 57% | 86% | 25% | 50% | 57% | 71% |
| Average perception | 60% | 65% | 53% | 50% | 48% | 50% | 53% | 59% | 54% | 56% |

*Table 4: Relative frequency of perception by SPSTs for the most relevant situations. Every situation below 70% in the post-assessment is listed in appendix a. The RS are ranked by total frequency in the post-assessment.*

The participants show a higher overall perception in the transfer vignette. This vignette contains only eight relevant situations in total but the participants also show a majority perception for

three of these situations. All participants mainly refer to the impulses of the teacher, without further discussing the students. RS 1 (teacher starts the lesson with a silent impulse) is perceived by almost every participant making it the most frequently perceived situation for both vignettes. The other two situations are RS 3 (Teacher challenges the ideas of the students) and RS 5 (Teacher again challenges the ideas of the students). Seminar A has the highest overall perception in the pre-assessment for both vignettes.

**Post-assessment**

Overall, the total frequency of perception increases between the pre- and post-assessment for both vignettes. The learning vignette shows a particularly noticeable change for RS5 (teacher asks follow-up question) and RS 11 (teacher connects the current and prior lesson), making them majority perceived situations in the post-assessment. This is also more in line with the advisors, who all show a reference to RS5 and a reference to RS11 in three cases.

Comparing the pre- and post-assessment for the transfer vignette shows only a little change overall. There is a slight increase in RS7 (teacher presents the guiding question) and RS8 (teacher presents the first task) making them both majority perceived situations, matching the perception of the advisors. However, the overall frequency of perceptions remains almost the same. Seminar A has the highest perception in the post-assessment for both vignettes. The highest increase between pre- and post-assessment can be observed in Seminar C for the learning vignette and Seminar D for the transfer vignette. Seminar B shows a slight decrease in the post-assessment for the transfer vignette.

**Interpretation**

**Pre-assessment**

The analysis of interpretations for the learning vignette shows that about a half of the criteria mentioned by the SPSTs are part of cognitive activation in the SEP framework (82 of 150). However, there are also criteria of dimensions I and V, that are frequently mentioned.

Two of the predetermined criteria were already frequently mentioned in the pre-assessment: “integrating students' ideas constructively into the lessons” and “activating prior knowledge”. The SPSTs also frequently used the criterion “supporting cognitive activating processes”. Although the task asked the participants to rate the cognitive activation, they frequently mentioned criteria of Dimension I mainly referring to the “selection and implementation of content” and dimension V referring to the “increase of student participation.

If the seminars are analyzed as individual cases, the same general tendencies for the selection of criteria are evident for almost all of them. Seminar B presents an exemption, because no criteria from dimension I are used and only one interpretation refers to dimension V. This makes it the only seminar which almost exclusively uses criteria of cognitive activation. Seminar C, on the other hand, stands out because dimensions I, II and V are used with almost equal frequency. However, the choice of individual criteria shows the same pattern as the other seminars.

| | Seminar A | | Seminar B | | Seminar C | | Seminar D | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Pre | Post | Pre | Post | Pre | Post | Pre | Post | Pre | Post |
| Learning vignette | | | | | | | | | | |
| I. Content selection and presentation | 8 | 12 | - | 7 | 18 | 11 | 8 | 8 | 35 | 38 |
| II. Cognitive activation | 32 | 39 | 19 | 27 | 18 | 20 | 13 | 20 | 82 | 106 |
| Integrating students' ideas constructively into the lessons | 7 | 8 | 5 | 12 | 5 | 6 | 4 | 14 | 21 | 40 |
| Activating prior knowledge | 4 | 10 | - | 4 | 6 | 3 | 4 | 2 | 14 | 19 |
| Cooperative work for a goal-oriented activation of students | 2 | 2 | 3 | 2 | - | 3 | 2 | - | 7 | 7 |
| Cognitive activating implementation of scientific methods | 2 | 2 | - | 1 | - | - | - | - | 2 | 3 |
| Total of the four predetermined criteria | 15 | 22 | 8 | 19 | 11 | 12 | 10 | 16 | 44 | 69 |
| V. Cutting-across instructional aspects aiming to maximize student-learning | 10 | 10 | 1 | - | 14 | 17 | 6 | 1 | 31 | 28 |
| Total | 50 | 69 | 20 | 34 | 52 | 52 | 28 | 30 | 150 | 185 |
| Transfer vignette | | | | | | | | | | |
| I. Content selection and presentation | 6 | 6 | 3 | 4 | 14 | 8 | 7 | 10 | 30 | 28 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| II. Cognitive activation | 23 | 34 | 19 | 17 | 18 | 17 | 7 | 14 | 67 | 82 |
| Activating prior knowledge | 10 | 12 | 4 | 7 | 4 | 4 | 1 | 2 | 19 | 25 |
| Integrating students' ideas constructively into the lessons | 6 | 7 | 5 | 4 | 6 | 6 | 3 | 5 | 20 | 22 |
| Cooperative work for a goal-oriented activation of students | - | 2 | - | - | 1 | 1 | - | 1 | 1 | 4 |
| Cognitive activating implementation of scientific methods | - | 1 | 1 | - | - | - | - | - | 1 | 1 |
| Total of the four predetermined criteria | 16 | 22 | 10 | 11 | 11 | 11 | 4 | 8 | 41 | 52 |
| V. Cutting-across instructional aspects aiming to maximize student-learning | 5 | 1 | - | - | 7 | 5 | 3 | 2 | 15 | 8 |
| Total | 37 | 43 | 22 | 21 | 40 | 30 | 17 | 26 | 116 | 120 |

*Table 5: Absolute number of different interpretations during pre-and post-assessment. Note: The table is reduced to the three mainly used dimensions and the four predetermined criteria of cognitive activation. Full table in appendix B.*

The transfer vignette shows a similar pattern in the overall use of criteria. Dimension II has the most frequent use (67 of 116) and is followed by dimensions I and V. On the level of individual criteria, the transfer vignette shows a majority use for the same criteria as the learning vignette.

An analysis of the individual seminars reveals that seminar B again stands out with no use of dimensions other than cognitive activation, except for three uses of dimension I. Seminar C shows an equally frequent use of dimension I and II with a lesser use of dimension V. Seminar D also shows an equally frequent use of dimensions I and II but has an overall smaller number of interpretations that are used for the evaluation, compared to the learning vignette.

**Post-Assessment**

Regarding the absolute numbers of interpretations (tab. 5) the seminars show an increase in the post-assessment of the learning vignette. This increase is mainly attributed to the criterion “Integrating students’ ideas constructively into the lessons”. The changes of dimension I in seminar B show an unusual pattern in the data. This dimension was not supposed to be evaluated

in the vignette and the participants did not show any references in the pre-assessment but started using it in the post-assessment.

The total number of interpretations stays almost the same in the transfer vignette, but the number of references to cognitive activation shows an increase. In this case no criteria stand out to mainly cause the changes.

On an individual level, the SPSTs within a seminar show differences in their changes. On average 69 % of all participants show an increase in the absolute number of interpretations regarding the learning vignette (tab.6). Seminar A and B show generally higher rates. The average increase of interpretations connected to cognitive activation can be seen in 73 % of all cases. The higher increase of cognitive activation compared to the total increase is due to a shift towards cognitive activation with a decrease for other criteria. This shift is not equal for all criteria of cognitive activation and not limited to the predetermined criteria. The average change per criterion shows that the total increase in interpretations connected to cognitive activation is in many cases limited to only a few criteria. Seminar D even shows an average decrease for the four criteria.

| | Percentage of SPSTs with an overall increase in the absolute number of interpretations | Percentage of SPSTs with an increased perception of criteria of cognitive activation | Average change of cognitive activation per criterion | Average change of the four predetermined criteria | Average changes for all criteria |
|---|---|---|---|---|---|
| Learning vignette | | | | | |
| Seminar A | 83% | 83% | 9% | 17% | 9% |
| Seminar B | 100% | 75% | 19% | 38% | 23% |
| Seminar C | 43% | 57% | 0% | 4% | 1% |
| Seminar D | 50% | 75% | -7% | -19% | -6% |
| Average | 69% | 73% | 5% | 10% | 7% |
| Transfer vignette | | | | | |
| Seminar A | 67% | 67% | 24% | 25% | 10% |
| Seminar B | 0% | 0% | -7% | -6% | -2% |
| Seminar C | 14% | 43% | -4% | 0% | -2% |
| Seminar D | 75% | 75% | 19% | 25% | 13% |
| Average | 39% | 46% | 8% | 11% | 5% |

*Table 6: Relative frequency of changes in SPSTs' interpretations in the post-assessment*

When comparing the four seminars they show a different pattern for the transfer vignette. The overall percentage of participants with an increase in interpretations related to cognitive activation is lower than the learning vignette, but it is still present for all seminars, except for seminar B. The average change per criterion is comparable between both vignettes and again shows, that the individual SPSTs focus on specific criteria while neglecting others.

**Evaluations by domain-specific advisors**

**Perception**

The learning vignette contains two situations that are perceived as relevant by all advisors: RS5 (teacher asks follow-up question) and RS 9 (teacher reminds the student of last lessons model experiment). RS 9 presents the only situation that is perceived as relevant by all advisors and most of the SPSTs. Some situations are perceived as relevant by the advisors, but are hardly noticed by the SPSTs: RS2 (student presents an inadequate idea), RS5 (teacher asks follow-up question) and RS6 (student is missing knowledge to answer the question correctly). RS 10 (teacher repeats students answers) is the only situation that is not addressed by any of the advisors but mentioned by some SPSTs. Advisors B and D have the highest perception (75 % of all RS), followed by advisor C (66 %). Advisor A has the lowest perception (33 %).

The transfer vignette contains three situations that are perceived as relevant by all advisors and two situations that are not mentioned at all. The situations that are mentioned by all, refer to the actions of the teacher, while the situations that are not mentioned show student actions. This is similar to the perception of SPSTs. On an individual level, advisors C and D perceive the same situations as relevant for cognitive activation. Advisor B perceives the same number of situations but differs in one aspect. Advisor A has the lowest number of perceptions but also differs in only one aspect compared to advisor C and D.

**Interpretation**

The advisors use almost the same number of criteria of dimension I and dimension II to evaluate the learning vignette (tab. 7). They mainly use the same criteria as the SPSTs by referring to “integrating students' ideas constructively into the lessons”, “supporting cognitive activating processes” and “Activating prior knowledge”. The high use of dimension I is mainly determined by advisors B and D, who both show more references to dimension I than dimension II.

