

Augmented Reality, haptisch-interaktiv oder illustrativ?

Entwicklung und Anwendung von Qualitätskriterien zum Vergleich von Modelltypen
und deren Einfluss auf kognitiv-affektive Lernprozesse im Anfangsunterricht Chemie

von

Tobias Bier

geboren in Ottweiler

Angenommene Dissertation zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

Fachbereich Natur- und Umweltwissenschaften

Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau

Berichterstatter:

Herr Prof. Dr. Björn Risch, RPTU Kaiserslautern-Landau

Frau Prof. Dr. Sandra Nitz, RPTU Kaiserslautern-Landau

Herr Prof. Dr. Markus Emden, PH Zürich

Tag der Disputation 11.04.2025

Zusammenfassung

Die Anwendung von Modellen ermöglicht das Verständnis komplexer Realitäten, die Vorhersage zukünftiger Entwicklungen sowie die Entwicklung effektiver Lösungsstrategien in Wissenschaft, Wirtschaft und Politik. Im naturwissenschaftlichen Unterricht sollen Schüler*innen befähigt werden, mithilfe abstrakter Modelle Phänomene zu beobachten und zu beschreiben, Hypothesen zu bilden und zu überprüfen sowie die Methoden der Erkenntnisgewinnung zu reflektieren. Im Fach Chemie stützen Teilchen- und Atommodelle das Verständnis für die Wechselbeziehung zwischen Stoff- und Teilchenebene. Der spiralcurriculare Wechsel zwischen verschiedenen Atom- und Bindungsmodellen trägt jedoch dazu bei, dass Lernende neue Modelle als verbesserte Wahrheit ansehen, statt deren parallele Existenz und die Trennung zwischen Modellvorstellung und Realität zu verstehen. Zudem ist der Übergang von der makroskopischen zur submikroskopischen Ebene für viele Lernende schwer nachvollziehbar. Klassische Anschauungsmodelle und Abbildungen mildern diese Schwierigkeiten nur bedingt, da sie oft als realitätsgetreue Abbilder missverstanden werden. Digitale Animationen bergen jedoch das Potenzial, die Verwechslungsgefahr von Modell und Realität zu verringern.

Im Rahmen einer Interventionsstudie wird der Einfluss unterschiedlicher Modelltypen auf kognitiv-affektive Lernendenvariablen untersucht. Schüler*innen der achten Klassenstufe führen Experimente zu Lösungs- und Diffusionsprozessen durch und nutzen anschließend zu deren Deutung einen von drei Modelltypen: illustrative (IL) Modelle, haptisch-interaktive (HI) Modelle und Augmented Reality (AR)-Modelle. In einer ersten Studie werden Kriterien zur Bewertung der Qualität von Anschauungsmodellen entwickelt und empirisch geprüft. Eine zweite Studie evaluiert die drei Modelltypen anhand dieser Kriterien mittels eines Expert*innenratings. Diese vorgestellte Bewertung der Modellqualität gewährleistet eine fundierte und vergleichbare Basis für die nachfolgende Untersuchung der Modelleffekte und stellt die interne Validität der Untersuchung sicher. In der abschließenden Interventionsstudie wird die Wirksamkeit der unterschiedlichen Modelltypen zum situationalen Interesse, Fachwissen sowie zum Modellverständnis der Schüler*innen untersucht. Die Ergebnisse zeigen, dass AR-Modelle und IL-Modelle zu einer signifikanten Verbesserung des Modellverständnisses führen, während HI-Modelle keine signifikante Steigerung bewirken. In allen Gruppen konnte eine signifikante Zunahme des Fachwissens beobachtet werden. Für das situationale Interesse der Schüler*innen zeigte sich kein signifikanter Unterschied zwischen den Modelltypen. Die konzipierten Experimente, Modelle und Lernumgebungen erweisen sich aufgrund ihrer hervorragenden Werte für die affektiv-kognitiven Kriterien als geeignete bildungspraktische Implikationen. Die Ergebnisse bieten wichtige Implikationen für die Gestaltung naturwissenschaftlichen Unterrichts, die Entwicklung kognitiv und affektiv förderlicher Modelle sowie Grundlagen für weiterführende Forschung zum Einfluss externer Repräsentationen auf das Modellverständnis.

Abstract

Models facilitate the understanding of complex realities, the prediction of future developments, and the formulation of effective solution strategies in science, economics, and politics. In science education, students should be empowered to observe and describe phenomena using abstract models, formulate and test hypotheses, and reflect on methods of knowledge acquisition. In chemistry, particle and atomic models support the understanding of the interrelationship between substance and particle levels. However, the spiral curricular transition between various atomic and bonding models contributes to learners viewing new models as improved truths, rather than understanding their parallel existence and the distinction between model conception and reality. Furthermore, the transition from the macroscopic to the submicroscopic level is often challenging for many learners to comprehend. Traditional visual models and illustrations only partially mitigate these difficulties, as they are often misinterpreted as accurate representations of reality. Digital animations, however, hold the potential to reduce the risk of confusing models with reality.

The influence of different model types on the cognitive-affective learner variables is investigated in an intervention study. Eighth grade students carry out experiments on solution and diffusion processes and then use one of three model types to interpret them: illustrative models, haptic-interactive models and augmented reality (AR) models. In a first study, criteria for evaluating the quality of illustrative models are developed and empirically tested. In a second study, the three model types are evaluated based on the previously defined criteria using an expert rating. The preceding assessment of model quality ensures a sound and comparable basis for the subsequent investigation of model effects. In addition, compliance with design and quality criteria during the conception of the different interventions ensures the internal validity of the study. The concluding intervention study examines how the different types of models affect students' situational interest, subject knowledge, and their understanding of the nature of models. The results show that the use of AR models or illustrative models leads to a significant improvement in model understanding, while haptic-interactive models do not lead to a significant increase. Subject knowledge increased significantly in all groups to the same extent. There was no significant difference between the model types for students' situational interest. The investigation's results provide important implications for the design of science education, the development of cognitively and affectively supportive models, and the foundations for further research on the influence of external representations on the understanding of the nature of models.

Inhaltsverzeichnis

1. Ausgangslage und Problemstellung	1
2. Theoretische Rahmung	5
2.1. Modelle und deren Funktionen	5
2.2. Klassifikation von Modellen in den Naturwissenschaften	23
2.3. Modellkompetenz und Modellverständnis im Kontext der Modellmethode	28
2.4. Modelltypabhängige Lernendenvariablen	46
3. Forschungsstand zu Modelltypen in Medienvergleichsstudien	62
3.1. Medienvergleichsstudien	62
3.2. Der Einfluss des Modelltyps auf den Lernprozess	64
4. Ableitung der Forschungsfragen und Hypothesen	71
5. Konzeption der Lernumgebung	75
5.1. Kriterien zur Konzeption der Lernmaterialien	75
5.2. Formulierung und Einordnung der Lernziele	81
5.3. Konzeption des Aufgabenmaterials	85
5.4. Ablauf der Erarbeitungsphase	86
5.5. Konzeption der Experimente	88
5.6. Konzeption der Modelle	92
6. Studie I – Entwicklung des Messinstruments „BeQuAn“ zur Bewertung der Qualität von Anschauungsmodellen	100
6.1. Formulierung der Items	100
6.2. Validierung des Messinstruments	102
6.3. Reliabilitätsbestimmung des Messinstruments	111
6.4. Implikationen aus Studie I für die weiteren Untersuchungen	113
7. Studie II Gütebestimmung der Qualität der drei Modelltypen haptisch-interaktiv, AR und IL	114
7.1. Methoden der Datenerhebung	114
7.2. Methoden der Datenauswertung	115
7.3. Ergebnisse	117
7.4. Diskussion zu Studie II	122
8. Studie III – Unterschiede in der Wirksamkeit der drei Modelltypen	124
8.1. Pilotierungsstudie	124
8.1.1. Testkonstruktion	126
8.1.2. Evaluation der Lernmaterialien und -umgebungen	127
8.2. Analyse der eingesetzten Messinstrumente	139
8.3. Interventionsstudie	149
8.3.1. Methoden der Datenerhebung	149
8.3.2. Fragebogenerhebung	150
8.3.3. Methoden der Datenauswertung	153

8.3.4.	Ergebnisse	155
8.3.5.	Diskussion zur Interventionsstudie	170
9.	Fazit und Ausblick	181
10.	Verzeichnisse	184
10.1.	Abbildungsverzeichnis	184
10.2.	Tabellenverzeichnis	186
10.3.	Literaturverzeichnis	189
11.	Anhang	214
11.1	Lernmaterialien	214
11.2	Fragebögen	214
11.3	Ergänzende methodische Inhalte	214
11.4	Darstellung des Eigenanteils und Eidesstattliche Erklärung	214
11.5	Danksagung	215

1. Ausgangslage und Problemstellung

(0) Die Bedeutung von Modellen

Der fachgerechte Umgang mit Modellen ist in diversen Lebensbereichen von entscheidender Bedeutung. Sei es in der Wissenschaft, der Wirtschaft oder der Politik. In Zeiten wachsender gesellschaftlicher Krisen erlangen Modelle eine besondere Bedeutung. Sie helfen dabei, die komplexe Realität zu verstehen, zukünftige Entwicklungen vorherzusagen und effektive Lösungsstrategien für Probleme zu entwickeln. Aufgrund ihres abstrakten und vereinfachenden Charakters weisen Modellierungen jedoch inhärente Unsicherheiten und Einschränkungen auf. Ihre eingeschränkte Interpretierbarkeit erfordert daher im Umgang mit den Modellen die Kompetenz, mögliche Unsicherheiten in der Interpretation zu berücksichtigen. Laut dem Beschluss der Kultusministerkonferenz soll der naturwissenschaftliche Unterricht Schüler*innen zum souveränen Umgang mit Modellen und Modellierungen befähigen (Sekretariat der Ständigen Konferenz der Kultusminister [SSKK], 2020). Dazu zählen folgende Kompetenzen:

- *„Phänomene der Natur, der Technik und des Alltags aus naturwissenschaftlicher Perspektive zu beobachten, mithilfe zunehmend abstrakter und komplexer Modelle zu beschreiben und naturwissenschaftliche Fragestellungen aus diesen abzuleiten*
- *Hypothesen zu bilden, diese zum Beispiel durch systematisches Beobachten, Experimente, Modellen, Simulationen bzw. theoretische Überlegungen zu prüfen [...]*
- *die Methoden der Erkenntnisgewinnung wie zum Beispiel systematische Beobachtungen, Experimente und Modelle in den Naturwissenschaften zu reflektieren und die Vor- und Nachteile sowie die Grenzen dieser Methoden zu bewerten.“* (SSKK, 2020, S. 9)

Die Vertiefung und Vernetzung der vorhandenen Kompetenzen der Lernenden und deren Erweiterung um eine Metaperspektive auf die Denk- und Arbeitsweisen der Naturwissenschaften wird dabei als der Kern naturwissenschaftlicher Kompetenz herausgestellt (SSKK, 2020).

(1) Modelle als zentrale Arbeitsmethode und Mittel zur Erkenntnisgewinnung

Modelle zählen zu den wichtigsten Arbeits- und Hilfsmitteln in den Naturwissenschaften. Speziell für den naturwissenschaftlichen Unterricht gilt der Einsatz von Modellen als eine der zentralen Methoden der Erkenntnisvermittlung und Erkenntnisgewinnung (Eschenhagen, Kattmann, & Rodi, 2008; Lehrer & Schauble, 2015). Für das Fach Chemie sind dabei das kognitive und praktische Arbeiten mit Teilchen- bzw. Atommodellen kennzeichnend ([KMK] K., 2005). Die Eigenschaften eines Stoffs werden dadurch erklärt, dass dieser aus bestimmten Teilchen aufgebaut ist. Diesen Teilchen werden wiederum eigene spezielle Merkmale zugeschrieben. Die Charakteristika der Stoffe ergeben sich also grundsätzlich aus den Eigenschaften der

Teilchenkollektive, die den Stoff in der jeweiligen Form ausmachen. Dadurch stellt die Betrachtungsweise der Phänomene eine durchweg wechselseitige Beziehung von Stoff- und Teilchenebene dar. Dies gilt als ein Hauptcharakteristikum der Denkweisen in der Chemie (Christen, 1990).