The four criteria that were supposed to be addressed during the seminar were not used by all advisors in their evaluations. “Integrating students' ideas constructively into the lessons” and “activating prior knowledge” were both mentioned by three advisors, while “cognitive activating implementation of scientific methods” and “cooperative work for a goal-oriented activation of students” were only mentioned by one advisor each. However, this differs from the actual implementation during the seminar.

| | **Advisor A** | **Advisor B** | **Advisor C** | **Advisor D** | total |
|---|---|---|---|---|---|
| Learning vignette | | | | | |
| I. Content selection and presentation | 2 | 19 | 3 | 11 | 35 |
| II. Cognitive activation | 6 | 12 | 9 | 7 | 34 |
| Integrating students' ideas constructively into the lessons | - | 5 | 6 | 3 | 14 |
| Activating prior knowledge | 2 | - | 1 | 1 | 4 |
| Cognitive activating implementation of scientific methods | - | 2 | - | - | 2 |
| Cooperative work for a goal-oriented activation of students | - | - | - | 1 | 1 |
| Total | 8 | 33 | 14 | 21 | 76 |
| Transfer vignette | | | | | |
| I. Content selection and presentation | 4 | 2 | 1 | 7 | 14 |
| II Cognitive activation | 7 | 3 | 4 | 2 | 16 |
| Integrating students' ideas constructively into the lessons | - | 1 | 1 | 2 | 4 |
| Activating prior knowledge | 1 | - | 2 | - | 3 |
| Cognitive activating implementation of scientific methods | - | - | - | - | - |
| Cooperative work for a goal-oriented activation of students | - | - | - | - | - |
| Total | 12 | 5 | 6 | 9 | 32 |

*Table 7: Criteria used by advisors to evaluate the vignettes. Note: The table is reduced to the two mainly used dimensions and the four predetermined criteria for cognitive activation. Full table in appendix b.*

It is particularly noticeable that fewer criteria are used to evaluate the transfer vignette. Dimension I and dimension II show a comparable number of uses and the emphasis is on the same criteria as seen in the learning vignette. However, none of the criteria that were supposed to be integrated in the seminar were used by all advisors and two of them were not mentioned in the evaluation of the transfer vignette at all.

**Stimulated recall**

The stimulated recall presents a deeper insight into the communication within an individual seminar. However, it is also possible to compare them on a generalized level by analyzing the criteria that are mentioned in the stimulated recall.

Seminars A and B each show a full agreement for only one of the six terms that were used in the stimulated recall but seminar B shows a high agreement overall. Seminar C has a full agreement for two terms and seminar D shows a full agreement for five terms. Overall, the SPSTs show the highest agreement for terms that refer to “integrating students' ideas constructively into the lessons” (high agreement of >70 % for 8 of 8), which was also the criterion mentioned the most during the stimulated recall (8 of 24).

| SeminarA | Seminar B | Seminar C | Seminar D |
|---|---|---|---|
| Supporting cognitive activating processes **67 %** | Cognitively constructive error culture **50 %** Integrating students' ideas constructively into the lessons **50 %** | Motivating embedding of content **29 %** Supporting cognitive activating processes **29 %** | Activating prior knowledge **50 %** Problem-based learning **50 %** |
| Supporting and integrating metacognitive processes **67 %** | Integrating students' ideas constructively into the lessons **75 %** | Integrating students' ideas constructively into the lessons **100 %** | Integrating students' ideas constructively into the lessons **100 %** |
| Integrating students' ideas constructively into the lessons **100 %** | Integrating students' ideas constructively into the lessons **100 %** | Structuring and focusing key aspects **43 %** | Connecting to previous or future lessons as well as to other subjects **100 %** |
| Connecting to previous or future lessons as | Activating prior knowledge **50 %** | Integrating students' ideas constructively into the lessons | Integrating students' ideas constructively into the lessons |

| | | | |
|---|---|---|---|
| well as to other subjects **83 %** | Selection and implementation of content **50 %** Integrating students' ideas constructively into the lessons **50 %** | **100 %** | **100 %** |
| Activating prior knowledge **50 %** Integrating students' ideas constructively into the lessons **50 %** | Selection and implementation of scientific methods **75 %** | Activating prior knowledge **29 %** Selection and implementation of content **29 %** Cognitive activating implementation of scientific methods **29 %** | Connecting to previous or future lessons as well as to other subjects **100 %** |
| Cognitive activation **50 %** | Integrating students' ideas constructively into the lessons **80 %** | Activating prior knowledge **71 %** | Structuring and focusing key aspects **100 %** |

*Figure 3: Criteria with the highest agreement in the stimulated recall, including percentage of participants using the criteria.*

**Discussion**

We present summarized answers to our research questions and discuss them against the background of current literature, to derive implications for SPST education.

**The SPST education seminars share a small common ground of cognitive activation**

The analysis of our first research question (To what extent are perceptions and interpretations of cognitive activation in video vignettes comparable in SPST education seminars?) showed a frequent reference to three situations for both vignettes with a common use of three criteria for cognitive activation: “Integrating students' ideas constructively into the lessons”, “activating prior knowledge” and “supporting cognitive activating processes”.

Even without any instructions, the SPSTs share a common ground in perceiving and interpreting cognitive activation. This might be based in their university education or experiences they made as students themselves, that influenced their view on classroom interactions (Blomberg et al. 2011). The three criteria seem to be accessible for many SPSTs in our vignettes. It would be plausible that these aspects are a specific focus in university education and SPSTs thus have especially high pedagogical-content knowledge (PCK) in this specific

area. Since PCK is connected to professional vision (Meschede et al. 2017), this would explain the already existing commonality in the pre-assessment. This explanation is supported by the localization of the seminars, that are all situated in the same federal state.

The SPSTs also regularly mentioned criteria that are connected to dimensions other than cognitive activation in the SEP framework. They might use these criteria as indicators of cognitive activation, even without explicitly stating this connection as a part of their implicit understanding of instructional quality (Praetorius et al. 2012, Taut and Rakoczy 2016). This seems plausible since we see a comparably strong use of two criteria outside of dimension II ("selection and implementation of content" & "increase of student participation").

**The seminars show a comparable change in professional vision which is mainly connected to criteria that were already present**

The analysis of our second research question (To what extent is the change in perceptions and interpretations regarding cognitive activation comparable within and between seminars?) showed larger changes in perceptions of the learning vignette. Two situations were added to the common ground and the overall perception of different situations increases. It seems obvious that SPSTs are able to point out more relevant situations after talking about the vignette. The transfer vignette also contains two additional situations in the common ground after the post-assessment, but shows a smaller change in perception overall. Some aspects might be transferable but the discussion of specific situations seems to have a stronger impact on the perception. This again shows that SPSTs have difficulties to perceive relevant situations, even when they should have the theoretical knowledge to assess them (Star & Strickland, 2008).

The interpretation shows limited changes for criteria of cognitive activation between pre- and post-assessment. This is connected to the ranking of the seminars. Although four specific criteria were supposed to be addressed during the seminar, not all advisors included them according the presented definitions. This provided unequal learning opportunities for the SPSTs and limited the possible changes in the post-assessment. Most of the SPSTs show an increase

in the use of criteria of cognitive activation in general but not for all four criteria. A comparison of changes between the criteria reveals, that this increase is often connected to a decrease in the use of other criteria. This shift in focus is supposed to happen, as long as the four criteria show an average increase. Seminars with a higher ranking (A and B) also show a higher average increase for the four criteria in the learning vignette and an overall higher absolute number of interpretations regarding the four criteria. The categories that were chosen for the scoring seem to be fitting indicators for the implementation of specific criteria in a SPST education seminar. The connection between score and changes in interpretation cannot be fully observed for the transfer vignette. Even though seminar A shows the expected change, seminar B and D contradict the previous observations. This might underline the importance of connecting theoretic concepts to observable behavior, which was possible for the learning vignette but not for the transfer vignette, leading to isolated knowledge about a concept without the opportunity to use it. However, the better performance of seminar D must be put into perspective, since the SPSTs of this seminar have the lowest interpretation rate and the seminar relatively small with only four participants in pre- and post-assessment. Thus, a small change in the post-assessment leads to higher relative values.

Some of the missing or divergent introductions of the four criteria might be indicators, that the advisors introduced the criteria from their point of view but did not match the definition of the SEP framework. The conceptualization and operationalization of instructional quality differs in many studies (Christ et al. 2022). This might also be the case for advisors and highlights the importance of a common reference for instructional quality.

The stimulated recall also showed difficulties in communication for many criteria. The main increase in use was seen for “Integrating students' ideas constructively into the lessons”, which also showed the highest agreement in the stimulated recall. This highlights the importance of successful communication for the training of professional vision.

**Advisors show an individual interpretation of cognitive activation that influences their seminar**

The analysis of our third research question (To what extent is the development of pre-service teachers' professional vision specific to their seminar?) shows a connection between the changes in professional vision and the ranking of the seminars but also individual influences of the advisors. The evaluations of the vignettes made by domain specific advisors differ from their implementation of our instructions. Although the advisors of seminar A and B did not include all four criteria in their evaluations, they implemented them in their seminars, leading to their high ranking. However, the high-ranking seminars also show changes that are not in line with the instructions. The advisors show an individual understanding of cognitive activation. They refer to similar situations in their evaluations but differences become more apparent when comparing interpretations. This is not surprising, considering the differences in evaluations of instructional quality seen in other research (Heinitz & Nehring, 2023). The commonalities between SPSTs and their advisors increase in the post-assessment. The advisors show an impact on perceptions by pointing out relevant situations but also on the use of criteria. Especially seminar B shows this impact because they did not use dimension I in the pre-assessment but start using it for the post-assessment, even though this was not part of the instructions. It seems plausible that differences between advisors might be transferred to their SPSTs and might even differ from standards set for SPST education.

**Implications for science pre-service teacher education**

The results of our study have implications regarding evaluations of cognitive activation, its implementation in SPST education and research.

**Evaluations of cognitive activation need a common basis and differences in understanding instructional quality should be explicitly addressed to improve SPST education**

Our study on cognitive activation highlights the necessity for a common framework to evaluate instructional quality (Heinitz & Nehring, 2023) and showed difficulties implementing specific criteria of cognitive activation in SPST education. The stimulated recall showed, that communication about cognitive activation was difficult during the seminar session. This can also be assumed for evaluations of instructional quality in SPST education. Advisors should be aware of these differences when talking about criteria of cognitive activation and provide specific definitions. The score for implementing the four predetermined criteria also contained the inclusion of definitions for the criteria which were not provided by all seminars. The changes in professional vision show a higher increase, when definitions were provided. This connection seems trivial but the use of subject specific terminology is not always based on the same theoretical background, which has to be acknowledged. Using a standardized framework in SPST education would help to improve transparency and continue the effort for a collaborative approach to instructional quality (Charalambous et al., 2021; Charalambous & Praetorius, 2022) as instructional quality is often assessed divergently in different studies (Christ et al. 2022). This would also offer common definitions for advisors and help them connect their criteria to the views of SPSTs to create a shared vision of teaching (Kang & van Es, 2019).

**Cognitive activation might be too complex or vague to be included in teacher education**

Our study shows general tendencies but also differences in the understanding of cognitive activation. This result is in accordance with research as cognitive activation is often described as difficult to assess and needs generally longer observation periods (Praetorius et al. 2014) or multiple raters for a reliable evaluation (Praetorius et al. 2012). It seems plausible to say that cognitive activation as a dimension widely used in research on instructional quality might be too difficult to incorporate into SPST education productively. However, cognitive activation in many cases described as an important prerequisite for high learning outcomes (Praetorius et al. 2018). Thus, not considering it as a part of instructional quality does not seem fit. It rather has

to be split into criteria with indicators of observable behavior and introduced stepwise in SPST education. In many cases this is a common practice but criteria are not based on the same theoretical background leading to differences in understanding, as seen in the stimulated recall. When breaking it down into criteria, educators and researchers would have to carefully pick relevant criteria that can broadly applied.

**A change in professional vision is favored by familiarity with a criterion**

Prior knowledge can be one of the most important factors explaining learning (Simonsmeier et al., 2021). Our study showed an increase in the use of criteria in the post-assessment. However, this increase is mainly limited to criteria that were already present in the pre-assessment. This limitation might be connected to missing theoretical background or missing learning opportunities during the seminar session. SPST education should acknowledge these limitations and provide sufficient support, when introducing criteria of instructional quality. Martin et al. (2023) highlighted the importance of prerequisites to support a learning with video vignettes. Criteria of instructional quality should be explicitly connected to observable behavior and also to other criteria, since SPSTs might have difficulties perceiving relevant situations that are connected to learning challenges for students (Heinonen et al., 2022). A clear connection to observable behavior might improve the accessibility for SPSTs and create comparable learning opportunities in SPSTs education. The ranking of seminars in our study also included the connection to examples for specific criteria and higher-ranking seminars showed a higher increase in professional vision.

All participants frequently refer to actions of the teachers in both vignettes and show fewer references to student actions. The students are mainly mentioned, whenever something is perceived to be problematic for cognitive activation and not when they are presented with positive examples. This might indicate two things: Either a missing theoretical background or encapsulation of knowledge. Certain actions by the teacher may have served as “good practice” examples and SPSTs might recognize them during a lesson but they cannot connect student

actions to criteria of cognitive activation. The advisors on the other hand might see an undisturbed flow of the lesson without a need to intervene or comment on the behavior (Berliner, 1989). The transfer vignette is generally regarded more positive and shows fewer remarks by the advisors, which supports this explanation. Even when asked they might not be able to fully explain their evaluation due to encapsulation of knowledge similar to medical experts (Schmidt & Rikers, 2007). Cognitive activation might be easier accessible, when it is not successful and students show obvious signs (e.g., boredom or confusion). A successful cognitive activation might be harder to pin to specific actions but it should be recognizable to enable SPSTs to learn from positive examples.

**A criteria-based training of instructional quality using observable indicators of cognitive activation might be helpful**

The post-assessment shows changes in perception and interpretation, but they do not always refer to cognitive activation as defined by the SEP framework. Even if the participants were describing indicators of cognitive activation from their point of view, they did not always match the definitions coming from research. These differences might be present in SPST education seminars and impact communication by lowering the common ground (Clark & Schaefer, 1989). This is also supported by our findings in the stimulated recall. A more transparent approach and an explicit connection between criteria of instructional quality and observable behavior might help SPSTs to transfer their knowledge.

The STAR model provided a systematic approach for analyzing different views on cognitive activation by focusing on teaching situations separately and linking observations with individual criteria. This does not exclude the possibility that several criteria may be relevant at the same time in a teaching situation, but looking at individual criteria facilitates access. The explicit linking of criteria for cognitive activation to specific teaching situations could provide a transparent approach for SPST education. Using different vignettes and connecting them with the same criteria leads to a continuous addition of indicators for that criterion. This might make

a transfer to new and unknown situations easier for SPSTs. A situation is then also connected to different criteria and correlations between criteria of instructional quality can be explicitly addressed.

**Studies about the evaluation of instructional quality could benefit from a more open approach to highlight differences**

Our study used an open approach to assess individual understandings of cognitive activation without providing a specific instrument for the participants. Instead, they had to describe their evaluations using their own terminology. By using the SEP framework and analyzing all statements with the STAR model, we were able to find commonalities between the participants but also many individual differences. This variance could have been masked in a more cohesive approach. Standardized instruments, involving rating items that make SPST react to given stimuli, might offer an easy approach to compare evaluations of instructional quality but might measure a different construct (Müller & Gold, 2022). Even though cognitive activation might need a more structured approach in teacher education, research on cognitive activation should always consider the differences in understanding, that were also found in our study.

**Limitations**

Although our findings are in accordance to other research, some limitations must be taken into account. First, there is no control group to compare the seminars in this observational study. SPSTs are constantly exposed to classroom observations during their education in the second phase. It would not be possible to find a SPST who was not exposed to any form of classroom observation or feedback during the timeframe of our study. The four seminars are considered separate cases in our study because their comparability is limited due to individual differences. Nevertheless, a control group might have provided deeper insights into the changes after the seminar session.

Second, our study was conducted in an open setting to better capture variance within the seminars. Thus, all analyses regarding cognitive activation are interpretations of statements

made by the participants. In order to substantiate our analyses, additional feedback by the participants regarding our interpretation of their written statements could have provided additional clarification. However, since comparable patterns could be found in all seminars, it can be assumed that the interpretation using the framework is reasonably close to what the participants wanted to express.

**Conclusion and Future Directions**

Cognitive activation is an important dimension of instructional quality and included in many studies. However, differing conceptualizations in research make it difficult to include in SPST education. The SPST education seminars had an impact on the professional vision of SPSTs, even with limited guidelines for their implementation into existing structures. This indicates the need for stronger guidelines but also shows the possibilities within the current system. Criteria of cognitive activation need to be connected to observable behavior to establish a common understanding and make it easier accessible. Cognitive activation might be too complex as a whole and should be implemented stepwise using a common ground.

Future research on the implementation of cognitive activation should provide indicators of observable behavior for all criteria. Criteria should be established detached from context to allow transferability and make them less dependent on specific situations.

**Ethics statement**

All participants were informed about the use of their data and were told they could terminate their participation at any point without consequences for their participation in the seminar. All data were collected anonymized. There were no further ethics requirements from our institution.

**References**


Berliner. (1989). New Directions for Teacher Assessment. In Pfleiderer J.(eds) Proceedings of the 1988 ETS Invitational Conference. Educational Testing Service. University of Michigan

Blomberg, G., Stürmer, K., & Seidel, T. (2011). How pre-service teachers observe teaching on video: Effects of viewers' teaching subjects and the subject of the video. *Teaching and Teacher Education*, 27(7), 1131–1140. https://doi.org/10.1016/j.tate.2011.04.008

Blomberg, G., Renkl, A., Sherin, G., Borko, H., & Seidel, T. (2013). Five research-based heuristics for using video in pre-service teacher education. _Journal for Educational Research, 90–114.

Blömeke, S., Gustafsson, J. E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. In _Zeitschrift fur Psychologie / Journal of Psychology_ (Vol. 223, Issue 1, pp. 3–13). Hogrefe Publishing. https://doi.org/10.1027/2151-2604/a000194

Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progress. *Learning and Instruction*, 79. https://doi.org/10.1016/j.learninstruc.2022.101600

Charalambous, C. Y., & Praetorius, A. K. (2022). Synthesizing collaborative reflections on classroom observation frameworks and reflecting on the necessity of synthesized frameworks. _Studies in Educational Evaluation_, _75_. https://doi.org/10.1016/j.stueduc.2022.101202

Charalambous, C. Y., Praetorius, A.-K., Sammons, P., Walkowiak, T., Jentsch, A., & Kyriakides, L. (2021). Working more collaboratively to better understand teaching and its quality: Challenges faced and possible solutions. _Studies in Educational Evaluation_, _71_, 101092. https://doi.org/10.1016/j.stueduc.2021.101092

Christ, A. A., Capon-Sieber, V., Grob, U., & Praetorius, A.-K. (2022). Learning processes and their mediating role between teaching quality and student achievement: A systematic review. _Studies in Educational Evaluation_, _75_, 101209. https://doi.org/10.1016/j.stueduc.2022.101209

Clark, H. H., & Schaefer, E. F. (1989). Contributing to Discourse. *Cognitive Science*, *13*(2), 259–294. https://doi.org/10.1207/s15516709cog1302_7

Heinitz, B., & Nehring, A. (2020). Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung. In Unterrichtswissenschaft (Vol. 48, Issue 3, pp. 319–360). Springer VS. https://doi.org/10.1007/s42010-020-00074-8

Heinitz, B., & Nehring, A. (2023). Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors. International Journal of Science Education. https://doi.org/10.1080/09500693.2023.2213382

Heinitz, B., & Nehring, A. (2024). Virtuelle Unterrichtshospitationen im Chemieunterricht - Eine Vernetzung der ersten und zweiten Phase der Lehrkräftebildung. *MNU Journal*, *3*, 182–190.

Heinonen, N., Katajavuori, N., Murtonen, M., & Södervik, I. (2022). Short pedagogical training in supporting university teachers' professional vision: A comparison of prospective and current faculty teachers. Instructional Science. https://doi.org/10.1007/s11251-022-09603-7

Hellermann, C., Gold, B., & Holodynski, M. (2015). Förderung von Klassenführungsfähigkeiten im Lehramtsstudium: Die Wirkung der Analyse eigener und fremder Unterrichtsvideos auf das strategische wissen und die professionelle Wahrnehmung. Zeitschrift für Entwicklungspsychologie Und Pädagogische Psychologie, 47(2), 97–109. https://doi.org/10.1026/0049-8637/a000129

Junker, R., Zucker, V., Oellers, M., Rauterberg, T., Konjer, S., Meschede, N., & Holodynski, M. (2022). *Lehren und Forschen mit Videos in der Lehrkräftebildung*. Waxmann.

Kang, H., & van Es, E. A. (2019). Articulating Design Principles for Productive Use of Video in Preservice Education. Journal of Teacher Education, 70(3), 237–250. https://doi.org/10.1177/0022487118778549

Korthagen, F. A. J., & Kessels, J. P. A. M. (1999). Linking Theory and Practice: Changing the Pedagogy of Teacher Education. In Educational Researcher (Vol. 28, Issue 4).

Kramer, C., König, J., Kaiser, G., Ligtvoet, R., & Blömeke, S. (2017). Der Einsatz von Unterrichtsvideos in der universitären Ausbildung: Zur Wirksamkeit video- und transkriptgestützter Seminare zur Klassenführung auf pädagogisches Wissen und situationsspezifische Fähigkeiten angehender Lehrkräfte. Zeitschrift Fur Erziehungswissenschaft, 20, 137–164. https://doi.org/10.1007/s11618-017-0732-8

Kulgemeyer, C., Kempin, M., Weißbach, A., Borowski, A., Buschhüter, D., Enkrott, P., Reinhold, P., Riese, J., Schecker, H., Schröder, J., & Vogelsang, C. (2021). Exploring the impact of pre-service science teachers' reflection skills on the development of professional knowledge during a field experience. International Journal of Science Education, 43(18), 3035–3057. https://doi.org/10.1080/09500693.2021.2006820

Kunz, H. & Uhl S. (2021). Allgemeine Ziele, Aufbau und Struktur des Vorbereitungsdienstes in den Bundesländern. In (Peitz J. & Harring M.) Das Referendariat. Ein systematischer Blick auf den schulpraktischen Vorbereitungsdienst (S. 15 – 27). Waxmann.

Lortie, D. (1975). Schoolteacher: A Sociological Study. London: University of Chicago Press.