(2) Die Probleme im Umgang mit Modellen erfordern ein etabliertes Modellverständnis

Besonders jüngere Schüler*innen mit mäßiger Abstraktionsfähigkeit weisen bei diesen Wechselprozessen Schwierigkeiten auf (Christen, 1990). Wie in den Bildungsstandards festgelegt, werden die Schüler*innen entlang eines Spiralcurriculums über mehrere Modelle mit steigender Komplexität geführt (MBWWK, 2014). So können kumulativ mehr Phänomene erklärbar gemacht und ein immer tieferes Verständnis für die Basiskonzepte generiert werden (ebd.). Dieser spiralcurriculare Wechsel der unterschiedlichen Atom- und Bindungsmodelle hat häufig zur Folge, dass bei Schüler*innen eine parallele Existenz der verschiedenen Modellvorstellungen verhindert wird (Saborowski, 2000). Lernende fassen die neu auftretenden Modelle lediglich als eine verbesserte Form der Wahrheit auf (ebd.). Die Trennung von Modellvorstellungen und Realität sowie das Wissen um die parallele Existenz mehrerer Modelle sind Kernkompetenzen eines adäquaten Modellverständnisses (z.B. Leisner, 2005a). Insbesondere der Aspekt der beschränkten Wiedergabe der Realität und die damit verbundenen Grenzen der Modellvorstellungen sollen dabei im Fokus des Chemieunterrichts stehen (Christen, 1990). Oftmals führen die spiralcurricular induzierten Konflikte bei den Schüler*innen zu Verwirrung und Frustration (Saborowski, 2000; Baumert et al. 2013; Blankenburg, Broman & Parchmann, 2016). Auch haben Untersuchungen zum Fachinteresse gezeigt, dass Schüler*innen mit besonders großen Erwartungen in den Chemieunterricht starten, dieses Interesse jedoch kontinuierlich und im Verhältnis zu den restlichen Schulfächern stärker abnimmt (Elster, 2005; Elster, 2007; Hoffmann & Lehrke, 1986). Probleme entstehen vor allem dann, wenn die Deutung von Phänomenen nicht mehr auf der Stoffebene, sondern auf der Teilchenebene erfolgt (Fischler & Lichtfeld, 1997).

(3) Die Einführung des Denkmodells „Teilchenmodell“ als kritische Schlüsselstelle im Chemieunterricht

Der thematische Einstieg in das Teilchenmodell als erste Modellvorstellung stellt einen entscheidenden Wendepunkt hinsichtlich des Interesses am und des Verständnisses im Chemieunterricht dar (Saborowski, 2000). Studien haben gezeigt, dass der gedankliche Übergang von der makroskopischen Ebene hin zu den Modellvorstellungen über das submikroskopische Diskontinuum für Lernende schwer nachzuvollziehen ist (Özmen, 2011a; Kirman Bilgin, Demircioğlu Yürükel, & Yiğit, 2017). Zudem haben Lernende über alle Altersstufen hinweg Probleme beim Verstehen und der Anwendung des Teilchenmodells (Andersson, 1990; Christen, 1990;

Aydeniz & Kotowski, 2012; Yan & Talanquer, 2015; Kirman Bilgin et al. 2017). Auch wenn diese Probleme schon länger bekannt sind, wurden sie bisher noch nicht zufriedenstellend gelöst und werden noch weiterhin in der naturwissenschaftsdidaktischen Forschung untersucht (Parchmann & Schwarzer, 2016; Bittorf, Hallier, Busch & Sieve, 2017; Sieve, Graulich, Caspari & Bittorf, 2017; Thomas, Struckmeier & Sieve, 2017). Anschauungsmodelle sind das klassische Hilfsmittel, um Schüler*innen den Übergang zu einem adäquaten Verständnis der Teilchenwelt zu erleichtern. Daneben gilt das Arbeiten mit diversen Arten von Modellen und Modelltypen als ein grundlegendes Anliegen des Chemieunterrichts (Keller, 1977; Christen, 1990). Solche Visualisierungen zur Teilchenebene konnten bereits als effektive Hilfsmittel zum Erlangen von Fachwissen zum Teilchenmodell evaluiert werden (Gabel, 1999; Williamson & José, 2009; Sanger, 2009; Williamson, 2011; Williamson, 2015). Die Schüler*innen weisen jedoch auch nach dem Gebrauch diverser Anschauungsmodelle weiterhin große Defizite insbesondere in ihren Vorstellungen zum Teilchenmodell auf (Aydeniz & Kotowski, 2012; Özmen, 2011a; Kirman Bilgin et al., 2017). Eine mögliche Ursache hierfür könnte im Modelltyp der Anschauungsmodelle liegen (Saborowski, 2000). In der Schulpraxis kommen zumeist Abbildungen bzw. Bildfolgen in Schulbüchern oder gegenständliche Modelle, wie zum Beispiel Kugel-Stab-Modelle, zum Einsatz (Härtig et al., 2021; Ostermann, Härtig, Kampschulte, Lindmeier, Ropohl & Schwanewedel, 2019). Diese Modelltypen vereint das Problem, dass sie als Idealisierung von realen Teilchen gesehen oder mit Vergrößerungen davon verwechselt werden (Saborowski, 2000). Es besteht keine Trennschärfe zwischen Realität und Modell, sodass die Schüler*innen die Modellvorstellungen als reale Sachverhalte verstehen (Jung, 1977; Harrison & Treagust, 1996). Laut Saborowski (2000) bergen digitale Animationen das Potential, die Gefahr der Verwechslung von Modell und Realität zu verringern (Saborowski, 2000). Durch Animationen werde es möglich, die Wirklichkeit zu simulieren und aufgrund dieser Imitation von real Erfahrbarem den Bruch mit der alltäglichen Realität zu vermeiden (ebd.). Darüber hinaus konnten bereits Vorteile von animationsgestützten, selbstgesteuerten Conceptual-Change-Lerneinheiten im Vergleich zu lehrerzentriertem Frontalunterricht festgestellt werden (Özmen, 2011b). Diese betreffen vornehmlich eine Verbesserung in den Bereichen Fachwissen sowie Vorstellungsänderungen zum Thema Modelle in der Chemie (ebd.). Bislang gibt es jedoch noch keine Forschung zur Auswirkung der Modelltypen auf komplexe Konstrukte, wie das Modellverständnis im Kontext makroskopischer Phänomene. Die Ursache hierfür liegt vor allem darin begründet, dass es bislang nur wenige AR-Animationen und haptische Modelle gibt, die solche Prozesse veranschaulichen. Insbesondere im Bereich der haptischen(-interaktiven) Modelle herrscht hier eine große Lücke. Daher stellt insbesondere der Vergleich von haptisch-interaktiven Modellen mit AR-Modellen ein neues Beobachtungsfeld dar.

Hier setzt das Forschungsprojekt an. Das übergeordnete Ziel besteht darin, im Rahmen einer Interventionsstudie den Einfluss von unterschiedlichen Modelltypen auf kognitiv-affektive Lernendenvariablen zu untersuchen. Das zentrale Konzept sieht vor, dass Schüler*innen Experimente zu Lösungs- und Diffusionsprozessen durchführen. Die Erarbeitung der Versuchsdeutung auf Teilchenebene erfolgt im Anschluss mithilfe von drei unterschiedlichen Modelltypen: Illustrative Modelle, haptisch-interaktive Modelle und Augmented Reality (AR) Modelle. Da adäquate Vorlagen für Schüler*innen-Experimente, AR- oder haptisch-interaktive Modelle nicht verfügbar sind, werden zunächst Experimente und Modelle neu konzipiert und weiterentwickelt.

Zur Gewährleistung der Vergleichbarkeit der Modellgüte der Modelltypen werden in Studie I Kriterien zur Bewertung der Qualität von Anschauungsmodellen deduziert und empirisch geprüft (Przywarra et al., 2021). Anhand der Ergebnisse aus Studie I sollen die Kriterien von Eschenhagen et al. (2008) sowie von Rauh und Stickel (1997) in eine valide Kurzskala zur Beurteilung der Modellqualität von Anschauungsmodellen mittels Expert*inneninterviews und anschließender qualitativer Inhaltsanalyse überführt werden.

In Studie II werden die Modelle der verschiedenen Modelltypen mithilfe des entwickelten Instruments einem Expert*innenrating unterzogen. Die Ergebnisse dienen der Gewährleistung einer gleichwertigen Qualität der entwickelten Modelle der drei verglichenen Modelltypen.

Im Anschluss werden in Studie III die Experimente und Modelle kriteriengeleitet für die Durchführung einer Unterrichtseinheit auf Basis der Conceptual Change Theorie evaluiert und optimiert. In der abschließenden Interventionsstudie wird der Frage nachgegangen wie sich die unterschiedlichen Modelltypen auf das *situationale Interesse*, das *Fachwissen* zu Löse- und Diffusionsprozessen und das *Modellverständnis* von Schüler*innen der achten Klassenstufe auswirken. Die Intervention umfasst drei Doppelstunden, in deren Verlauf die Schüler*innen jeweils zwei Experimente durchführen. Experimentalgruppe I (EG I) erarbeitet sich die Deutungen mit haptisch-interaktiven Modellen, Experimentalgruppe II (EG II) mit AR-Modellen und die Experimentalgruppe III (EG III) verwendet Legebilder mit Illustrationen.

Im Rahmen der vorliegenden Untersuchung erfolgt sowohl eine Analyse der Entwicklung der Proband*innen als auch ein Vergleich der drei Experimentalgruppen hinsichtlich der genannten abhängigen Variablen. Als Kontrollvariablen werden die Konstrukte *Fachinteresse Chemie*, *Fähigkeitsselbstkonzept*, *Cognitive Load*, *System Usability* und die *kognitiven Fähigkeiten* begleitend erhoben.

2. Theoretische Rahmung

2.1 Modelle und deren Funktionen

- Es gibt keine einheitliche Definition des Modellbegriffs. Modelle werden durch Abbildungs-, Verkürzungs- und Subjektivierungsmerkmale charakterisiert und sind stets zweckorientiert.
- In den Naturwissenschaften werden Modelle stets als Repräsentationen mit denkökonomischen Vorteilen verstanden.
- Modellierungen und der Modelleinsatz sind zweckorientiert und durch die Eigenschaften der Repräsentation und des Originals bedingt.
- Wissenschaftlich-realistische Perspektiven zum Modellbegriff basieren auf dem Vergleich von Modell und Original; - Konstruktivistisch-empirische Positionen fokussieren die Intentionen von Modellierenden und Modellnutzenden.
- Naturwissenschaftsdidaktische Modelle werden in hypothetisch-realistische und direkt-realistische Modelle unterteilt, wobei Teilchendarstellungen und Modellverständnis den hypothetisch-realistischen Modellen zugeordnet werden.
- Mentale Modelle stellen persönliche, interne Repräsentationen dar, die sich von etablierten naturwissenschaftlichen Modellen unterscheiden können.
- Anschauungsmodelle dienen als Werkzeuge zur Erkenntnisgewinnung und -vermittlung in den Naturwissenschaften.
- Gütekriterien für Anschauungsmodelle sind Fachliche Korrektheit, Redundanzfreiheit, Vollständigkeit, Übersichtlichkeit und Verständlichkeit.
- Das Teilchenmodell für die Sekundarstufe I postuliert, dass Materie aus unsichtbaren, kleinsten Teilchen besteht, die in ständiger Bewegung sind, elastische Zusammenstöße haben und bei Temperaturänderungen ihre mittlere kinetische Energie verändern.

Modelle kommen in den unterschiedlichsten Formen und in den verschiedensten Kontexten vor. Sie spielen dabei in vielen Lebensbereichen eine zentrale Rolle und nehmen nicht zuletzt in den Wissenschaften eine Schlüsselposition ein (Dirks & Knobloch, 2008). Die Vielfalt und Vielzahl der existierenden Modelle und den unterschiedlichen Wissenschaftsfeldern ist kaum zu überschauen (ebd.). Dabei nehmen vor allem Modellrechnungen, Visualisierungen und Simulationen eine immer bedeutsamere Stellung ein, insbesondere wenn es um Betrachtungen komplexer Phänomene und Systeme geht (ebd.).

Der Modellbegriff

Der Begriff "Modell" wird in vielen verschiedenen Kontexten verwendet und kann auf verschiedene Arten interpretiert werden. In Abhängig der unterschiedlichen Kontexte, Zusammenhänge und Situationen ist in der Verwendung des Modellbegriffs keine gemeinsame Basis festzustellen (Dirks & Knobloch, 2008). Daraus ergeben sich wiederum zahlreiche Definitionen für den Modellbegriff. Den internationalen Diskurs um die didaktische Betrachtung von Modellen resümieren Oh & Oh (2011, S. 1112) damit, dass der Modellbegriff nicht eindeutig definierbar sei. Die Vielzahl von Modellbegriffen und Definitionen bringt auch eine Vielzahl von Modellmerkmalen mit sich, anhand derer Modelle charakterisiert werden können. Stachowiak

(1973) hat in einer empirischen Analyse zur generischen Verwendung des Modellbegriffs drei grundsätzliche Merkmale herausgearbeitet:

- **Abbildungsmerkmal:** Modelle sind Repräsentationen von bestimmten natürlichen oder künstlichen Originalen.
- **Verkürzungsmerkmal:** Modelle decken nie alle Charakteristika ihres repräsentierten Ursprungssystems ab. Es werden nur die Kennzeichen des Originals abgebildet, die dem Modellierer und den Nutzern der Modelle als relevant erscheinen.
- **Subjektivierungsmerkmal:** Beim Erstellen eines Modells wird ein bestimmter Zweck verfolgt. Der Modellierer versucht Aussagen über die Realität zu gewinnen, ohne in vollem Umfang über die Realität zu verfügen. Die Funktionen werden weiterhin durch die gedanklichen oder tatsächlichen Handlungen und innerhalb der zugeordneten Zeitspannen eingeschränkt.