Martin, M., Farrell, M., Seidel, T., Rieß, W., Könings, K. D., van Merriënboer, J. J. G., & Renkl, A. (2023). Knowing what matters: Short introductory texts support pre-service teachers' professional vision of tutoring interactions. Teaching and Teacher Education, 124. https://doi.org/10.1016/j.tate.2023.104014

Meschede, N., Fiebranz, A., Möller, K., & Steffensky, M. (2017). Teachers' professional vision, pedagogical content knowledge and beliefs: On its relation and differences between pre-service and in-service teachers. Teaching and Teacher Education, 66, 158–170. https://doi.org/10.1016/j.tate.2017.04.010

Müller, M.M., Gold, B. (2022). Videobasierte Erfassung wissensbasierten Verarbeitens als Teilprozess der professionellen Unterrichtswahrnehmung – Analyse eines geschlossenen und

offenen Verfahrens. *Z Erziehungswiss* **26**, 7–29 (2023). https://doi.org/10.1007/s11618-022-01128-6

Praetorius, A. K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? Learning and Instruction, 22(6), 387–400. https://doi.org/10.1016/j.learninstruc.2012.03.002

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. _Learning and Instruction_, _31_, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. _ZDM - Mathematics Education_, _50_(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. _ZDM - Mathematics Education_, _50_(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A. K., Herrmann, C., Gerlach, E., Zülsdorf-Kersting, M., Heinitz, B., & Nehring, A. (2020). Teaching quality in different subject matters in German-speaking countries—Inbetween genericness and subject-specificity. _Unterrichtswissenschaft_, _48_(3), 409–446. https://doi.org/10.1007/s42010-020-00082-8

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

Sherin, M. G. (2001). Developing a professional vision of classroom events: Teaching elementary school mathematics. In Beyond classical pedagogy (S. 75–93). Erlbaum.

Sherin, M. G. (2007). The development of teachers' professional vision in video clubs. In *Video research in the learning sciences* (pp. 383-395). Erlbaum.

Sherin, M. G., & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. Journal of Teacher Education, 60(1), 20–37. https://doi.org/10.1177/0022487108328155

Simonsmeier, B.A., Flaig, M., Deiglmayr, A., Schalk, L. & Schneider, M. (2022) Domain-specific prior knowledge and learning: A meta-analysis, Educational Psychologist, 57:1, 31-54, DOI: 10.1080/00461520.2021.1939700

Star, J. R., & Strickland, S. K. (2008). Learning to observe: Using video to improve preservice mathematics teachers' ability to notice. Journal of Mathematics Teacher Education, 11(2), 107–125. https://doi.org/10.1007/s10857-007-9063-7

Steffensky, M., Gold, B., Holdynski, M., & Möller, K. (2015). Professional Vision of Classroom Management and Learning Support in Science Classrooms—Does Professional Vision Differ Across General and Content-Specific Classroom Interactions? International Journal of Science and Mathematics Education, 13(2), 351–368. https://doi.org/10.1007/s10763-014-9607-0

Stürmer, K., Könings, K. D., & Seidel, T. (2013). Declarative knowledge and professional vision in teacher education: Effect of courses in teaching and learning. British Journal of Educational Psychology, 83(3), 467–483. https://doi.org/10.1111/j.2044-8279.2012.02075.x

Stürmer, K., Könings, K. D., & Seidel, T. (2014). Factors Within University-Based Teacher Education Relating to Preservice Teachers' Professional Vision. Vocations and Learning, 8(1), 35–54. https://doi.org/10.1007/s12186-014-9122-z

Sunder, C., Todorova, M., & Möller, K. (2016). Kann die professionelle Unterrichtswahrnehmung von Sachunterrichtsstudierenden trainiert werden? – Konzeption und Erprobung einer Intervention mit Videos aus dem naturwissenschaftlichen Grundschulunterricht. Zeitschrift Für Didaktik Der Naturwissenschaften, 22(1), 1–12. https://doi.org/10.1007/s40573-015-0037-5

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. Learning and Instruction, 46, 45–60. https://doi.org/10.1016/j.learninstruc.2016.08.003

Yin, R. K. (2018). Case Study Research and Applications: Design and Methods (6th ed.). Thousand Oaks, CA: Sage.

**Declaration of Interest Statement**

No potential conflict of interest was reported by the author(s).

### *11.5. Beitrag 5: Virtuelle Unterrichtshospitationen im Chemieunterricht – Eine Vernetzung der ersten und zweiten Phase der Lehrkräftebildung*

Heinitz & Nehring (2024). Virtuelle Unterrichtshospitationen im Chemieunterricht – Eine Vernetzung der ersten und zweiten Phase der Lehrkräftebildung. *Der mathematische und naturwissenschaftliche Unterricht : MNU*, 182-190.

Zusammenfassung:
Angehende Lehrkräfte stehen vor der Herausforderung, die theoretisch orientierten Inhalte ihres Studiums mit der Unterrichtspraxis zu verknüpfen. Die Webseite „VirtU-net Chemie“ der Leibniz Universität Hannover bietet Videovignetten für chemiedidaktische Lehrveranstaltungen in Universität und Referendariat, um dabei zu unterstützen. Eine phasenübergreifende Nutzung dieser Vignetten kann die Kohärenz der Ausbildung erhöhen und dazu beitragen, die Lücke zwischen Theorie und Praxis zu verkleinern.

CRediT Author Statement zur Eigenleistung:
Benjamin Heinitz: Conceptualization**,** Writing — original draft, review & editing, Investigation, Data Curation, Formal analysis, Visualization

# Virtuelle Unterrichtshospitationen im Chemieunterricht

## Eine Vernetzung der ersten und zweiten Phase der Lehrkräftebildung

Benjamin Heinitz – Andreas Nehring

**Angehende Lehrkräfte stehen vor der Herausforderung, die theoretisch orientierten Inhalte ihres Studiums mit der Unterrichtspraxis zu verknüpfen. Die Webseite „VirtU-net Chemie“ der Leibniz Universität Hannover bietet Videovignetten für chemiedidaktische Lehrveranstaltungen in Universität und Referendariat, um dabei zu unterstützen. Eine phasenübergreifende Nutzung dieser Vignetten kann die Kohärenz der Ausbildung erhöhen und dazu beitragen, die Lücke zwischen Theorie und Praxis zu verkleinern.**

### 1 *Einführung: Unterrichtswahrnehmung als Teil der Professionalisierung angehender Chemielehrkräfte*

In einer Unterrichtsstunde finden viele Handlungen gleichzeitig statt. Lehrkräfte stehen dadurch vor der stetigen Herausforderung, die für den Unterrichtsverlauf wichtigen Informationen herauszufiltern und entsprechend zu reagieren. Gezieltes Wahrnehmen und Interpretieren von relevanten Handlungen im Unterricht sind somit zentrale Fähigkeiten, die zur Professionalität von Chemielehrkräften gehören. In der Forschung werden diese Fähigkeiten mit dem Begriff der professionellen Unterrichtswahrnehmung verbunden (Sherin, 2001). Diese setzt voraus, dass eine Unterrichtshandlung als relevant wahrgenommen und basierend auf dem fachlichen, fachdidaktischen und pädagogischen Wissen der Lehrkraft interpretiert wird. Wird daraus wiederum eine konkrete Handlung abgeleitet, bezeichnet man dies als *Knowledge based Reasoning* (Sherin, 2001). Blömeke Gustafsson & Shavelson (2015) beschreiben diesen Prozess als situationsspezifisch und unterteilen ihn in die Wahrnehmung (*Perception*), Interpretation (*Interpretation*) und Entscheidungsfindung (*Decision Making*), die wiederum durch die Voraussetzungen der Lehrkräfte (*Disposition*) beeinflusst werden und letztendlich in einer beobachtbaren Handlung (*Performance)* der Lehrkraft münden. Empirische Befunde bestätigen die Erwartung, dass eine professionelle Unterrichtswahrnehmung zur Unterrichtsqualität beiträgt (Blömeke, Jentsch, Ross, Kaiser & König, 2022), indem fachliches und fachdidaktisches Wissen auf konkrete Unterrichtssituationen bezogen werden kann. Die Forschung zeigt jedoch auch, dass diese Prozesse bei erfahrenen Lehrkräften häufig unbewusst ablaufen (Berliner, 1989). Bei angehenden Lehrkräften müssen sie – angesichts der Vielzahl von Informationen – zunächst angebahnt und gefördert werden.

### 2 *Videovignetten in der Lehrkräftebildung*

Einer Unterrichtsstunde gehen viele Überlegungen in der Planung voraus, bei der auch unterschiedliche Unterrichtsabläufe antizipiert werden müssen Im Unterricht selbst laufen dann viele Situationen parallel ab, die durch die Lehrkraft gezielt evaluiert werden müssen, um eine angemessene Handlung zu generieren. Zusätzlich müssen die eigenen Handlungen stetig reflektiert werden. Dadurch liegt eine große Menge an Informationen vor, die gerade für angehende Lehrkräfte schnell überfordernd sein kann. Videovignetten (kurze aufbereitete Ausschnitte von Unterrichtsaufzeichnungen) haben sich deshalb in den letzten Jahren als ein hilfreiches Werkzeug in der Lehrkräftebildung etabliert. Sie können so geschnitten und gewählt werden, dass Komplexität reduziert wird und relevante Handlungen im Unterricht leichter identifizierbar und interpretierbar werden (Beck, King & Marshall, 2002). Videovignetten eignen sich deshalb, um die professionelle Unterrichtswahrnehmung in regelmäßiger Anwendung kontinuierlich zu trainieren (Sherin & van Es, 2009).

Besonders für die fachdidaktische Ausbildung in den Universitäten und den Ausbildungsseminaren der zweiten Phase der Lehrkräftebildung sind Videovignetten relevant, da sie theoretisches Wissen mit konkreten Handlungen der Unterrichtspraxis verknüpfen können. Die Übertragung in die spätere Praxis und den eigenen Unterricht wird erleichtert, wenn bereits Bezüge zu praktischen Beispielen hergestellt wurden. Damit wird einem trägen Wissen entgegengewirkt und das theoretische Wissen des Studiums in der Praxis leichter abrufbar.

Dementsprechend finden Videovignetten in der universitären Ausbildung von angehenden Lehrkräften bereits eine breite Anwendung (Blomberg, Renkl, Sherin, Borko & Seidel, 2013; Junker et al., 2022). Dabei werden sie sowohl für die Förderung von fächerübergreifenden Merkmalen der Unterrichtsqualität (z.B. Gold, Förster & Holodynski, 2013) als auch für die Förderung von fachspezifischen Merkmalen erfolgreich eingesetzt (z.B. Sunder, Todorova & Möller, 2016). Eine gezielte Vernetzung der ersten und zweiten Phase der Lehrkräftebildung mit Hilfe von Videovignetten, ist ein bisher selten umgesetztes Vorhaben. Dabei bietet die Vernetzung von Theorie und Praxis über den Bezug zu konkreten Unterrichtshandlungen ein hohes Potential für die Ausbildung angehender Lehrkräfte.



MNU-Journal – Ausgabe 03.2024 – ISSN 0025-5866 – © Verlag Klaus Seeberger, Neuss

**I. Auswahl und Thematisierung von Inhalten**

- Auswahl und Einbindung von Fachinhalten
- Auswahl und Einbindung naturwissenschaftlicher Denk- und Arbeitsweisen
- Lernen im Kontext und Verknüpfung zu socio-scientific issues
- Motivierende Einbettung der Inhalte
- Zielklarheit
- Strukturieren und Fokussieren von Schlüsselaspekten
- Horizontale und vertikale Vernetzung
- Strukturierter Ablauf und Sequenzierung der Stunde
- Progression innerhalb der Stunde
- Angemessene Repräsentation der Inhalte
- Nutzung präziser Fachsprache
- Vermeidung inhaltlicher Fehler
- Fachlich adäquate Einbindung von Inhalten und Denk- und Arbeitsweisen
- Adäquate didaktische Reduktion unter Berücksichtigung zukünftiger fachlicher Lernschritte

**II. Kognitive Aktivierung**

- Auswahl herausfordernder Lerngelegenheiten
- Problemlösendes Lernen
- Nutzung multipler Repräsentationen und Lösungswege
- Einsatz komplexer und vernetzender Aufgaben
- Kognitiv aktivierender Einsatz naturwissenschaftlicher Denk- und Arbeitsweisen
- Aktivierung von Vorwissen
- Kognitiv-konstruktive Fehlerkultur
- Explizierung von Denkprozessen
- Konstruktive Einbindung von eigenen Ideen und Schülervorstellungen in den Unterricht
- Unterstützung kognitiv aktivierender Prozesse
- Unterstützung und Einbindung metakognitiver Prozesse
- Explizierung von naturwissenschaftlichen Denkweisen
- Kooperatives Arbeiten zur zielführenden Aktivierung der Schüler/innen

**III. Unterstützung des Übens**

- Wiederholende Anwendung von Fachinhalten und Methoden
- Konstruktiver Umgang mit Fehlern und Schwierigkeiten von Schüler/inne/n beim Üben

**IV. Formatives Assessment**

- Unterrichtsbezogene Rückmeldung
- Regelmäßige Überprüfung des schülerseitigen Verständnisses
- Konstruktives Feedback
- Konstruktive Nutzung des Feedbacks

**V. Unterstützung des Lernens aller Schüler/innen**

- Autonomie der Schüler/innen
- Selbstwahrnehmung der Schüler/innen
- Individualisierung/Differenzierung
- Adaptive Erleichterung
- Förderung der Schülerbeteiligung
- Lernbegleitende Unterstützung der Schüler/innen

**VI. Sozio-emotionale Unterstützung**

- Unterstützende Lehrer/in-Schüler/in-Interaktion
- Beziehung der Schüler/innen untereinander
- Inhalt (Antizipation von Angst, Respekt oder Ekel)

**VII. Klassenführung**

- Allgegenwärtigkeit
- Prävention
- Intervention
- Zeitliche Strukturierung der Stunde
- Phasenübergänge
- Pacing
- Sicherheit
- Vorbereitete Umgebung

Abb. 1. Das SEP-Framework – unterteilt in sieben Dimensionen der Unterrichtsqualität (übersetzt und adaptiert, HEINITZ & NEHRING, 2023)

## 3 VirtU-net Chemie – Plattform für virtuelle Unterrichtshospitationen

Die Webseite VirtU-net Chemie der Leibniz Universität Hannover stellt Videovignetten für Seminare der ersten und zweiten Phase der Lehrkräftebildung zur Verfügung. Diese sind in Zusammenarbeit mit verschiedenen Lehrkräften für die Nutzung auf VirtU-net Chemie entstanden und dürfen zum Zweck der Lehrkräftebildung verwendet werden. VirtU-net bietet dabei vorrangig Chemieunterricht an, allerdings werden im Rahmen der Vignetten auch fächerübergreifende Merkmale thematisiert. Zusätzlich gibt es eine Verknüpfung zur Videoplattform der Englischdidaktik („VirtU"), wodurch die professionelle Unterrichtswahrnehmung auch fächerübergreifend betrachtet werden kann.

Die Vignetten auf VirtU-net Chemie sind so gewählt, dass sie ein breites Spektrum unterschiedlicher Themen und Unterrichtssituationen enthalten und werden stetig ergänzt.

Die Videovignetten werden als Gelenkstelle zwischen Theorie und Praxis genutzt, um die erste und zweite Phase der Ausbildung zu vernetzen. Ein wiederholtes Aufgreifen derselben Vignetten ermöglicht es, die auf Wissen beruhende Wahrnehmung und Interpretation in früheren Lernphasen stärker zu fördern und in späteren Lernphasen vor allem das Generieren von Handlungsoptionen in den Vordergrund zu stellen, ohne dabei alle Grundlagen neu zu erarbeiten. Der zunehmende Praxisbezug bildet sich in den Aufgaben ab, die zu jeder Vignette verfügbar sind.

## 4 Was zeichnet die Qualität von Chemieunterricht aus? – Die theoretische Grundlage von VirtU-net Chemie

Um einen Beitrag dazu zu leisten, Theorie und Praxis zu verknüpfen, baut VirtU-net Chemie auf bestehenden Theorien und Evidenzen der Unterrichtsqualität auf. Als Grundlage für die Merkmale dient das *Science Education Perspectives* (SEP)-Framework (Heinitz & Nehring, 2023), welches sowohl generische als auch fachspezifische und hybride Merkmale der Unterrichtsqualität berücksichtigt. Das Framework wurde auf Basis eines Literaturreviews zu Qualitätsmerkmalen der Naturwissenschaftsdidaktiken erstellt und bildet die Unterrichtsqualität entsprechend umfassend ab (Heinitz & Nehring, 2020). Das SEP-Framework gliedert sich in sieben Dimensionen der Unterrichtsqualität, welche durch insgesamt 50 Merkmale weiter unterteilt werden (Abb. 1). So wird z.B. die Dimension I „Auswahl und Thematisierung von Inhalten" durch die Merkmale „Auswahl und Einbindung von Fachinhalten" oder auch „Motivierende Einbettung der Inhalte" weiter ausdifferenziert. Wichtig ist hierbei, dass ein guter Unterricht sich nicht durch eine hohe Ausprägung aller Merkmale auszeichnet, sondern abhängig vom Ziel der Unterrichtsstunde unterschiedliche Merkmale relevant sein können. So enthält z.B. nicht jede Stunde eine Möglichkeit zur „wiederholenden Anwendung von Fachinhalten und Methoden" (Dimension III, Abb. 1), dennoch kann dieses Merkmal relevant sein, wenn bestimmte Abläufe eingeübt werden sollen. Entsprechend muss die Auswahl der Merkmale zur Beurteilung flexibel angepasst werden, was auch Teil einer professionellen Unterrichtswahrnehmung ist. Es wird

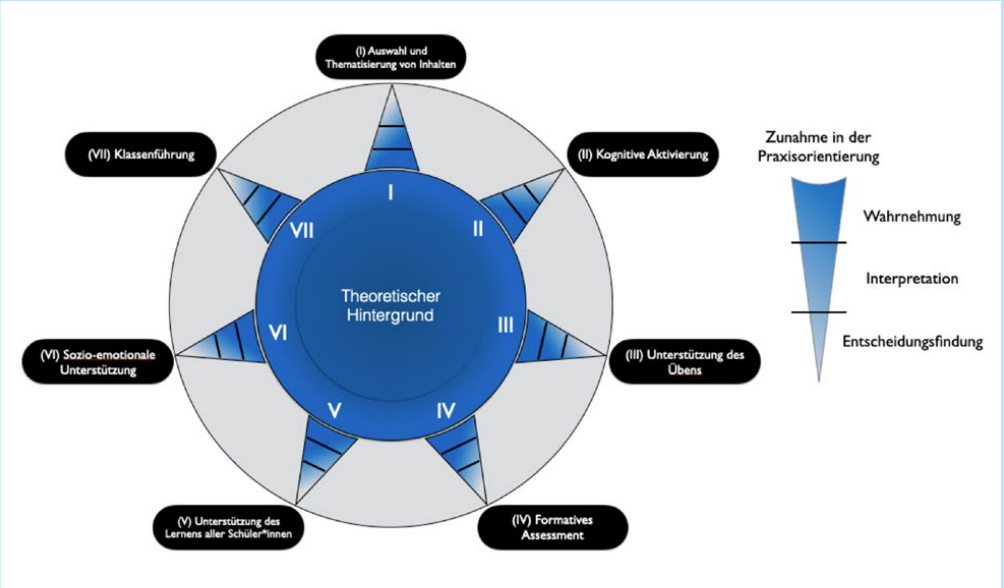


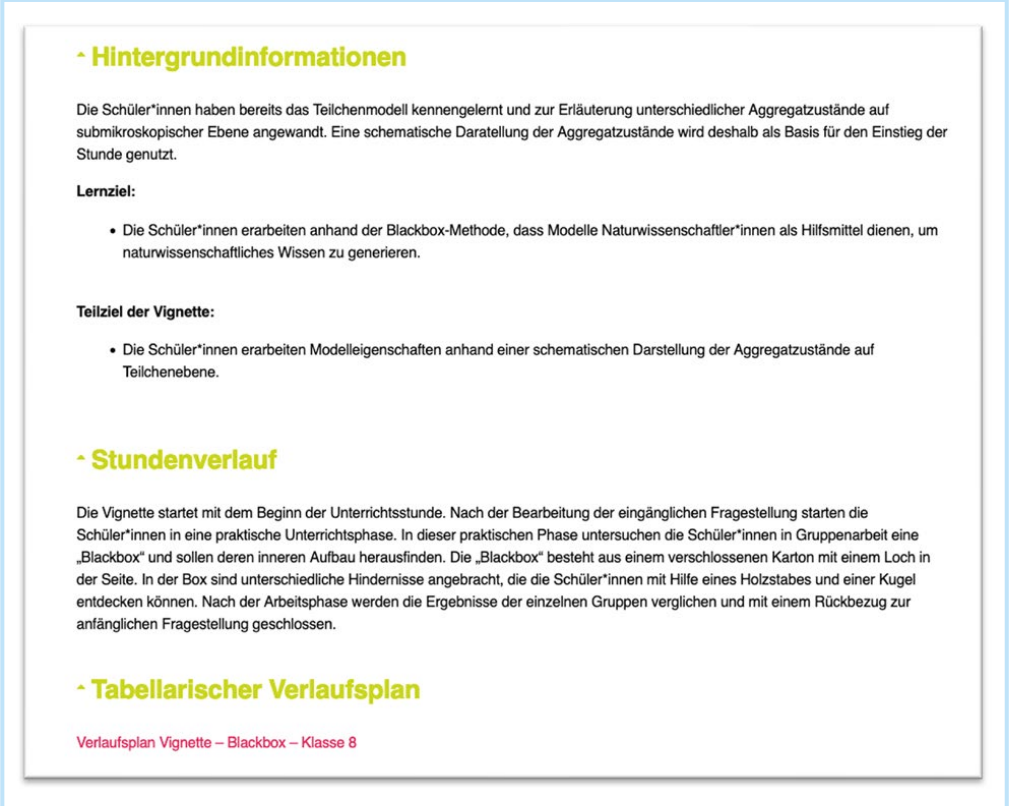

Abb. 2. STAR-Model zur kriterienorientierten Strukturierung der professionellen Unterrichtswahrnehmung auf VirtU-net Chemie.

folglich nicht immer das gesamte SEP-Framework genutzt, sondern innerhalb des Frameworks auf bestimmte Merkmale fokussiert.

Durch die Verwendung einheitlicher Merkmale können Parallelen zwischen Unterrichtssituationen einfacher aufgezeigt und eine gemeinsame Fachsprache der Unterrichtsqualität entwickelt werden. Ein Merkmal wird schrittweise mit konkreten Unterrichtssituationen verknüpft und erhält dadurch einen immer stärkeren Praxisbezug. Dadurch kann auch ein Transfer auf eine unbekannte Unterrichtssituation erleichtert werden. Jedes Merkmal benötigt zur Beurteilung ein bestimmtes theoretisches Wissen, das zuvor aktiviert oder zunächst erarbeitet werden muss. Damit bilden die Merkmale eine Brücke zwischen Theorie und Praxis.