In der (naturwissenschaftsdidaktischen) Forschung wird der Modellbegriff zunehmend aus einer epistemologischen Perspektive betrachtet (Gilbert & Justi, 2016; Mahr, 2015; Passmore et al., 2014). Ein Objekt wird als Modell aufgefasst, wenn es von einem Subjekt als solches entwickelt (Ritchey, 2012), genutzt (Giere, 2010) oder aufgefasst wird (Mahr, 2015). Mahr (2015) bietet eine konsequent epistemologische Perspektive auf den Modellbegriff als Begründung dafür, ein Objekt als Modell zu definieren (s. Abbildung 1). Indem er Bedingungen formuliert, unter denen ein Objekt als ein Modell verstanden werden kann, löst Mahr (2008a) das Ambiguitätsproblem des Modellbegriffs auf (Digel, 2020). Mit diesen Kriterien für das Modellsein umgeht Mahr (2008a) die Frage „Was ist ein Modell?“. Stattdessen bildet sich ein subjektbezogenes Urteil über das Modellsein basierend auf der Erscheinungsform, der Herstellungs- und Anwendungsperspektive. In einem ersten kognitiven Schritt werden die Objektmerkmale, die Erkenntnisse über das Objekt und die Auffassungen zum Objekt selektiert sowie generalisiert und eine neue Repräsentation induktiv gebildet (= Herstellungsperspektive). In einem zweiten Schritt wird der Inhalt der Repräsentation im Zuge einer Deduktion zweckgerichtet auf einen Anwendungsbereich oder einen anderen Gegenstand übertragen (= Anwendungsperspektive). Weiterhin unterscheidet Mahr (2008a) zwischen einem gedachten, subjektbezogenen Modell und einem weitgehend gegenständlichen, externen Modellobjekt, das dieses Modell repräsentiert und vom Subjekt gebildet wird. Die Beziehungen zwischen dem Modellobjekt und dem Modell bzw. dem repräsentierten Objekt sowie zwischen dem Modellobjekt und dem modellbildenden und -nutzenden Subjekt, gelten als die zentralen Aspekte des Modellseins. Je nach Kontext der Modellnutzung und -auffassung des Subjekts kann das Modellobjekt sowohl „Modell von etwas“ (Abbild) als auch „Modell für etwas“ (Vorbild) sein

(Gouvea & Passmore, 2017; Mahr, 2015). Die Entscheidung über dieses Urteil des Modellseins, basiert auf den konstruktiven Beziehungen zwischen der Schaffung eines Modells und seiner Anwendung. Aus der Herstellungsperspektive gilt ein Modell als Abbild während es aus der Anwendungsperspektive als Vorbild dient (Mahr, 2015). Ein „Modell von etwas“ ist dabei das Resultat des Induktionsprozesses, bei dem die Merkmale und Regelinhalte des Ausgangsobjekts in einer neuen Repräsentationsform, dem Modellobjekt, als Inhalt des Modells fungieren (Mahr, 2008a). Von einem „Modell für etwas“ spricht Mahr (2008a), wenn das externe Modellobjekt als Medium in einer deduktiven Vermittlungssituation dient. In beiden Fällen dient das Modell dem Transport von Inhalten, dem „Cargo“ (Mahr, 2008b; 2015). Dieser Cargo beinhaltet die Schnittmenge der Anwendungs- und der Herstellungsperspektive des Modells (Mahr, 2008b).

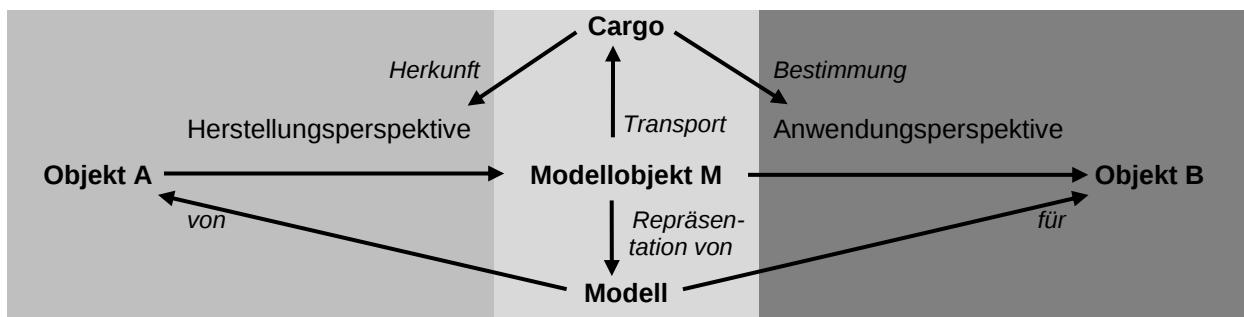


Abbildung 1: Urteil über das Modellsein (verändert nach Mahr, 2008a; Upmeier zu Belzen & Krüger, 2010)

Der naturwissenschaftliche Modellbegriff

In den Naturwissenschaften gelten Modelle als die zentralen Arbeits- & Hilfsmittel (Krüger et al., 2018). Da Modelle hier mehrere Funktionen erfüllen, werden sie auch stets aus verschiedenen Blickwinkeln betrachtet. Bislang gibt es auch für den naturwissenschaftlichen Modellbegriff keine eindeutige Definition (Oh & Oh, 2011; Gehlen 2016). Zudem existieren keine speziellen Merkmale und Charakteristika, mit deren Hilfe sich ein Modell eindeutig als solches definieren lässt (Upmeier zu Belzen & Krüger, 2010). Mittelstraß (2005, S. 65) versucht eine allgemeine Rahmung zu liefern: „Modelle sind Nachbildungen eines realen oder imaginären Gegenstandes mit dem Ziel, etwas über diesen oder mit diesem zu lernen“. Auch wenn über die Bedeutung von Modellen ein allgemeiner Konsens herrscht, sorgen Versuche einer einheitlichen Definition des Modellbegriff stets für Debatten (Gehlen, 2016). Dies hat eine erschwerte Charakterisierung des naturwissenschaftlichen Modellbegriffs zur Folge. Als Gemeinsamkeit für die Verwendung des Modellbegriffs in den Naturwissenschaften, heben Oh & Oh (2011) die Nutzung des Begriffs Repräsentation hervor. Ein Modell wird als Repräsentation aufgefasst, die Teilgebiete naturwissenschaftlicher Theorien abbildet (Schwarz &

Gwekwerere, 2007) sowie als System von Objekten, das ein anderes System repräsentiert (Gilbert, 2004). Weiterhin haben alle Perspektiven auf Modelle gemeinsam, dass Modelle auf gewissen Modellvorstellungen beruhen, die vor allem aus denkökonomischen Gründen heraus entstanden sind (Steinbuch, 1977). Da sie stets Vereinfachungen der Realität sind, ermöglichen Modelle es Sachverhalte leichter zu erfassen und Probleme einfacher zu lösen (Eschenhagen et al., 2008). Darüber hinaus ist allen Modellen gemein, dass diejenigen Aspekte der Realität, die in das Modell mit einfließen und der Zweck des Modells sich gegenseitig bedingen (Keller, 1977). Man spricht hierbei vom zweifach finalen Charakter der Modelle (ebd.). Gleichzeitig erfolgen sowohl die Modellierung als auch der Einsatz von Modellen stets zweckorientiert (Eschenhagen, Kattmann & Rodi, 2008). In den Naturwissenschaften gelten Modelle dabei als theoretische Rekonstruktionen der Realität und dienen der Erkenntnisgewinnung und Erkenntnisvermittlung (Eschenhagen, Kattmann, & Rodi, 2008). Die Mehrheit der Autoren im naturwissenschaftlichen Bereich versuchen den Begriff zu skizzieren, indem sie entweder die Eigenschaften der Modelle aufzählen oder einen Verwendungszweck definieren (Keller, 1977). Diesbezüglich haben sich zwei unterschiedliche Perspektiven auf den Modellbegriff entwickelt - wissenschaftlicher Realismus und konstruktiver Empirismus (Fischer et al., 1995). Annäherungen an den Modellbegriff auf Basis des Vergleichs zwischen Modell und Original werden als wissenschaftlich-realistische Sichtweise bezeichnet. Nahezu alle wissenschaftlich-realistischen Definitionen stimmen darin überein, dass Modelle stets als Gegenbegriff zu den Originalen stehen und als vereinfachte Darstellung dieser angesehen werden (Fischer, Grube & Reisin, 1995; Eschenhagen, Kattmann & Rodi, 2008; Lehnert & Köhler, 2012). So definieren Lehnert & Köhler (2012) Modelle als vereinfachte Abbilder einer komplexeren Realität. Diese entsprechen weitestgehend dem Original, unterscheiden sich aber in mindestens einer der folgenden Eigenschaften:

- dem Stoff aus dem die Modelle aufgebaut sind
- der Dimension (größer oder kleiner Dargestellt)
- dem Abstraktionsgrad (Unwichtiges auslassen und Wichtiges hervorheben)

Neben der Fokussierung auf Unterschiede zum Original beschreiben van Driel & Verloop (1999) sieben grundsätzliche Eigenschaften für alle naturwissenschaftlichen Modelle:

1. Modelle stehen stets in Relation zu den Originalen, die von diesen Modellen repräsentiert werden. Unter Originalen werden hierbei Phänomene, Objekte, Prozesse oder Systeme verstanden.
2. Modelle dienen als wissenschaftliche Hilfsmittel, um Informationen über Originale zu erschließen, die weder direkt beobachtbar noch messbar sind. - Demnach gilt ein Modell, das mit den Sinnen erfahrbar ist und didaktisch reduziert dargestellt wurde im engeren Sinne nicht als wissenschaftliches Modell - (Beerenwinkel & Parchmann, 2008).
3. Modelle können nicht unmittelbar mit den Originalen, die durch sie repräsentiert werden, interagieren.
4. Modelle weisen zu ihren Originalen bestimmte Analogien auf, die das Testen von Hypothesen über das unzugängliche Original ermöglichen. Dadurch lassen sich neue Informationen über das Original erschließen. Daher nehmen Modelle eine Schlüsselposition für die Erkenntnisgewinnung ein (z.B. Justi & Gilbert, 2002; Wellnitz & Mayer, 2012; Reith & Nehring, 2020).
5. Modelle und ihre Originale unterscheiden sich stets in spezifischen Punkten voneinander. Dabei gelten Modelle im Vergleich zum Original generell als vereinfacht. Neben bestimmten Aspekten, die weggelassen werden, besitzen Modelle auch gewisse Eigenschaften, die ihre Originale nicht aufweisen.
6. Während der Konzeption der Modelle entsteht ein Kompromiss zwischen den Unterschieden und den Analogien zum Original. Dieser Kompromiss wird stets in Abhängigkeit vom Kontext für jede Situation neu geschlossen.
7. Die Konzeption von Modellen zeichnet sich als iterativer Prozess aus. Können am Original empirisch gewonnene Daten nicht mithilfe des Modells erklärt werden, bedarf es einer Anpassung des Modells. Das angepasste Modell wird anschließend im Zuge weiterer Untersuchungen am Original auf seine Güte hin untersucht.

Um die Merkmale von Modellen nach Stachowiak (1973) gänzlich zu erfüllen, müssen naturwissenschaftliche Modelle laut Eschenhagen, Kattmann & Rodi (2008) folgende Kriterien erfüllen:

- Das Kriterium der Ähnlichkeit und Entsprechung setzt voraus, dass das Modell in wesentlichen Eigenschaften mit dem Original übereinstimmt. Daher sind die Modelle den Originalen in den Hauptmerkmalen ähnlich.
- Das Prinzip der Einfachheit und Adäquatheit stellt einen Kompromiss zwischen diesen beiden Aspekten dar. Das Modell soll zum einen die wesentlichen Merkmale des Originals abbilden. Zum anderen soll das Modell einfacher als das Original sein.
- Unter dem Kriterium der Exaktheit und Fruchtbarkeit wird verstanden, dass das Modell so exakt sein soll, um unter bestimmten Bedingungen Prognosen über das Original zu ermöglichen.

Upmeier zu Belzen und Krüger (2010) wählen die Kriterien von Rischer, Grube und Reisin (1995) als wissenschaftlich-realistischen Orientierungsrahmen zur weiteren Charakterisierung von Modellen im naturwissenschaftlichen Unterricht (Tabelle 1). Zunächst wird eine Vergleichbarkeit von Modell und Original vorausgesetzt. Die Güte des Modells wird hierbei in Relation zu einem objektiv existierenden Original bewertet. Das reale Original existiert dabei in Unabhängigkeit von seiner Beziehung zum Modell, wodurch der Vergleich von Original und Modell erst sinnvoll wird. Das Modell repräsentiert das Original überwiegend so, wie es in der Realität existiert. Im Hinblick auf die Kriterien Korrektheit und Vollständigkeit kann das Modell möglicherweise fehlerhaft und unvollständig sein. Im Falle einer Revision des Modells hat dies keine Auswirkungen auf die Vorstellungen über das Original. Darüber hinaus lässt sich unabhängig vom Modellzweck beurteilen, ob das Original vom Modell korrekt und vollständig abgebildet wird. Diese Herangehensweise ist aus erkenntnistheoretischer Sicht problematisch (Upmeier zu Belzen & Krüger, 2010), da Modelle als Mittel zur Erkenntnisgewinnung stets nur anhand des Zwecks und der Zielgruppe bewertbar sind (Giere, 2010). Daher werden Modelle speziell im naturwissenschaftsdidaktischen Kontext aus der Perspektive des empirischen Konstruktivismus betrachtet (Fischer et al., 1995). Dieses Konzept vertritt den Ansatz, dass Modelle adäquat zu Erfahrungen mit beobachtbaren Phänomenen aufzufassen sind (Giere, 2010). Aufgrund fehlender Relevanz bzw. Konsequenzen einer Entscheidung über die Beziehung zwischen Modell und Original, nehmen empirisch konstruktivistische Ansätze dazu keine Beurteilung vor (Upmeier zu Belzen & Krüger, 2010). Upmeier zu Belzen & Krüger (2010) leiten daraus unterschiedliche Konsequenzen, sowohl für die Interpretation der Herstellungs- und Anwendungsperspektive als auch für die Interpretation der Kompatibilität von Modellen mit

empirischen Daten ab. Anstatt der Annahme einer generellen Ähnlichkeit zwischen Original und Modell ist das subjektive Ähnlichkeitskonstrukt des urteilenden Subjekts als die Repräsentation eines Interpretationsprozesses zu verstehen (Giere, 2010; Upmeyer zu Belzen & Krüger, 2010). Unter Einnahme dieser Perspektive können die Kriterien Vergleichbarkeit und Unabhängigkeit nicht vorausgesetzt werden (Upmeyer zu Belzen & Krüger, 2010). Modell und Original sind dann voneinander abhängig und sich beiderseitig konstituierend (Fischer et al., 1995). Die Modellierung umfasst hierbei gleichermaßen die Konstruktion eines wissenschaftlichen Denkmodells zu einem Original sowie die Konstruktion eines mentalen Modells über das Original (Fischer et al., 1995; Giere, 2010). In diesem Kontext wird nun die Angemessenheit des Modells zur Erfüllung seines Zwecks zum Kriterium (Fischer et al., 1995). Dadurch wird es erforderlich sowohl das Verkürzungsmerkmal als auch das Subjektivierungsmerkmal in Bezug auf das Original sowie die Angemessenheit der Vorstellungen zum Original zu hinterfragen und gegebenenfalls zu revidieren (Fischer et al., 1995). Modelle sind daher nicht als objektive Repräsentationen der Realität anzusehen, sondern als Hilfsmittel um erfolgreich mentale Modelle über die Wirklichkeit zu generieren (Upmeyer zu Belzen & Krüger, 2010).

Tabelle 1: Kriterien und Konsequenzen zum Verhältnis zwischen Modell und Original nach Fischer et al., 1995, verändert nach Upmeyer zu Belzen & Krüger, 2010, S.41

Wissenschaftsphilosophischer Ansatz		
<u>Kriterium</u>	<u>Wissenschaftlicher Realismus</u>	<u>Konstruktiver Empirismus</u>
Vergleichbarkeit	vorausgesetzt	irrelevant
Unabhängigkeit	vorausgesetzt	Nicht vorausgesetzt, potenziell abhängig
Korrektheit	entscheidbar	Nicht entscheidbar
Vollständigkeit	Entscheidbar	Nicht entscheidbar
Angemessenheit	Nur auf das Modell bezogen	Auf das Modell und das vorgestellte Original bezogen
Konsequenzen		
Konstruktion	Modell wird verändert, bis Güte erreicht ist	Modell und Original verändern sich beide im Urteil des Modellierers
Güte des Modells	Güte ist festgelegt und orientiert sich am Original	Interessen, Ziele und Vorstellungen des Modellierenden entscheiden über Qualität des Modells, nicht das Original
Bedeutung	Status des Originals	Status des Modellierenden

Erkenntnistheoretische Funktionen von Modellen

Da der Modellbegriff durch das Subjektivierungsmerkmal und seinen doppelt finalen Charakter gekennzeichnet ist, bringen Modelle auch immer mehrere Funktionen mit sich. Innerhalb dieser Arbeit wird sich auf die Funktionen im Bereich der Anwendung als Lehr- Lernmodelle beschränkt.

Mit der Begründung, dass jede Form der Erkenntnis stets das Produkt einer Modellierung sei, konstatiert die Modelltheorie nach Stachowiak (1973) eine weitumfassende erkenntnistheoretische Position zum Modellbegriff. Der erkenntnistheoretische Ansatz beschäftigt sich mit der Frage, wie man zu Erkenntnis bzw. Wissen über die Welt gelangen kann bzw. welche Rahmen- und Gelingensbedingungen für eine Erkenntnis gegeben sein müssen (Baumann, 2006). Rehfus (2003) formuliert als Rahmung zwei gegensätzliche Strömungen. Die antirealistische Sichtweise geht davon aus, dass die Gegenstände der Beobachtung überhaupt nur durch diese Beobachtung existieren. Realistische Sichtweisen gehen davon aus, dass die Gegenstände in Unabhängigkeit von der Beobachtung existieren. Aufgrund der Prämisse, dass Naturwissenschaften sich mit der Beschreibung der realen Welt beschäftigen, sind für naturwissenschaftliche Betrachtungen nur realistische Positionen relevant. Ausgehend vom Gesichtspunkt, dass die Beobachtungsgegenstände unmittelbar bzw. direkt wahrgenommen werden können, unterscheidet man zwischen naivem (hypothetischem) und direktem (kritischem) Realismus. Nimmt man an, dass der real existierende Gegenstand nicht vollkommen direkt wahrnehmbar ist, sondern durch geistige Gegenstände, wie Ideen oder Theorien erweitert werden muss, wird eine hypothetisch-realistische Position eingenommen. Auf Basis der erkenntnistheoretischen Interessensgegenstände von sowohl Erkenntnisgewinnung als auch der Erkenntnis an sich, sind sowohl der Zugang zur Erkenntnis über die Eigenschaften des Originals als auch die des Modells von Interesse. Da Anschauungsmodelle direkt erfahrbare sind, die dahinterstehenden Theorien und Konzepte jedoch häufig nur indirekt, ist es notwendig sowohl direkt- als auch hypothetisch-realistische Modelle zum Umgang mit Modellen zu betrachten. Steinbuch (1977) formuliert auf Basis der kritisch-rationalistischen Erkenntnisposition Stachowiaks (1973) ein direkt-realistisches Modell zur Erkenntnis durch die Modellmethode in den Naturwissenschaften. Laut diesem Modell wird ein Original als komplexer Sachverhalt aus der Realität in einem abstrakten Denkmodell abgebildet. Für dieses Denkmodell werden nur für den jeweiligen Zusammenhang relevante Informationen genutzt. Diese Informationen („das Wesentliche“) werden über einen Filter vom Rest der Informationen getrennt. Weiterhin wird das Denkmodell um weitere, notwendige Informationen, wie zum Beispiel Naturgesetze, ergänzt. Damit steht dem Bewusstsein ein erweitertes Denkmodell für künftige Erkenntnisprozesse zur Verfügung. Durch die Herstellung eines konkreten Anschauungsmodells, lässt sich das Modell erneut in die Realität übertragen (s. Abbildung 2).

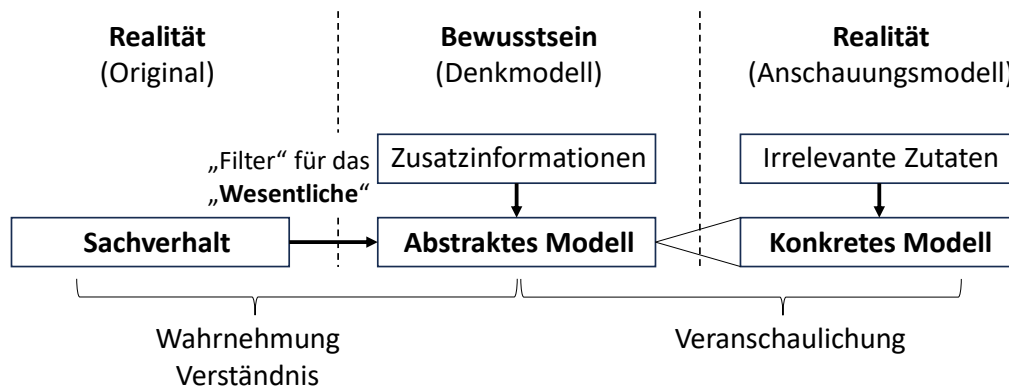


Abbildung 2: Schema zum "Denken in Modellen" (nach Steinbuch, 1977)

Hypothetisch-realistische Positionen zum Modellverständnis (Leisner, 2005a) werden durch die hypothetische Natur des Modellbegriffs als basales Hilfsmittel zur Erkenntnisgewinnung und Theorienentwicklung begründet (Abel, 2008). Ein indirekt realistisches Modell zum Umgang mit Modellen wurde von Kircher (1977, 1995) entwickelt (s. Abbildung 3). Im Folgenden wird es am Beispiel eines Kochsalz-Kristalls und den zugehörigen Modellen erläutert:

- Das Original *O* ist wird von einem natürlichen Steinsalzkrystal mit gut ausgeprägter Würfelform, geraden Kanten, glatten Flächen und rechten Winkeln repräsentiert.
- Das gewählte Modell stellt ein Raumgitter-Modell dar, in dem Natrium-Ionen durch die kleinen, blauen und die Chlorid-Ionen durch die großen, grünen Kugeln dargestellt werden.
- Dem/der Lernenden als Subjekt *S* wird nun ein neuer Zugang zu Informationen über das Original *O* durch das Modell *M* bereitgestellt. Somit dient das Modell *M* als Vermittler zwischen dem Original *O* und dem Subjekt *S*.

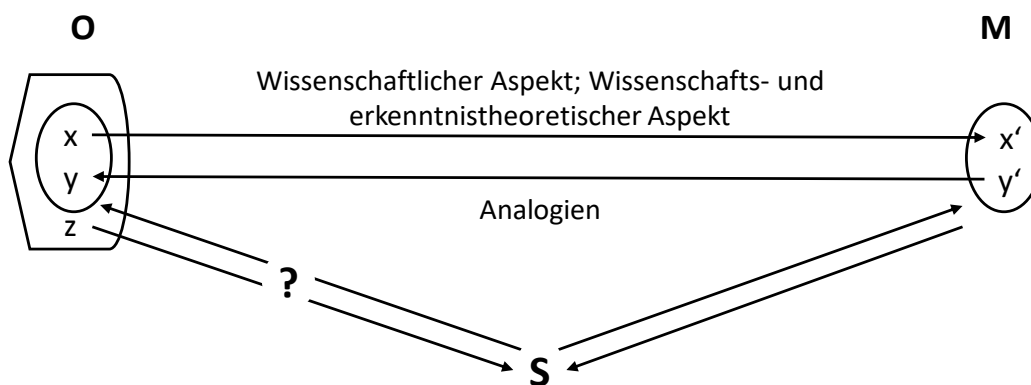


Abbildung 3: Schema der Modell-Original-Subjekt-Relation (nach Kircher, 1995)

Den Eigenschaften *x* und *y* des Originals werden im Modell entsprechende Modelleigenschaften *x'* und *y'* zugeordnet. Sei *x* die räumliche Struktur der Ionen im Kristall, so wäre *x'* die entsprechende Anordnung der Kugeln im Modell. Die Eigenschaft *y* sei das Verhältnis der

Radien der beiden Ionenarten im Kochsalzkristall. Dann ist y' das entsprechende Größenverhältnis der Kugeln im Raumgitter-Modell. Somit sind x und y die durch das Modell dargestellten Parameter. Stachowiak (1973) bezeichnet diese Parameter als Abbildungsmerkmale und Steinbuch (1977) als das Wesentliche des Modells. Es gibt auch Eigenschaften z , zu denen es kein entsprechendes Merkmal im Modell gibt. In Bezug auf das Kochsalzbeispiel wären dies zum Beispiel der salzige Geschmack oder die Wasserlöslichkeit. In diesen Eigenschaften spiegelt sich das Verkürzungsmerkmal wider (Stachowiak 1973). Je nach Intention zum Einsatz des Modells ist auch zumeist nicht gewünscht, alle Informationen in das Modell zu übertragen. Weiterhin kann es Eigenschaften geben, die zwar im Modell aber nicht im Original zu finden sind. Ein Beispiel hierfür wären die Stäbe im Kristallgittermodell. Diese Eigenschaften sind laut Steinbuch (1977) irrelevante Zutaten. Im Modell selbst sind die Merkmale als Informationen jedoch an die Repräsentation in Form des Modellinhalts gebunden (Mahr, 2008a).