Ein gemeinsames Framework der Unterrichtsqualität erleichtert auch das Erkennen von Beziehungen zwischen Merkmalen. Das SEP-Framework bietet damit einen strukturierten Übergang zwischen fachlichem bzw. fachdidaktischem Wissen, Qualitätsmerkmalen und der professionellen Unterrichtswahrnehmung.

Die Verknüpfung der drei Aspekte wird durch das *Systematic and Transferable Approach for Ratings of Instructional Quality* (STAR)-Model (Abb. 2) verdeutlicht. Das Modell bietet eine Möglichkeit, um die Analyse einer Unterrichtshandlung transparent darzustellen und wird auf VirtU-net Chemie als theoretische Grundlage genutzt. Ausgehend von der theoretischen Grundlage wird eine Unterrichtshandlung entlang einer der sieben Dimensionen analysiert. Die Analyse staffelt sich in die Unterpunkte *Wahrnehmung*, *Interpretation* und *Entscheidungsfindung*, wobei Aufgaben zu jedem Unterpunkt vorliegen. Mit dem Abschluss der *Entscheidungsfindung* soll eine mögliche Handlung generiert werden, die zwar nicht direkt umgesetzt, aber zumindest theoretisch besprochen werden kann. Dies kann einen Diskussionsanlass in den Seminaren der Lehrkräftebildung bieten, um somit den iterativen Prozess der professionellen Unterrichtswahrnehmung (Wöhlke, 2020) bereits in der Theorie anzubahnen. Das Modell verdeutlicht aber auch, dass eine Unterrichtshandlung ebenso entlang unterschiedlicher Qualitätsmerkmale analysiert werden kann. Eine Handlung kann zwar mit Blick auf ein bestimmtes Merkmal geplant sein, aber durch andere Personen mit anderen Merkmalen beurteilt werden.



## Hintergrundinformationen

Die Schüler*innen haben bereits das Teilchenmodell kennengelernt und zur Erläuterung unterschiedlicher Aggregatzustände auf submikroskopischer Ebene angewandt. Eine schematische Daratellung der Aggregatzustände wird deshalb als Basis für den Einstieg der Stunde genutzt.

**Lernziel:**

- Die Schüler*innen erarbeiten anhand der Blackbox-Methode, dass Modelle Naturwissenschaftler*innen als Hilfsmittel dienen, um naturwissenschaftliches Wissen zu generieren.

**Teilziel der Vignette:**

- Die Schüler*innen erarbeiten Modelleigenschaften anhand einer schematischen Darstellung der Aggregatzustände auf Teilchenebene.

## Stundenverlauf

Die Vignette startet mit dem Beginn der Unterrichtsstunde. Nach der Bearbeitung der eingänglichen Fragestellung starten die Schüler*innen in eine praktische Unterrichtsphase. In dieser praktischen Phase untersuchen die Schüler*innen in Gruppenarbeit eine „Blackbox" und sollen deren inneren Aufbau herausfinden. Die „Blackbox" besteht aus einem verschlossenen Karton mit einem Loch in der Seite. In der Box sind unterschiedliche Hindernisse angebracht, die die Schüler*innen mit Hilfe eines Holzstabes und einer Kugel entdecken können. Nach der Arbeitsphase werden die Ergebnisse der einzelnen Gruppen verglichen und mit einem Rückbezug zur anfänglichen Fragestellung geschlossen.

## Tabellarischer Verlaufsplan

Verlaufsplan Vignette – Blackbox – Klasse 8

Abb. 3. Screenshot der Hintergrundinformationen zur Vignette „Blackbox – Klasse 8 – Unterrichtseinstieg" im Themenbereich „Arbeiten mit Modellen" auf VirtU-net Chemie

## 5 Aufbau und Navigation auf der Webseite

Das zentrale Element der Webseite ist die Auswahlübersicht der Unterrichtsvignetten. Diese können über ein Kategoriensystem nach spezifischen Themen (z.B. Säure-Base; Arbeit mit Modellen; Chemische Reaktion) und Unterrichtsphasen (z.B. Einstieg; Erarbeitungsphase; Auswertung) sortiert werden. Sobald eine Unterrichtsvignette ausgewählt wurde, können die Informationen zum Hintergrund der Stunde aufgerufen werden (Abb. 3). Hierbei werden u.a. das Thema der Stunde, die Klassenstufe und das Lernziel sowie das Teilziel der Vignette genannt. Um die Vignette einfacher einordnen zu können, wird diese durch eine kurze Beschreibung im Stundenverlauf verortet. Hierbei werden die weiteren Abläufe lediglich skizziert, sodass eine Beurteilung zum Erreichen der Lernziele bei den Nutzenden liegt. Zusätzlich gibt es auch Vignetten, bei denen die späteren Abschnitte derselben Stunde angesehen werden können, um die Folgen bestimmter Unterrichtssituationen nachzuvollziehen. Durch den schriftlichen Stundenverlauf wird lediglich der geplante Ablauf zugänglich gemacht, der sich jedoch vom tatsächlichen Verlauf unterscheiden kann.

Alle Videovignetten können direkt in Seminaren der Lehrkräftebildung eingesetzt werden, wobei die konkrete Einbindung an den individuellen Einsatzzweck angepasst werden kann. Die Vignette „Blackbox – Klasse 8 – Unterrichtseinstieg“ (Abb.3) kann z.B. genutzt werden, um die notwendigen Voraussetzungen für den Umgang mit dem Teilchenmodell zu erarbeiten. Alle zusätzlichen Informationen sind zunächst hinter ausklappbaren Menüs verborgen, sodass z.B. auch das zugrundeliegende Modell aus der Vignette abgeleitet werden kann, ohne dass dieses vorgegeben ist. Bei Bedarf stehen Arbeitsaufträge und eine manualbasierte Kodierung der Unterrichtsqualität zu jeder Vignette zur Verfügung, sodass sie auch ohne weitere Rahmung nutzbar sind. Diese können durch die Navigation am unteren Rand der Seite jederzeit aufgerufen werden. Die Videovignette ist dabei jeweils so eingebettet, dass sie unabhängig von den zusätzlichen Materialien abgespielt werden kann. Die Menüleiste am oberen Rand der Webseite bietet jederzeit den Übergang zur gesamten Auswahl der Videovignetten und zu den theoretischen Grundlagen von VirtU-net Chemie.

## 6 Aufgaben zur Vernetzung von Theorie und Praxis

In Zusammenarbeit mit Vertreter/innen der ersten und zweiten Phase, wurden Aufgaben zu den Videovignetten entwickelt, die sich am STAR-Model orientieren und direkt in den Seminaren genutzt werden können. Die Aufgaben staffeln sich und gehen sowohl auf fachliches und fachdidaktisches Wissen (Abb. 4) als auch auf konkrete Beobachtungsaufträge (Abb. 5) für die jeweiligen Unterrichtssituationen ein. Zunächst wird das notwendige theoretische Wissen (re-)aktiviert, um dann zielführend mögliche Handlungen abzuleiten.

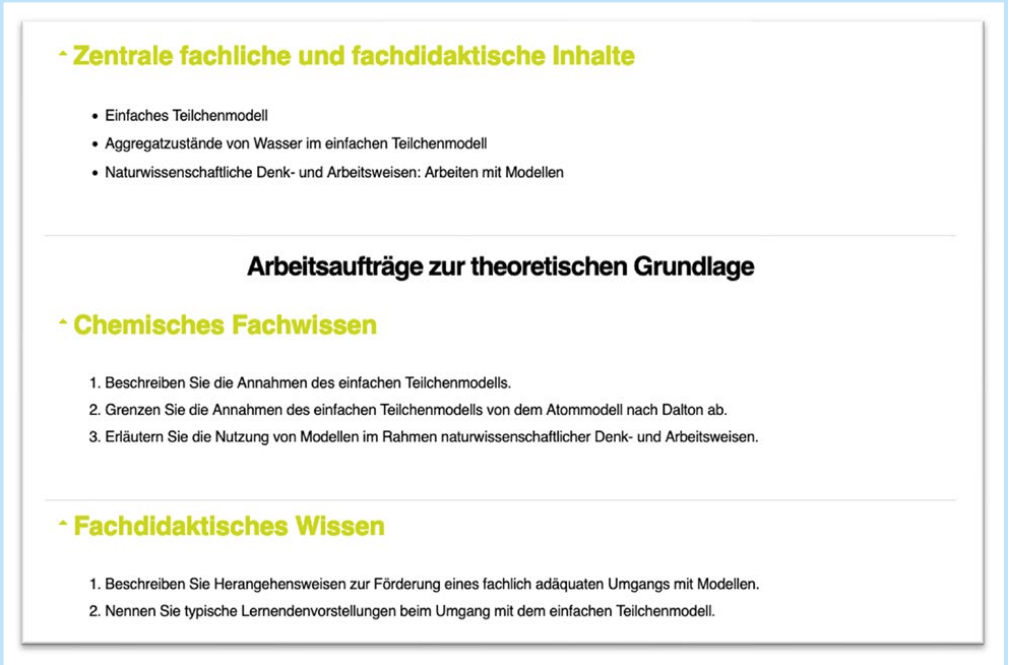



Abb. 4. Screenshot der Arbeitsaufträge zum fachlichen und fachdidaktischen Wissen der Vignette „Blackbox – Klasse 8 – Unterrichtseinstieg“ im Themenbereich „Arbeiten mit Modellen“ auf VirtU-net Chemie

Die Einbettung der Aufgaben soll im Folgenden am Beispiel der Vignette „Blackbox – Klasse 8 – Unterrichtseinstieg“ näher erläutert werden.

### Zusammenfassung der Vignette:

Die Vignette zeigt den Einstieg einer Unterrichtsstunde, in der die Lehrkraft mit einem stillen Impuls startet. Dazu zeigt sie eine Abbildung, auf der die Aggregatzustände mit Hilfe des einfachen Teilchenmodells dargestellt sind. Die Schüler/innen äußern unterschiedliche Interpretationen der Abbildung und beziehen sich dabei auf das Teilchenmodell und die Aggregatzustände. Die Lehrkraft lenkt das Gespräch mit Rückfragen und erarbeitet den Modellcharakter der Abbildung. Abschließend wird die Fragestellung der Unterrichtsstunde durch die Lehrkraft vorgegeben: „Wie kommen Naturwissenschaftler zu ihren Erkenntnissen?“

### Aufgaben zum chemischen Fachwissen und fachdidaktischen Wissen:

Für die Beurteilung der Vignette ist es zunächst wichtig, dass die Annahmen des einfachen Teilchenmodells bekannt und abgrenzbar zu anderen Denkmodellen sind, um die Interaktion zwischen den Schüler/inne/n und der Lehrkraft zu beurteilen. Somit zielen die Fragen zum chemischen Fachwissen (Abb. 4) auch auf Inhalte ab, die nicht direkt Teil der Unterrichtsstunde sind, sondern notwendige Voraussetzungen darstellen und ggf. in vorangegangenen Unterrichtsstunden behandelt wurden. Die Fragen zum fachdidaktischen Wissen greifen diese Inhalte auf und fokussieren auf deren Einbindung in den Unterricht.