Saborowski (2000, S.65ff) erweitert das M-O-S-Schema nach Kircher (1995) um die Perspektiven Steinbuchs (1977) durch eine Unterteilung in Modell-, Realitäts- und Subjektbereich (vgl. Abbildung 4). Der Realitätsbereich umfasst dabei neben dem Original (O) noch das Experiment (E) und das Anschauungsmodell als Methoden. Der Modell- und Subjektbereich spiegeln das Bewusstsein im Schema Steinbuchs wider. Die Besonderheit an der Perspektive Saborowskis ist, dass sie sich auf chemische Modellvorstellungen zur Struktur der Materie bezieht und es daher keine Analogien zwischen dem Original und dem Anschauungsmodell (AM) gibt. Da das Original nicht direkt vom Subjekt (S) erfahrbare ist, müssen theoretische Modellannahmen (M) getroffen werden, die Analogien zum Original aufweisen. Im Zuge des Erkenntnisprozesses definiert das, aus den Modellannahmen resultierende, Denkmodell die Planung eines Experiments. Das Subjekt nimmt die konkrete Durchführung des Experiments vor. Das Ergebnis des Experiments liefert Informationen über das Original. Diese können durch Messinstrumente etc. meist direkt erhoben werden, jedoch nur unter Zuhilfenahme des Denkmodells interpretiert werden. Zwischen dem Denkmodell und dem Anschauungsmodell liegen Analogien vor. Um die meist sehr abstrakten Gesichtspunkte des Modellbereichs besser nachvollziehen zu können, kann das Subjekt direkt auf das Anschauungsmodell als Hilfsmittel zurückgreifen.

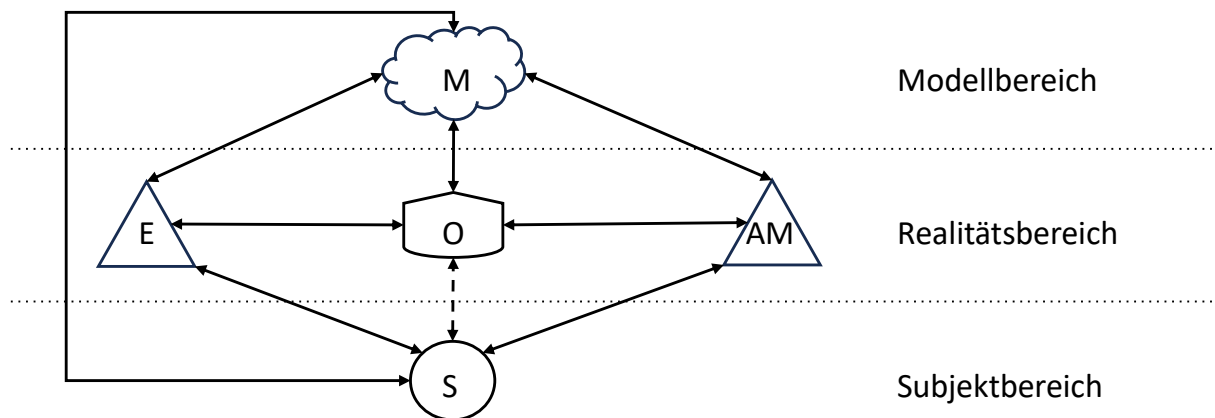


Abbildung 4: erweitertes M-O-S-Schema zu Modellen zu Struktur der Materie (nach Saborowski, 2000)

Während Saborowski (2000), Mahr (2008a) und Kircher (1995, 2015) den Modellbegriff aus rein wissenschafts- und erkenntnistheoretischen Blickwinkeln betrachten, betonen kognitionspsychologische Ansätze für den Zugang zu empirischen Phänomenen noch die Komponente der mentalen Modelle, die dem Modellierenden als Ausgangspunkt gedient haben (Digel, 2020; Terzer, 2013).

Mentale Modelle

Anders als die übergeordnete Perspektive mit Fokus auf die Beziehungen zwischen Subjekt, Objekt und Modell fokussiert sich dieser kognitionspsychologische Ansatz auf die Wahrnehmung der realen Welt (=Objekte) sowie der kognitiven Verarbeitung durch das Subjekt (Digel, 2020). Erstmals wurde diese Perspektive in der Theorie der mentalen Modelle beschrieben (Johnson-Laird, 1980). Der Begriff Modelle wird dort als eine mentale Repräsentation verstanden. Während etablierte naturwissenschaftliche Modelle aufgrund ihrer Kohärenz zu gesichertem naturwissenschaftlichem Wissen als vollständige und präzise Repräsentationen gelten (Norman, 2014), stellen mentale Modelle als persönliche, interne Repräsentationen dar (Greca & Moreira, 2000). Ähnlich wie beim Modellbegriff gibt es auch für den Begriff mentale Modelle keine einheitliche Konzeptualisierung (Greca & Moreira, 2000). Es gibt zwei unterschiedlich motivierte, theoretische Strömungen: Der „theoretical approach“ nach Johnson-Laird (1983) und der „instructional approach“ nach Gentner und Stevens (2014). Johnson-Laird (1983) sieht mentale Modelle als rein situative, im Arbeitsgedächtnis vorläufig vorliegende, ikonische Repräsentationen auf Basis abstrakter Elemente. Diese Repräsentationen grenzen sich von rein bildlichen und zweidimensionalen, mentalen Bildern und Propositionen in Form von Inferenzregeln unterliegenden Zeichenfolgen ab. Mentale Modelle unterscheiden sich sowohl in ihrer Struktur als auch in ihrer Dynamik: Sie sind ikonisch zur repräsentierten Struktur, stehen

repräsentativ für eine reelle Möglichkeit und gelten für das Subjekt als wahr bzw. als konsistent darin, Phänomene zuverlässig abzubilden. Stets ausgehend von einer Problemsituation verläuft die Ausbildung mentaler Modelle über drei Phasen (Held et al., 2006). In der Modellkonstruktion wird ein mentales Modell auf Basis der gegebenen Informationen gebildet. Im Zuge der Modellinspektion werden Informationslücken in der Problembeschreibung identifiziert und daraus Schlussfolgerungen gezogen. Die Modellvariation prüft die Gültigkeit dieser Schlussfolgerungen für die jeweilige Situation und berücksichtigt dabei alternative mentale Modelle. Während beim „theoretical approach“ logisches Denken und die Kompetenz zu Diskursen in konkreten Problemlöseanwendungen im Vordergrund stehen (Johnson-Laird, 2005), setzt der „instructional approach“ allgemeiner bei der Bildung von Wissensstrukturen und den mentalen Prozessen zum Verständnis komplexer Domänen an (Nitz & Fechner, 2018). Darüber hinaus unterscheiden sich die beiden Ansätze in ihrem Verständnis über den Charakter des vorliegenden Wissens. Gentner & Stevens (2014) bezeichnen mentale Modelle als stabile Strukturen, die aktiviert und genutzt werden können. Hier entstehen die mentalen Modelle durch Analogiebildung zu einem bereits bekannten und strukturell ähnlichen Sachverhalt unter Adaption der konzeptuellen Struktur (Gentner, 1983). Mentale Manipulationen am mentalen Modell ermöglichen es dem Subjekt Vorhersagen und Erklärungen zu treffen (Greca & Moreira, 2000). Greca und Moreira (2002) nutzen diesen Ausgangspunkt, um den Prozess der naturwissenschaftlichen Erkenntnisgewinnung auf den Verstehensprozess einer naturwissenschaftlichen Theorie abzubilden. Beim Verstehen eines Phänomens beeinflusst ein konzeptuelles Denkmodell, dem phänomenologisch und konzeptuell eine naturwissenschaftliche Theorie zugrunde liegt, das persönliche, mentale Modell. Im Anschluss wird dieses mentale Modell durch ein mathematisches Modell um die abstrakten, syntaktischen Strukturen der Theorie erweitert. Die Beeinflussung der mentalen Modelle durch Konfrontation des Subjekts mit dem konzeptuellen Modell führt nicht zwangsläufig zu einer identischen, internalen Repräsentation dieses konzeptuellen Modells (Greca & Moreira, 2000). Den Lernenden fällt es schwer, das konzeptuelle Modell als solches zu erkennen. Zudem fehlt es ihnen an ausreichendem Fachwissen und Modellverständnis, um das Modell zu interpretieren und den Modellcharakter und seine Implikationen deuten zu können (ebd.). Daher ist das mentale Modell auch nicht unbedingt konsistent zum konzeptuellen Bezugsmodell (ebd.). Die Charakterisierung von mentalen Modellen stimmt darin mit dem naturwissenschaftsdidaktischen Verständnis von *Vorstellungen* überein (Nitz & Fechner, 2018). So charakterisieren Gropengießer und Marohn (2018, S. 51) Vorstellungen als „subjektiv“ gebildete, „gedankliche Konstrukte“ eines Individuums. Auch im Modell des Modellseins berücksichtigt Mahr (2015) das Subjekt sowohl bei der Unterscheidung zwischen mentalem Modell und Modellobjekt als auch bei deren Entstehung und Anwendung mit.

Modelle und Erkenntnisgewinnung

Modelle zählen zu den wichtigsten Arbeits- und Hilfsmitteln in den Naturwissenschaften. Der Einsatz von Modellen im naturwissenschaftlichen Unterricht gilt als eine der zentralen Methoden der Erkenntnisvermittlung und Erkenntnisgewinnung (Eschenhagen, Kattmann, & Rodi, 2008; Lehrer & Schauble, 2015). Damit kommt dem Zusammenhang zwischen der Modellmethode und dem Kompetenzbereich Erkenntnisgewinnung eine besondere Bedeutung zu. Als zweckmäßig konstruierte Hilfsmittel erfüllen Modelle vorrangig zwei wichtige Funktionen. Aufgrund ihrer Anschauungsfunktion (Terzer, 2008) fungieren Modelle als Medien zur Kommunikation anerkannten Wissens, um zuvor beobachtete Phänomene durch Modelle zu erklären. Weiterhin dienen Modelle als Untersuchungsinstrumente und kommen als Forschungswerkzeuge zum Einsatz, die es auf Grundlage der Modellinhalte ermöglichen, Schlussfolgerungen in Bezug auf das Original abzuleiten (Crawford & Cullin, 2005; Krell et al., 2014). Anhand dieser beiden Funktionen ermöglichen Modelle es auch, Prognosen zu Phänomenen zu erstellen (Krell et al., 2014; Upmeyer zu Belzen & Krüger, 2010), die experimentell überprüft werden können (Christen, 1990). Zudem werden geeignete Modelle genutzt, um experimentell gewonnene Ergebnisse zu interpretieren oder Hypothesen zu prüfen (SSKK, 2020). Darüber hinaus werden Modellen noch weitere Funktionen zugeschrieben (Eschenhagen et al. 2008; Lehnert & Köhler, 2012; Nerdel, 2017). Als vereinfachte Abbilder ermöglichen Modelle es Sachverhalte leichter zu erfassen und erleichtern das Lösen von Problemstellungen (Eschenhagen et al., 2008). Diese Erleichterung des Verständnisses durch eine strukturelle Reduktion wird als denkökonomische Funktion der Modelle bezeichnet (ebd.). Denkmodelle erfüllen dabei den Zweck Phänomene bzw. deren Parameter miteinander logisch zu verknüpfen und sie damit zu erklären (Christen, 1990). Weiterhin ermöglichen Modelle durch ihren hypothetischen und vorläufigen Charakter den Prozess der Problemfindung und -eingrenzung (Eschenhagen et al., 2008). Dies wird als die heuristische Funktion von Modellen bezeichnet (ebd.).

Laut Köhler & Meisert (2012) besteht ein Ziel der naturwissenschaftlichen Erkenntnisgewinnung darin Modelle zu entwickeln, die Phänomene erklärbar machen. Weiterhin sollen die Modelle mit den messbaren Variablen des Phänomens und anderen bestehenden Modellen vereinbar sein. Ausgehend von einer Fragestellung werden hierzu hypothetische Modelle entwickelt und im Anschluss einer systematischen Überprüfung unterzogen. Die Überprüfung beginnt in der Regel mit dem Treffen von Vorhersagen, die mit erhobenen Daten verglichen werden. Sind Daten und Vorhersagen miteinander vereinbar, liegt ein Indiz für die Richtigkeit des Modells vor. Bei einem Widerspruch von Daten und Vorhersagen, muss das Modell entsprechend weiterentwickelt werden. Zur Entwicklung von Modellen im Unterricht empfehlen sich folgende vier Punkte:

1. Modelle entwickeln
2. Aus Modellen Vorhersagen ableiten
3. Modelle prüfen
4. Modelle weiterentwickeln.

Jeder dieser vier Schritte kann als sinnvolles Lernangebot dazu dienen Kompetenzen naturwissenschaftlicher Theorieentwicklung zu entwickeln oder zu fördern (Meisert, 2014).

Auf Basis der Modellperspektiven Mahrs (2008; 2015) entwickelten Krell et al. (2016) ein Prozessschema der naturwissenschaftlichen Erkenntnisgewinnung im Kontext der Modellmethode. Auf der Ebene der Herstellungsperspektive dient zum Beispiel ein beobachtetes Phänomen als Ausgangspunkt des Erkenntnisprozesses. Die Lernenden bilden unter Einbeziehung von Kontext und Nutzen des Modellgegenstands sowie des vorhanden Vorwissens ein initiales mentales Modell (Clement, 2000). Ausgehend von diesem initialen Modell und in Abhängigkeit der vorliegenden Präkonzepte bzw. des Vorwissens prüfen die Lernenden die Passung der ihnen bereits bekannten, kohärenten Modelle (Krüger et al., 2018). Im Anschluss generieren die Lernenden gegebenenfalls neue Verknüpfungen und Modellelemente (ebd.). Dieser Prozess wird zudem durch eine Prüfung auf Passung und Kohärenz zum Phänomen begleitet. Abschließend erfolgt die Externalisierung eines oder mehrerer Modelle. Liegen diese Modelle als Externalisierung vor, geht der Prozess in die Anwendungsperspektive über. Anhand der externalisierten Modelle können Hypothesen zu einer konkreten Fragestellung generiert werden, die sich wiederum experimentell überprüfen lassen. Passen die Hypothesen zu den Daten aus dem Experiment, führt dies über die Verifizierung des Modells zu einem Erkenntnisgewinn in (ebd.).

Gütekriterien für Modelle

Für naturwissenschaftsdidaktische Anschauungsmodelle gilt der Vorsatz, dass diese stets nur so gut sind, wie ihr Einsatz im Lernprozess und die daran anschließende Modellkritik (Eschenhagen, Kattmann & Rodi, 2008). Folglich gibt es in der naturwissenschaftsdidaktischen Literatur keine klar beschriebenen Gütekriterien, anhand derer sich Anschauungsmodelle bewerten lassen. Eschenhagen, Kattmann und Rodi (2008) haben allgemeine Anforderungen an naturwissenschaftsdidaktische Modelle formuliert, die zur Qualitätsbestimmung herangezogen werden können:

- Das Kriterium der **Ähnlichkeit und Entsprechung** setzt voraus, dass das Modell in wesentlichen Eigenschaften mit dem Original übereinstimmt. Daher sind die Modelle den Originalen in den Hauptmerkmalen ähnlich.
- Das Prinzip der **Einfachheit und Adäquatheit** stellt einen Kompromiss zwischen diesen beiden Aspekten dar. Das Modell soll zum einen die wesentlichen Merkmale des Originals abbilden. Zum anderen soll das Modell einfacher als das Original sein.
- Unter dem Kriterium der **Exaktheit und Fruchtbarkeit** wird verstanden, dass das Modell so exakt sein soll, um unter bestimmten Bedingungen Prognosen über das Original zu ermöglichen.

Aufgrund fehlender Kriterien bedienen sich Autoren in den Naturwissenschaftsdidaktiken für Annäherungen an den Modellbegriff häufig an Konzepten aus mehr oder weniger verwandten Fachrichtungen. So wählen Upmeier zu Belzen & Krüger (2010) zur Einordnung des Modellbegriffs den Ansatz des theoretischen Informatikers Bernd Mahr. Auf Basis der Unterscheidung zwischen „ein Modell von etwas“ und „ein Modell für etwas sein“ formuliert Mahr (2008) drei allgemeine Kriterien zur Bewertung Modellen. Zunächst muss das Modell seine Funktion als Medium erfüllen. Das bedeutet, es muss alle Informationen über das Original, wofür es ein Modell ist, zugänglich machen, die benötigt werden damit das Modell seinen Zweck erfüllen kann. Als zweites Kriterium definiert Mahr (2008), dass das Modell „ein Garant von Konsistenz sein“ soll. Mahr versteht darunter, dass das Modell keine inhaltlichen oder logischen Widersprüche enthält. Dadurch soll vermieden werden, dass die Anwendung des Modells zwangsläufig zu Widersprüchen führt. Ferner soll das Modell über eine ausreichende pragmatische Eignung verfügen. Demnach müsse das Modellobjekt, wovon oder wofür das Modell steht, angemessen und ausreichend repräsentiert werden.

Weiterhin liefert die Unterscheidung zwischen wissenschaftlich-realistischer und konstruktiv-empirischer Position der Informatiker Fischer, Gruber und Reisin (1995) Kriterien (vgl. Tabelle 1), anhand derer sich Modelle bewerten lassen (Upmeier zu Belzen & Krüger, 2010). Der

konstruktiv-empirische Ansatz dieses Modells setzt den Fokus auf Modelle als Mittel zur Erkenntnisgewinnung (Fischer et al., 1995) und ist daher zur Bewertung von Erkenntnismodellen vorzuziehen. Das Konstrukt der Modellkompetenz beinhaltet jedoch neben den Dimensionen, der Erkenntnisgewinnung, auch in eine Fachwissensdimension (vgl. Tabelle 33). Da diese Fachwissenskomponente der Anschauungsmodelle eindeutig als (un-)korrekt bzw. (un-)vollständig bezeichnet werden kann, ist die wissenschaftlich-realistische Betrachtung von Anschauungsmodellen nach Fischer et al. (1995) im Kontext der Modellkompetenz vorzuziehen. Eine Einnahme beider Perspektiven ist demnach unbedingt erforderlich, um die Modellqualität beurteilen zu können. Darüber hinaus sollte eine Beurteilung jeglicher Kriterien der Modellgüte lediglich in Abhängigkeit vom intendierten Zweck des Modells erfolgen (Giere, 2010).

Die Wirtschaftsinformatiker Rauh & Stickel (1997) haben Gütekriterien zur Bewertung der Modellqualität formuliert, die auf die Beurteilung von naturwissenschaftsdidaktischen Anschauungsmodellen übertragen werden können:

- Das erste Kriterium für eine gute Modellqualität ist das Kriterium der **Korrektheit**. Dieses ist wiederum in die semantische und die syntaktische Korrektheit unterteilt. Unter syntaktische Korrektheit fällt die korrekte (grammatikalische) Anordnung der passenden Begrifflichkeiten. Die semantische Korrektheit umfasst die inhaltliche Korrektheit des abgebildeten Sachverhalts. Sie ist folglich dann gegeben, wenn keine inhaltlichen Fehler bei der Darstellung des Phänomens gemacht wurden. Zu umfangreiche oder unzulässige Beziehungen zwischen den einzelnen Elementen des Modells reduzieren ebenfalls die semantische Korrektheit.
- Ein weiteres Gütekriterium für Modelle ist die **Redundanzfreiheit**. Darunter wird verstanden, dass das Modell keine Informationen enthält, die zum Erfüllen des Zwecks des Modells überflüssig sind. Folglich sind überflüssige Komponenten und Informationen innerhalb des Modells zur Einhaltung der Redundanzfreiheit zu vermeiden.
- Eine weitere Bedingung für ein gutes Modell ist das Erfüllen des Kriteriums der **Vollständigkeit**. Das Modell muss hierbei alle Informationen beinhalten, die zum Erreichen des Ziels benötigt werden. Ist ein Aspekt des Modells überpräsent führt dies zu einer verringerten **Übersichtlichkeit** und es liegt ein Verstoß gegen das Prinzip der Genügsamkeit vor. Jedes Merkmal des Modells darf also nur so stark ausgeprägt sein, dass das Modell noch übersichtlich ist.
- Ist das Modell nicht nur übersichtlich, sondern auch ausdrucksfähig, wird das Kriterium der **Verständlichkeit** erfüllt. Im Vordergrund steht hierbei die Frage, inwiefern das Modell dem Beobachter Informationen über die Eigenschaften und Zusammenhänge der

Modellaspekte vermitteln kann. Ist ein Inhalt überpräsent, hat dies eine verringerte **Übersichtlichkeit** zur Folge und stellt darüber hinaus einen Verstoß gegen das Prinzip der Genügsamkeit bzw. dem Gesetz der Einfachheit (Winnenburg, 1993) dar.

Diskussion der Gütekriterien für Modelle

Die Perspektiven und Kriterien zur Bewertung von Modellen nach Fischer et al. (1995), Mahr (2008), Eschenhagen et al. (2008) sowie Upmeyer zu Belzen & Krüger (2010) basieren allesamt auf normativen Annahmen über die Modellgüte. Die Autoren berichten weder von praktischen Anwendungen zur Bewertung von Modellen noch wurden die Kriterien validiert oder in empirischen Studien erprobt. Dem gegenüber gelten für die Qualitätsmerkmale nach Rauh & Stickel (1997) messbare Werte und Kriterien zur Effektivität der getesteten Modelle in der Erfüllung ihrer Funktionen (Pitschke, 2011). Weiterhin werden diese Kriterien sowohl in der Praxis zur Messung und Vorhersage der Performanz industrieller und wirtschaftlicher Modelle als auch in der empirischen Forschung zur Qualitätsmessung bereits eingesetzt (Rupp, 2007). Zudem zeigen sich die Gütekriterien nach Rauh & Stickel (1997) als kohärent zu den bisher genannten normativen Ansätzen. Folglich können diese Kriterien weitgehend zu einer quantitativen Beurteilung von naturwissenschaftlichen Anschauungsmodellen herangezogen werden. Die syntaktischen Aspekte des Kriteriums Korrektheit finden jedoch keine Berücksichtigung, da Anschauungsmodelle (speziell, die in Studie II und III verwendeten) keine sprachlichen Aspekte beinhalten.

Das undifferenzierte Teilchenmodell in der Chemie

Das Teilchenmodell für die Sekundarstufe I basiert auf den Grundannahmen der kinetischen Gastheorie und ist laut Fischler & Lichtfeld (1997) wie folgt charakterisiert. Bausteine der Materie sind (kleinste) Teilchen, die man einzeln nicht sehen kann. Aussagen über die Form der Teilchen können nicht getätigt werden. Ebenso findet keine Differenzierung zwischen Atomen, Ionen und Molekülen statt. Zur Beschreibung und Interpretation ihres Verhaltens ist diese Angabe jedoch nicht relevant. Zwischen den Teilchen befindet sich aufgrund eines Vakuums absolute Leere. Die Teilchen befinden sich in ständiger Bewegung. Die Zusammenstöße zwischen den Teilchen sind vollkommen elastisch. Dadurch bleibt die mittlere kinetische Energie einer definierten Stoffmenge konstant. Wird die Temperatur der Materie erhöht, bedeutet dies eine Steigerung der mittleren kinetischen Energie der Teilchen. Bei Gasen können die Anziehungskräfte zwischen den Teilchen aufgrund der relativ großen Abstände vernachlässigt werden. Bei Flüssigkeiten sind die Anziehungskräfte stark genug, um ein festes Volumen zu definieren. In Festkörpern sind die Anziehungskräfte zwischen den Teilchen so groß, dass die Teilchen feste Plätze haben, die sie nicht verlassen können. Beim Wechsel der Aggregatzustände wird die zugeführte bzw. abgegebene Energie zum Überwinden bzw. Aufbringen der Anziehungskräfte umgesetzt. Dadurch bleibt die Temperatur der Materie bei diesen Vorgängen unverändert.