Da die Aufgaben auch vor der Betrachtung der Vignetten behandelt werden können, sind die Formulierungen allgemeiner gehalten und erlauben ein breites Spektrum unterschiedlicher Einsatzzwecke. So kann die Aufgabe „Beschreiben Sie Herangehensweisen zur Förderung eines fachlich adäquaten Umgangs mit Modellen.“ auch in die Planung einer ganzen Unterrichtsstunde integriert werden.

Bei der Aufgabe „Nennen Sie typische Lernendenvorstellungen beim Umgang mit dem einfachen Teilchenmodell.“ könnte ebenso in unterschiedliche Richtungen argumentiert werden und damit eine Diskussion mit eher theoretischem Fokus vorbereitet werden. Die Rahmung kann hierbei passend zum jeweiligen Einsatzzweck im Vorfeld festgelegt werden.

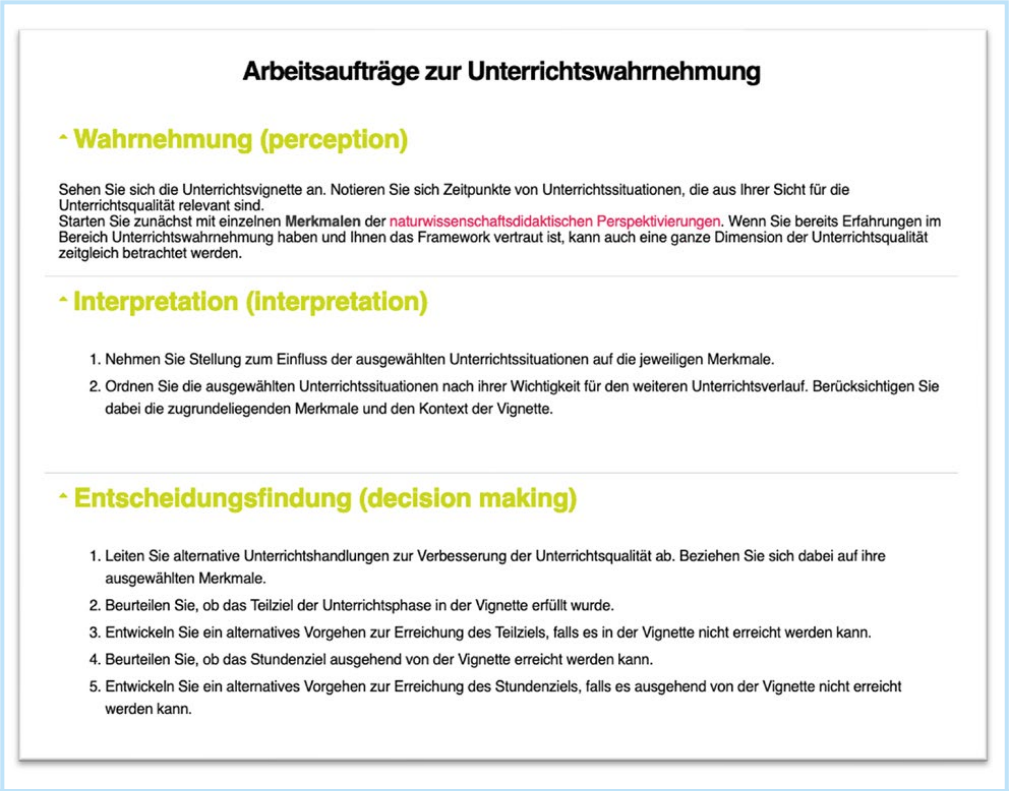



Abb. 5. Screenshot der Arbeitsaufträge zur Unterrichtswahrnehmung in der Vignette „Blackbox – Klasse 8 – Unterrichtseinstieg“ im Themenbereich „Arbeiten mit Modellen“ auf VirtU-net Chemie

### Aufgaben zur Unterrichtswahrnehmung:

Die Arbeitsaufträge zur Unterrichtswahrnehmung setzen die direkte Arbeit mit der Vignette voraus. Dabei wird die professionelle Wahrnehmung entlang von Merkmalen der Unterrichtsqualität trainiert. Besonders für den Einstieg in die Analyse von Unterrichtsaufzeichnungen bietet es sich an, zunächst mit einzelnen Merkmalen des SEP-Frameworks zu arbeiten (z.B. „Notieren Sie Zeitpunkte in der Vignette, die aus Ihrer Sicht relevant sind für die Unterstützung und Einbindung metakognitiver Prozesse."). Hierbei kommen bei der Vignette unterschiedliche Situationen infrage, da dieses Kriterium für einen reflektierten Umgang mit Modellen besonders wichtig ist. So wird z.B. von einer Schülerin die Abbildung zu Beginn der Stunde interpretiert als „der erste Teil ist fest, der zweite ist flüssig und der dritte ist gasförmig", worauf die Lehrkraft erwidert „ich sehe da eigentlich nur Kreise".
Diese Situation bietet einen Anlass, über die Umsetzung des Kriteriums zu diskutieren und die Frage aufzuwerfen, inwiefern die Handlung der Lehrkraft zur Erreichung des Lernziels beiträgt.

Sobald der Umgang mit dem Framework bekannt ist, können auch mehrere Merkmale gemeinsam für die Beurteilung der Unterrichtqualität herangezogen werden. Damit kann auch eine Diskussion über das Zusammenspiel einzelner Merkmale, deren Voraussetzungen oder über die Priorisierung in bestimmten Unterrichtssituationen angebahnt werden. So kann die beispielhaft genannte Aussage der Lehrkraft nicht nur mit einem Blick auf das Modellverständnis interpretiert werden, sondern auch mit Blick auf die „unterstützende Lehrer/in-Schüler/in-Interaktion". Die Aussage der Lehrkraft ist provokativ formuliert, um einen reflektierten Umgang der Schüler/innen mit dem Modell zu unterstützen. Sie kann aber auch durchaus negativ durch die betreffende Schülerin aufgefasst werden, deren Aussage in diesem Fall nicht wertgeschätzt wird.

Jede Handlung in den Vignetten kann mit einem oder auch mit mehreren Merkmalen der Unterrichtsqualität verknüpft werden. Wenn später weitere Vignetten analysiert werden, ergeben sich Transfermöglichkeiten für die Anwendung des theoretischen Wissens.

**Manualbasierte Kodierung der Vignette**

▸ **Dimension I - Auswahl und Thematisierung von Fachinhalten**

▾ **Dimension II - Kognitive Aktivierung**

| Perception | | Interpretation | |
|---|---|---|---|
| **Zeit** | **Unterrichtshandlung** | **Kodierung** | **Bewertung der Unterrichtshandlung mit Erläuterung** |
| 0:30 | Einstieg – Stummer Impuls, unterstützt durch das Schema von Aggregatzustandsänderung im Teilchenmodell | Aktivierung von Vorwissen | Die Schüler*innen erhalten keine weiteren Arbeitsanweisungen und müssen sich die Aufgabe selbst erschließen. Vermutlich ist der Inhalt bekannt (mehrere Meldungen), weshalb an dieser Stelle auf das Vorwissen der Schüler*innen zurückgegriffen wird. |
| 0:45 | S. beschreibt Schema als „das sind die drei Aggregatzustände"<br><br>L. geht nicht auf die Aussage ein | Explizierung von Denkprozessen | Hier wäre eine weitere Exploration der Denkweise von S. möglich gewesen. Worin besteht der Zusammenhang zwischen dem Schema und der Interpretation? Überprüfung des Modellverständnisses |
| 0:55 | S. beschreibt Schema als „fest, flüssig und gasförmig"<br><br>L.: „Ich sehe da eigentlich nur Kreise." | Unterstützung und Einbindung metakognitiver Prozesse | Erläuterung der Schüler*innen zu den Aggregatzuständen wird kontrastiert. Schüler*innen haben interpretiert, was im Schema abgebildet wird und L. beschreibt die Abbildung als einfache Beobachtung. Hier wird eine metakognitive Denkleistung eingefordert, da sich die Schüler*innen bewusst werden müssen, dass sie ein Denkmodell nutzen, um das Anschauungsmodell zu interpretieren. |

Abb. 6. Screenshot der manualbasierten Kodierung einer Videovignette auf VirtU-net Chemie

In einer weiteren Vignette im Themenfeld „chemische Reaktion“ zeigt sich in der Aussage eines Schülers, dass dieser eine Hybridvorstellung zwischen dem Teilchenmodell und dem Atommodell nach Dalton entwickelt hat. Eine stärkere Einbindung metakognitiver Prozesse, hätte diese Situation beeinflussen können und stellt somit eine direkte Verknüpfung zwischen zwei unterschiedlichen Unterrichtsstunden dar, die jedoch vor demselben theoretischen Hintergrund diskutiert werden können. Zusätzlich kann diese Situation genutzt werden, um eine mögliche Handlung für einen konstruktiven Umgang mit der aufgetretenen Lernendenvorstellung abzuleiten. Die manualbasierte Kodierung der Vignetten bietet dabei ein großes Potential für die Vernetzung unterschiedlicher Unterrichtssituationen, da die beobachtbaren Merkmale direkt benannt werden.

Die Aufgaben zur Interpretation und Entscheidungsfindung sind jeweils abhängig von den gewählten Merkmalen im Bereich der Wahrnehmung. Die Aufgaben stellen die gewählten Situationen außerdem in einen Zusammenhang mit den zentralen Inhalten (Abb. 4) und den Teilzielen (hier: Die Schüler/innen erarbeiten Modelleigenschaften anhand einer schematischen Darstellung der Aggregatzustände auf Teilchenebene), bzw. dem Lernziel (hier: Die Schüler/innen erarbeiten anhand der Blackbox-Methode, dass Modelle Naturwissenschaftlern als Hilfsmittel dienen, um naturwissenschaftliches Wissen zu generieren).

Das Lernziel der Unterrichtsstunde wird in allen Fällen genannt. Die Frage, ob es erreicht wird, kann jedoch im Rahmen einer einzelnen Vignette in der Regel nicht vollständig beurteilt werden. Die Aufgaben zur Unterrichtswahrnehmung nehmen zwar Bezug auf die Lernziele, bieten dabei jedoch viel Raum für Interpretation. In einigen Fällen können die Fortsetzungen der Vignetten einen vertieften Einblick in die Auswirkung bestimmter Handlungen bieten. Damit die Vignetten für sich allein genommen ebenfalls mit Blick auf ein Ziel beurteilt werden können, werden Teilziele für alle Vignetten genannt.

Zwischen den Aufgaben der ersten und zweiten Phase der Lehrkräftebildung gibt es keine scharfe Trennlinie. So können auch in der ersten Phase bereits Arbeitsaufträge zur Entwicklung alternativer Unterrichtshandlungen eingebunden oder in der zweiten Phase theoretische Grundlagen erneut aufgearbeitet werden.