2.2 Klassifikation von Modellen in den Naturwissenschaften

Trotz der fundamentalen Bedeutung von Modellen existiert bislang kein allgemeingültiges System zur Klassifikation von Modellen in fächerübergreifenden Diskursen (Mittelstraß, 2005; Krüger et al., 2018). Um Typen von Modellen genauer zu klassifizieren und zu definieren, gibt es in der fachdidaktischen Literatur mehrere Ansätze von Klassifikationssystemen auf Basis unterschiedlicher Typologien (Krüger et al., 2018). Keller (1977) unterteilt Modelle zunächst auf Basis des zutreffenden Erlebnisbereiches. Er grenzt die allgemeine bzw. alltägliche Erlebniswelt von der naturwissenschaftlichen ab. Boulter & Buckley (2000) unterscheiden Modelle einerseits nach dem Modus der Repräsentation (mathematisch, sprachlich, visuell, konkret) und andererseits nach den Merkmalen der Repräsentation. Auf Ebene der Repräsentationsmerkmale wird zwischen deterministischen und probabilistischen, statischen und dynamischen sowie quantitativen und qualitativen Modellen unterschieden. Grob werden dabei die Modi Denkmodelle oder auch abstrakte Modelle von Anschauungsmodellen (auch materielle Sachmodelle, Graf, 2002) getrennt (Eschenhagen, Kattmann, & Rodi, 2008). Denkmodelle sind rein kognitive Konstrukte, die sich direkt aus den vorherrschenden, akzeptierten Modellvorstellungen ergeben (Keller, 1977; Eschenhagen, Kattmann, & Rodi, 2008). Denkmodelle lassen sich wiederum in Modelle anschaulicher Natur (z.B. das Kern-Hülle-Atommodell) und rein abstrakte Konstrukte, die sich dem weltlichen Erfahrungsbereich entziehen (z.B. das quantenchemische Atommodell) unterteilen (Christen, 1990). Der Umgang in sämtlichen naturwissenschaftlichen Disziplinen ist dadurch geprägt, dass die Denkmodelle ständig durch neue Erkenntnisse erweitert oder verändert werden (Barke et al., 2015; Barke & Harsch, 2001a). Die meisten naturwissenschaftlichen Modelle zeichnen sich dadurch aus, dass sie zumindest auf mikroskopischer Ebene mit dem zugehörigen Original verglichen werden können. Da Stoff-Teilchen sich jedoch der Wahrnehmung entziehen, kann zu den chemischen Schlüsselthemen Teilchenmodell und chemische Strukturen kein direkter Vergleich mit dem Original erfolgen. Hier werden Modelle oft auf Basis bereits bekannter Gesetze oder experimenteller Daten gebildet (Caspari et al., 2018). Anschauungsmodelle hingegen finden sich auf der Wahrnehmungsebene wieder (Keller, 1977). Sie werden aus den entsprechenden Denkmodellen heraus erschaffen und sind direkt erfahrbar. Harrison und Treagust (2000) unterscheiden im Kontext der Kontroverse zwischen Abstraktheit und Konkretheit analoge Modelle (theoretisch, mathematisch, ikonisch und symbolisch, Simulation, Konzept-Prozessmodell, Skala, Tabelle, Diagramm, Karte) von synthetischen und mentalen Modellen. Diese Kategorisierungen sind phänomenologischer (Ritchey, 2012) bzw. ontologischer (Oh & Oh, 2011) Natur und bieten ein kriterienbezogenes Ordnungssystem, das sich an der Art der Repräsentation orientiert (Krüger et al., 2018). Die Funktion der Modelle im Prozess der Erkenntnisgewinnung wird dabei jedoch gänzlich außer Acht gelassen (ebd.). Fleige et al. (2012) schließen diese Lücke, indem sie die

Objektperspektive, die Herstellungsperspektive und die Anwendungsperspektive nach Mahr (2008a, 2008b, 2015) auf Modelle als Erkenntnismethode übertragen. Die Objektperspektive legt den Fokus auf das Modellobjekt und seine Beschreibung (z.B. zweidimensionale Darstellungen als Zeichnungen oder Zeichnungen gegenüber dreidimensionalen Gegenständen). Der ursprüngliche Gegenstand, auf dem das Modell beruht, sprich das Original oder die zugrundeliegende Theorie, wird als Ausgangsobjekt (Upmeyer zu Belzen & Krüger, 2010) bezeichnet. Dieses Ausgangsobjekt bleibt aus der Objektperspektive zunächst unbeachtet. Erst die Herstellungsperspektive beschäftigt sich mit dem, was das Modellobjekt darstellt. Das originale Phänomen und das Modell werden in Bezug auf verschiedene Eigenschaften wie Größe, Funktionen, Relationen etc. miteinander verglichen. Diese Informationen sind häufig bekannt und können vom Nutzer reproduktiv aus dem Modell gezogen werden. Auf der Ebene der Anwendungsperspektive kann durch das Modell und mithilfe seiner Nutzung etwas Neues über Phänomene, Strukturen oder Prozesse erfahrbar gemacht werden. Mithilfe des Modells lassen sich gewonnene Vermutungen und Prognosen sowie weiterführende Fragestellungen überprüfen. Diese Perspektive erlaubt einen Transfer zu weiteren Arbeitsweisen, insbesondere dem Experimentieren. In allen Fällen dient das Modell jedoch dem Transport von Wissen und Inhalten. Vereinfacht ausgedrückt, basiert dieser Transport darauf, dass Modelle Repräsentationen von bzw. für etwas sind (Oh & Oh, 2011). Anhand der Art und Weise, wie das Wissen transportiert wird, unterteilt Lindemann (1999) Modelle in:

- 1) Strukturmodelle: Hierzu zählen z.B. Molekülbaukästen und Gittermodelle.
- 2) Funktional ähnliche Modelle: Zu dieser Gruppe gehören zum Beispiel Modellexperimente.
- 3) Ikonische Modelle: Hierzu zählen bildliche Darstellungen aber auch Metaphern wie zum Beispiel die Modellvorstellung der „Elektronenwolken“.
- 4) Symbolische Modelle: Zum Beispiel chemische Formeln oder Reaktionsgleichungen.
- 5) Mathematisch logische Modelle wie zum Beispiel mathematische Formulierungen des chemischen Gleichgewichts.
- 6) Simulationsmodelle wie zum Beispiel Computersimulationen chemischer Abläufe.

Neben diesen Formen von Modellen kann es auch Modelle geben, die in einer Mischform vorliegen. Chemische Strukturformeln können zum Beispiel sowohl ein symbolisches Modell sein als auch in Form von Keilstrichformeln Informationen über den räumlichen Aufbau der Moleküle geben und somit Eigenschaften eines Strukturmodells aufweisen (Lindemann, 1999).

Upmeier zu Belzen (2013) beschreibt weitere Kriterien zur Klassifikation von Modellen in Bezug auf ihre erkenntnisgewinnende Funktion. Hierbei ergeben sich aus verschiedenen Kriterien zur Systematisierung unterschiedliche Modelltypen. Synonym zum Begriff Objektperspektive nach Fleige et al. (2012) verwendet Upmeier zu Belzen (2013) die Bezeichnung „Herstellung der Modelle“ Tabelle 2 gibt einen Überblick über die unterschiedlichen Kategorien zur Klassifikation von Modellen.

Kohärent zum Urteil über das Modellsein nach Mahr (2008a, 2008b; 2015) werden Modelle in der Naturwissenschaftsdidaktik anhand ihres Zwecks in zwei grobe Kategorien eingeteilt (Upmeier zu Belzen und Krüger 2010; Krüger et al. 2018; Krüger & Upmeier zu Belzen, 2021): Modelle mit einem kommunikativen Zweck, um als Medium aktuelles Wissen zu repräsentieren und Modelle als Forschungswerkzeuge zum Zweck der Erkenntnisgewinnung. Werden Modelle als Medien eingesetzt, helfen sie dabei, Strukturen zu beschreiben bzw. zu erkennen sowie deren Beteiligung oder Funktion innerhalb von Prozessen zu erklären (Krüger & Upmeier zu Belzen, 2021). Der Einsatz von Modellen als Forschungswerkzeuge bedeutet darüber hinaus, Voraussagen über das zukünftige Verhalten von Phänomenen unter gewissen Bedingungen herzuleiten und empirisch zu überprüfen (Krüger et al., 2018; Krüger & Upmeier zu Belzen, 2021). Dies ermöglicht es den Lernenden zu neuen Einsichten zu kommen und befähigt sie damit zu neuen Erklärungsansätzen (Krüger et al. 2018). Die Modellierung bringt immer Entscheidungen über die Auswahl gewisser, relevanter Attribute mit sich. Diese betreffen das Verkürzungsmerkmal (Stachowiak, 1973), den Cargo (Mahr, 2008b) oder den Abstraktionsgrad (Lehnert & Köhler, 2012). Folglich unterscheiden sich Modelle im Unterricht nicht nur hinsichtlich der Auswahl relevanter Attribute von ihren Originalen. Vielmehr sind sie in der Regel auch das Ergebnis einer didaktischen Reduktion. Unter Berücksichtigung der Relevanz für die Zielgruppe bezeichnen Eilks und Bindernagel (2008) diesen Ansatz als Curriculum-Modelle. Solche Curriculum-Modelle sollten hinsichtlich des Grades ihrer Vereinfachung, ihrer Komplexität und des Abstraktionsgrades dem Wissensstand der Lernenden angepasst sein. Darüber hinaus sollten sie so gestaltet sein, dass an sie im Sinne einer Weiterentwicklung sowohl fachlich als auch modellhaft erweiterbar sind (De Jong & Taber, 2013).

Tabelle 2: Kriterien zur Klassifikation von Modellen (verändert nach Upmeyer zu Belzen, 2013; erweitert nach Keller, 1977; Christen, 1990; Eschenhagen, Kattmann, & Rodi, 2008; Fleige et al., 2012; Meisert, 2014)

Kriterium zur Systematisierung	Ausprägung des Modells	
Erlebniswelt	Allgemeine Welt	Naturwissenschaftliche Welt
Abstrakter Modellbereich/ Realitätsbereich	Denkmodelle Kognitiver Bereich	Anschauungsmodelle Wahrnehmungsbereich
Objektperspektive/ Herstellung der Modelle	Virtuelle Modelle	Materielle Modelle / Anschauungsmodelle können in verschiedenen Dimensionen dargestellt sein: - Zweidimensional (bildlich) - Dreidimensional (physisch)
Abbildung der Eigenschaften der Anschauungsmodelle	Strukturmodelle veranschaulichen die Merkmale des Aufbaus (z.B. der Anatomie und Morphologie in der Biologie).	Funktionsmodelle weisen Prozesse und Zusammenhänge auf. Sie unterscheiden sich optisch vom Original.
Konstruktionsprozess	Analogmodelle: ein Gegenstand aus der Realität wird zum Original in Beziehung gesetzt und dient somit als Modell.	Homologmodelle: bilden häufig das Originalobjekt in entsprechender Proportion ab.
Zielebene	Nachahmungsmodelle: Originalnahe Vorstellungen sollen phänomenbezogen gefördert werden.	Erkenntnismodell: Verdeutlichung von Funktionsweisen und gesetzmäßigen Beziehungen.
Art der Anwendung/ Position im Erkenntnisprozess	Lehr-Lernmodelle dienen zur Veranschaulichung als Medium; dienen der Erkenntnisgewinnung im Rahmen der naturwissenschaftlichen Arbeitsweise.	Forschungsmodelle dienen im Forschungsprozess als datengenerierendes Medium der Erkenntnisgewinnung.
Anschaulichkeit	Modelle stellen eine konkrete Verbildlichung der Realität dar.	Modelle sind nur von abstrakter Natur. Lassen sich nicht mit real Erfahrbarem erläutern.

Klassifikation der Modelltypen zur Interventionsstudie dieser Arbeit

Auf Basis des Klassifizierungssystem nach Upmeyer zu Belzen (2013) lassen sich die in dieser verwendeten Modelltypen Haptisch-Interaktiv (HI), Augmented-Reality (AR) und Illustrativ (IL) wie folgt charakterisieren. Die für diese Arbeit entwickelten Curriculum-Modelle thematisieren die naturwissenschaftliche Erlebniswelt in Form von Teilchendarstellungen. Sie sind allesamt Anschauungsmodelle, die primär als Funktionsmodelle Prozesse und sekundär Strukturmerkmale auf Teilchenebene veranschaulichen. Da sich die Modelle auf zuvor im Real-Experiment gezeigte Phänomene beziehen, fungieren sie als Analogmodelle und bedürfen einer reflektierten Modellkritik. Als Erkenntnismodelle verdeutlichen sie die Funktionsweisen und gesetzmäßigen Beziehungen von Diffusions- und Löseprozessen auf Teilchenebene und dienen damit als Lehr-Lernmodelle der Erkenntnisgewinnung beim naturwissenschaftlichen Arbeiten. Aufgrund der Behandlung der Teilchenebene ist zu beachten, dass diese Modelle rein abstrakter Natur sind und nicht direkt durch real erfahrbare Phänomene erklärt werden können. Unterschiede zwischen den Modelltypen ergeben sich nur aus der Herstellungs- bzw. Objektperspektive. Während die AR-Modelle als virtuelle 3D-Modelle eingestuft werden, handelt es sich bei den HI-Modellen und IL-Modellen um materielle Modelle mit dreidimensionaler Darstellung.