Die manualbasierte Kodierung (Abb. 6) der Videovignetten wird als separates Begleitmaterial angeboten. Diese enthält potentiell relevante Unterrichtssituationen (*Wahrnehmung*) sowie eine Einschätzung mit Bezug zur Unterrichtsqualität (*Interpretation*) mit Angabe von Qualitätsmerkmalen. Die Kodierung bietet damit viele Lösungsmöglichkeiten für die Aufgaben zur Unterrichtswahrnehmung und kann nach der Bearbeitung der Aufgaben für einen Abgleich genutzt werden. Sie kann jedoch auch durch Lehrende für die Vorbereitung eines Seminars genutzt werden, um individuelle Analyseschwerpunkte für den Einsatz einer Vignette festzulegen. Das Kodiermanual enthält Beschreibungen für alle Merkmale des SEP-Frameworks (Abb. 1) und bildet somit viele Facetten der Unterrichtsqualität ab. Diese treten teilweise parallel auf und können gemeinsam oder getrennt thematisiert werden.

## 7 Zugang zur Plattform

Der Zugang zur Webseite ist aus datenschutzrechtlichen Gründen nur mit einer Anmeldung möglich. Grundsätzlich können alle Personen, die in der Lehrkräftebildung tätig sind, nach Zustimmung zur Nutzungsvereinbarung, einen Zugang zur Webseite erhalten. Alle zusätzlichen Materialien dürfen unter Berücksichtigung der Creative Commons-Lizenz heruntergeladen und bearbeitet werden.
Eine Anmeldung erfolgt direkt über die Plattform
https://virtu-net.idn.uni-hannover.de/
Für die Nutzung der Webseite fallen keine Kosten an.

## 8 Ausblick zur weiteren Entwicklung der Plattform

Die Plattform wird stetig durch weitere Videovignetten ergänzt, wobei auch Vorschläge für spezifische Themen und Unterrichtsabschnitte eingereicht werden können. Neben der Bereitstellung von Videovignetten soll langfristig auch ein Austausch zwischen unterschiedlichen Ausbildungsorten ermöglicht werden. Hierzu werden Möglichkeiten implementiert, um Videovignetten oder auch Aufgabenstellungen direkt zu kommentieren und auf die Kommentare anderer Personen einzugehen. VirtU-net Chemie soll weiterführend auch durch 360°-Aufnahmen ergänzt werden, um eine neue Perspektive auf den Unterricht zu ermöglichen. Hierbei wird aktiv mit der noch stärkeren Informationsfülle der Vignetten gearbeitet und gleichzeitig eine Immersion in den Chemieunterricht ermöglicht, um eine gezielte professionelle Wahrnehmung der Unterrichtsqualität zu trainieren. Eine fächerübergreifende Nutzung ist im Konzept der Webseite vorgesehen und wird bereits durch die Thematisierung von fächerübergreifenden Merkmalen sowie die Verknüpfung mit dem Fach Englisch (VirtU) umgesetzt.

## Literatur


Beck, R. J., King, A. & Marshall, S. K. (2002). Effects of videocase construction on preservice teachers' observations of teaching. *Journal of Experimental Education*, *70*(4), 345–361. https://doi.org/10.1080/00220970209599512

Berliner, D. C. (1989). Implications of studies of expertise in pedagogy for teacher education and evaluation. In *New directions for teacher assessment: Proceedings of the 1988 ETC invitational conference. 39–68. Princeton, NJ: Educational Testing Service.*

Blomberg, G., Renkl, A., Sherin, M. G., Borko, H. & Seidel, T. (2013). Five research-based heuristics for using video in pre-service teacher education. *Journal for Educational Research*, 90–114. https://doi.org/10.25656/01:8021

Blömeke, S., Gustafsson, J. E. & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie / Journal of Psychology* (Vol. 223, Issue 1, pp. 3–13). Hogrefe Publishing. https://doi.org/10.1027/2151-2604/a000194

Blömeke, S., Jentsch, A., Ross, N., Kaiser, G. & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progress. *Learning and Instruction*, 79. https://doi.org/10.1016/j.learninstruc.2022.101600

Gold, B., Förster, S. & Holodynski, M. (2013). Evaluation eines videobasierten Trainingsseminars zur Förderung der professionellen Wahrnehmung von Klassenführung im Grundschulunterricht. *Zeitschrift für Pädagogische Psychologie*, *27*(3), 141–155. https://doi.org/10.1024/1010-0652/a000100

Heinitz, B. & Nehring, A. (2020). Quality of instruction in science education – a systematic review including goals, contents, and methods of science education. *Unterrichtswissenschaft*, 48(3), 319–360. Springer VS. https://doi.org/10.1007/s42010-020-00074-8

Heinitz, B. & Nehring, A. (2023). Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors. *International Journal of Science Education*. https://doi.org/10.1080/09500693.2023.2213382

Junker, R., Zucker, V., Oellers, M., Rauterberg, T., Konjer, S., Meschede, N. & Holodynski, M. (2022). *Lehren und Forschen mit Videos in der Lehrkräftebildung*. Münster; New York: Waxmann 2022.

Sherin, M.G. (2001). Developing a professional vision of classroom events: Teaching elementary school mathematics. In Wood, T., Nelson, B.& Warfieeld, S. (Eds.*), Beyond classical pedagogy: Teaching elementary school mathematics*, 75–93. New York: Routledge.

Sherin, M.G. & van Es, E.A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, *60*(1), 20–37. https://doi.org/10.1177/0022487108328155

Sunder, C., Todorova, M. & Möller, K. (2016). Kann die professionelle Unterrichtswahrnehmung von Sachunterrichtsstudierenden trainiert werden? – Konzeption und Erprobung einer Intervention mit Videos aus dem naturwissenschaftlichen Grundschulunterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, 22 (1), 1–12

Wöhlke, C. (2020). *Entwicklung und Validierung eines Instruments zur Erfassung der professionellen Unterrichtswahrnehmung angehender Physiklehrkräfte*. (Volume 298). Berlin: Logos Verlag.

*Benjamin Heinitz ist Mitarbeiter im Projekt „VirtU-net Chemie" der Leibniz Universität Hannover und derzeit wissenschaftlicher Mitarbeiter an der Universität Münster, Fliednerstr. 21, 48149, Münster, b.heinitz@uni-muenster.de*

*Prof. Dr. Andreas Nehring ist Leiter des Projekts „VirtU-net Chemie" der Leibniz Universität Hannover und derzeit Professor für Didaktik der Naturwissenschaften (Schwerpunkt Chemiedidaktik) an der Leibniz Universität Hannover, Am Kleinen Felde 30, 30167, Hannover, nehring@idn.uni-hannover.de*

## *11.6. Lebenslauf & Publikationsliste*

Beruflicher Werdegang

| | |
|---|---|
| **Seit 10/2023** | **Wissenschaftlicher Mitarbeiter**<br>Institut für Psychologie in Erziehung und Bildung - Entwicklungspsychologische Voraussetzungen für Erziehung und Unterricht (AE Prof. Dr. Holodynski) - Universität Münster |
| **04/2021 – 09/2023** | **Betreuung des Projekts „VirtU-net Chemie - Plattform für virtuelle Unterrichtshospitationen“ zur Vernetzung von Studium und Referendariat**<br>Institut für Didaktik der Naturwissenschaften - Fachgebiet Didaktik der Chemie (AG Prof. Dr. Nehring) - Leibniz Universität Hannover |
| **04/2019 09/2023** | **Wissenschaftlicher Mitarbeiter**<br>Institut für Didaktik der Naturwissenschaften - Fachgebiet Didaktik der Chemie (AG Prof. Dr. Nehring) - Leibniz Universität Hannover |

Akademischer Werdegang

| | |
|---|---|
| **Seit 04/2020** | **Promotionsstudium**<br>Institut für Didaktik der Naturwissenschaften - Fachgebiet Didaktik der Chemie (AG Prof. Dr. Nehring) - Leibniz Universität Hannover |
| **10/2016 - 04/2019** | **Studium - Lehramt an Gymnasien mit den Fächern Chemie und Biologie**<br>Leibniz Universität Hannover |
| **10/2012 – 10/2016** | **Studium - Fächerübergreifender Bachelor mit den Fächern Chemie und Biologie**<br>Leibniz Universität Hannover |
| **06/2011** | **Gymnasiale Schulausbildung**<br>Gymnasium Johanneum Lüneburg |

Publikationen

**Peer-Reviewed**

1. Heinitz, B. & Nehring, A. (2023) Instructional quality in science teacher education: comparing evaluations by chemistry pre-service teachers and their advisors, *International Journal of Science Education* 10.1080/09500693.2023.2213382
2. Heinitz, B., Szogs, M., Förtsch, C., Korneck, F., Neuhaus, B., & Nehring, A. (2022). Unterrichtsqualität in den Naturwissenschaften: Eine vergleichende Gegenüberstellung von Ansätzen zwischen Fachspezifik und Generik. *Zeitschrift für Didaktik der Naturwissenschaften (ZfDN)*, *28*(1), [10]. doi.org/10.1007/s40573-022-00146-5
3. Heinitz, B., & Nehring, A. (2020). Kriterien naturwissenschaftsdidaktischer Unterrichtsqualität – ein systematisches Review videobasierter Unterrichtsforschung. *Unterrichtswissenschaft*. https://doi.org/10.1007/s42010-020-00074-8
4. Praetorius, A.K., Herrmann, C., Gerlach, E., Zülsdorf-Kersting, M., Heinitz, B., & Nehring, A. (2020). Unterrichtsqualität in den Fachdidaktiken im deutschsprachigen Raum – zwischen Generik und Fachspezifik. *Unterrichtswissenschaft* 48, 409–446. https://doi.org/10.1007/s42010-020-00082-8

**Weitere Beiträge**

5. Heinitz, B. & Nehring, A. (2024) Virtuelle Unterrichtshospitationen im Chemieunterricht – Eine Vernetzung der ersten und zweiten Phase der Lehrkräftebildung. *Der mathematische und naturwissenschaftliche Unterricht : MNU*, 182-190.

6. Heinitz, B., Nehring, A. (2023). Wie beurteilen Referendar*innen kognitive Aktivierung in Videovignetten? Eine explorative Beobachtungsstudie. In H. van Vorst (Hrsg.) *Lernen, Lehren und Forschen in einer digital geprägten Welt* (S. 250 – 253). GDCP-Tagungsband
7. Heinitz, B., Nehring, A. (2022). Unterrichtsqualitätseinschätzungen im Referendariat – Fach- und Seminarleiter*innen im Vergleich mit Referendar*innen. In S. Habig & H. van Vorst (Hrsg.) *Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen* (S. 448 – 451). GDCP-Tagungsband.
8. Heinitz, B., Nehring, A. (2021). Beurteilung von Unterrichtsqualität im Referendariat. In S. Habig (Hrsg.) *Naturwissenschaftlicher Unterricht und Lehrerbildung im Umbruch?* (S. 230-233). GDCP-Tagungsband
9. Heinitz, B., Nehring, A. (2020). Naturwissenschaftsspezifische Unterrichtsqualität - ein systematisches Review im Spiegel der Ziele, Inhalte und Methoden der naturwissenschaftlichen Unterrichtsfächer In S. Habig (Hrsg.) *Naturwissenschaftliche Kompetenzen in der Gesellschaft von morgen* (S. 202 – 205). GDCP-Tagungsband.
10. Heinitz, B., Nehring, A. (2019). Facetten von Kompetenzorientierung auf Stundenebene. In C. Maurer (Hrsg.) *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe* (S. 520 – 523). GDCP-Tagungsband.

**In Vorbereitung**

Heinitz, B. & Nehring, A. (submitted) Perception and Interpretation of Cognitive Activation in Science Pre-Service Teacher Education – The Necessity of Establishing a Common Understanding