2.3 Modellkompetenz und Modellverständnis im Kontext der Modellmethode

- Modellkompetenz ist sowohl eine fachliche als auch fachübergreifende Schlüsselkompetenz.
- Internationale Modellkompetenzmodelle gliedern sich in deklarative, epistemologische und prozedurale Dimensionen und unterscheiden sich darin, ob die Nutzung der Modellmethode oder auch das Modellieren mit betrachtet wird:
 - Der Teildimension Modellverständnis als „Nature of Models“ kommt eine übergeordnete Rolle zu und stellt die zentrale epistemologische Komponente dar.
- Alternative Schüler*innenvorstellungen zum Teilchenkonzept liegen zumeist zu den Konzepten Wechsel der Aggregatzustände als Verschwinden, Verlagerung, Wandlung, Transmutation und chemische Interaktion sowie zum inneren Aufbau der Materie als eine Übertragung makroskopischer und Kontinuumseigenschaften auf die Teilchen vor.

Zur Definition der Modellkompetenz wurde für diese Arbeit der, den Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss (KMK, 2005; 2024), sowie den Kompetenzmodellen für den naturwissenschaftlichen Unterricht (Leisner-Bodenthin, 2006; Mayer, 2007; Meisert, 2008; Upmeyer zu Belzen & Krüger, 2010) zugrundeliegende Kompetenzbegriff nach Weinert (2001) verwendet. Weinert (2001, S.27) definiert Kompetenzen als bei Individuen verfügbare oder durch sie erlernbare kognitive und affektive Fertigkeiten und Fähigkeiten bestimmte Probleme zu lösen. In Bezug auf den schulischen Unterricht unterscheidet Weinert (2001) zwischen folgenden Kompetenzen:

- Fachliche Kompetenzen, die zum Beispiel chemischer oder fremdsprachlicher Natur sein können.
- Schlüsselkompetenzen sind fächerübergreifende Kompetenzen wie zum Beispiel Problemlösefähigkeit, Kommunikation oder Selbststeuerung beim Lernen.
- Als Handlungskompetenzen beschreibt Weinert (2001) die damit einhergehende motivationale, volitionale und soziale Bereitschaft, um Probleme in unterschiedlichen Situationen effektiv und verantwortungsbewusst zu bewältigen. Damit sind auch affektive Kompetenzen den Handlungskompetenzen zugehörig.

Die Kernelemente einer Kompetenz, die sowohl Wissen zu den Modellinhalten als auch Kenntnisse über Modelle und deren Limitierung sowie zur Anwendung der Modelle beinhaltet, werden den ersten beiden Arten von Kompetenzen zugeschrieben (Leisner, 2005a). Moralische, emotionale und motivationale Aspekte finden dabei keine Berücksichtigung (ebd.). In Anlehnung daran leiten Klieme et al. (2000) drei Merkmale fächerübergreifender Kompetenzen im Kontext naturwissenschaftlicher Schulfächer ab:

- Funktionalität: Indikator für das Vorliegen einer Kompetenz ist das Bewältigen bestimmter Anforderungen.

- Domänenspezifität: Kompetenzen beziehen sich stets auf einen bestimmten Bereich von Situationen und Kontexten.
- Kompetenzen als Dispositionen: Kompetenzen sind begrenzt verallgemeinerbar – Das Zuschreiben von Kompetenzen geht über den Befund konkreter Einzelleistungen hinaus.

Auf dieser Basis definiert Leisner (2005b) Modellkompetenz als ein System aus Kenntnissen und erlernbaren Fähigkeiten, die zur Disposition der Lernenden führen, um beim Umgang mit naturwissenschaftlichen Modellen den Anforderungen des schulischen Niveaus zu entsprechen.

Andere Ansätze erweitern ihre Definition von Modellkompetenz um Modellierungstätigkeiten (Gehlen, 2016). Laut Upmeyer zu Belzen & Krüger (2010, S.49) umfasst Modellkompetenz daher „[...] die Fähigkeiten, mit Modellen zweckbezogen Erkenntnisse gewinnen zu können und über Modelle mit Bezug auf ihren Zweck urteilen zu können, die Fähigkeiten, über den Prozess der Erkenntnisgewinnung durch Modelle und Modellierungen [...] zu reflektieren sowie die Bereitschaft, diese Fähigkeiten in problemhaltigen Situationen anzuwenden.“

Kompetenzmodelle zum Erkenntnisprozess mit der Modellmethode

Im Hinblick auf die lernendenzentrierte Anwendung der Modellmethode unterscheidet man vorrangig Kompetenzmodelle zum Umgang mit Nachahmungsmodellen von denen zum Arbeiten mit Erkenntnismodellen. In der naturwissenschaftsdidaktischen Forschung existieren bislang lediglich Kompetenzmodelle, die sich explizit mit der Modellkompetenz zu Erkenntnismodellen befassen. Daher wird sich im Rahmen dieser Arbeit ausschließlich auf Kompetenzmodelle zu Erkenntnismodellen bezogen.

Bei der internationalen Betrachtung der Modellkompetenz, lassen sich zunächst Kompetenzmodelle, die Fachwissenskomponenten nicht miteinbeziehen (Mayer, 2007; Upmeyer & Krüger, 2010) von solchen, die Fachwissen als Dimension beinhalten, unterscheiden (Hodson, 1992; Leisner, 2005a; Meisert, 2008; Henze, 2007; Schwarz et al., 2009). Als Grundlage für die Dimensionierung der Kompetenzmodelle inklusive Fachwissensdimension lässt sich eine Unterteilung in eine Fachwissenskomponente und die Komponente der Erkenntnisgewinnung (vgl. Abb. 2) vornehmen (Hodson, 1992; Leisner, 2005a; Leisner-Bodenthin, 2006; Meisert, 2008). Es werden drei Zielebenen eines naturwissenschaftlichen Unterrichts mit Erkenntnismodellen unterschieden (Hodson, 2014). Hodson (1992, S.548f) bezeichnet diese als „learning science“ für das Erwerben naturwissenschaftlichen Fachwissens, „learning about science“ im Sinne eines Eindrucks über die Bedeutung und das Wesen der Naturwissenschaften sowie „doing science“ als das Anwenden von fachspezifischen Arbeitsweisen, Methoden der

Erkenntnisgewinnung und des Problemlösens. Diese Genese kann als Grundlage für eine Präskription der Modellkompetenz genutzt werden. Digel (2020) charakterisiert diese Dimensionen als deklarative, epistemologische und prozedurale Aspekte der Modellkompetenz. Der Beitrag von Modellen und Modellierungsprozessen zu wissenschaftlichen Bildungsfragen deckt alle drei Ziele der naturwissenschaftlichen Bildung nach van Driel und Verloop (2002) ab. Zum einen lassen sich zuvor beobachtete Phänomene durch Modelle erklären - „learning science“. Zum anderen ist es möglich mit ihrer Hilfe Vorhersagen über Ereignisse zu treffen, die im Anschluss experimentell überprüft werden (Christen, 1990) – „doing science“. Weiterhin vermitteln Modelle durch ihren hypothetischen und vorläufigen Charakter (Eschenhagen et al., 2008) wichtige Prinzipien von Nature of Science (Reith & Nehring, 2020). Analog zur Unterteilung von Hodson (1992) nehmen Henze et al. (2007) eine Differenzierung der Modellkompetenz in die Dimensionen „learn scientific models“, „nature of models“ und „act of modelling“ vor. Meisert (2008) unterteilt Modellkompetenz in die drei Teilkompetenzen Modellwissen, Modellverständnis und Modellarbeit. Auch bei Schwarz et al. (2009) findet sich eine Unterteilung in „subject matter expertise“, „epistemological understanding“ und „practice of scientific knowledge“.

Deklarativer Aspekt der Modellkompetenz „learning science“

Nach Meisert (2008) umfasst die deklarative Dimension (Modellwissen) die Kenntnis über grundlegende naturwissenschaftliche Modelle, die im Sinne eines gesicherten Fachwissens den Zugang zu den entsprechenden modellbasierten Konzepten der Fachwissenschaft ermöglicht. Damit ist die Teilkompetenz Modellwissen dem Kompetenzbereich Fachwissen der nationalen Bildungsstandards zuzuordnen (KMK, 2005). Digel (2020) versteht unter dem deklarativen Aspekt der Modellkompetenz vornehmlich konzeptuelles Wissen, das zwar auf Faktenwissen basiert, jedoch darüber hinaus geht. Da Modelle ein Verbund mehrerer Konzepte sind (Gilbert & Justi, 2016), können einzelne Fakten nicht mehr isoliert betrachtet werden, wenn diese als Modelleigenschaften Teil des Modells sind (Digel, 2020). Die Rolle von Modellen fokussiert sich in dieser Teildimension auf etablierte Denkmodelle, die auch als „consensus models“ (Justi & Gilbert, 2002) und „scientific models“ (Henze et al., 2007) bezeichnet werden. Unter dieser ontologischen Perspektive, werden Modelle gewöhnlich als Anschauungsobjekte für fachwissenschaftliches Lernen verstanden (Chiu & Lin, 2019). So sehen Schwarz et al. (2009) Modelle unter dem deklarativen Aspekt als vereinbarte Repräsentationen, die durch Fokussieren auf Schlüsselaspekte ein System abstrahieren und simplifizieren, um Phänomene erklären oder vorhersagen zu können. Jong et al. (2015) rücken den Theoriebezug noch stärker in den Fokus. Sie verstehen Modelle als Konstrukte, die Wissenschaftler zur Formulierung

von Theorien bilden (Jong et al., 2015). Durch Erweiterung ihres Kompetenzmodells um die Perspektive des „Abduktiven Schließens“ schreiben Krüger & Upmeyer zu Belzen (2021) dem Fachwissen eine Bedeutung für die Modellbildung im Zuge der Herstellung eines Modells und der deduktiven Modellanwendung zu. Unter abduktivem Schließen verstehen Krüger und Upmeyer zu Belzen (2021) Phänomene theoriebasiert oder kreativ durch kausale Rückschlüsse bestmöglich zu erklären. Die Herstellung eines Modells bildet beim abduktiv-schließenden Erklären die Basis für einen möglicherweise folgenden deduktiven Überprüfungsprozess (ebd.).

Prozeduraler Aspekt

Die spezifischen Kompetenzen zur Modellanwendung- und Herstellung im Zuge von naturwissenschaftlichen Erkenntnisprozessen werden der prozeduralen Teilkompetenz Modellarbeit zugeordnet. In den Bildungsstandards findet sich diese somit im Kompetenzbereich Erkenntnisgewinnung (Meisert, 2008) wieder. Während Leisner-Bodenthin (2006) sich beim prozeduralen Aspekt auf die Anwendung des Inhalts eines Modells und des Modellverständnisses beschränkt, erweitert Meisert (2008) die prozedurale Komponente in Anlehnung an den „act of modelling“ (Henze et al., 2007) um die Fähigkeiten eigenständig Modelle zu entwickeln und weiterzuentwickeln. Die eigenständige Modellentwicklung weist besonderes Potential für den Lern- und Erkenntnisprozess auf (Meisert, 2008). Durch autonome Modellentwicklungsprozesse erhalten die Lernenden einen konkreten Einblick in eine der zentralen Aktivitäten der wissenschaftlichen Erkenntnisgewinnung (ebd.). Zudem kann die direkte Thematisierung des Erkenntnisprozesses die Entwicklung mentaler Modelle fördern (Tamayo Alzate & Sanmartí Puig, 2007). Traditionell betont die prozedurale Perspektive, das Entwickeln, Testen, Verändern und gegenständliche Repräsentieren von Modellen (Chiu & Lin, 2019). Krüger & Upmeyer zu Belzen (2021) verstehen hierbei die Testung eines Modells als das Aufstellen von Hypothesen auf Basis des Modells sowie deren Überprüfung während der Modellanwendung. Diese Anwendung des Modells als Forschungswerkzeug adressiert, im Unterschied zur Anwendung des Modells als verständniserzeugendes Medium, primär prozedurale und epistemologische Aspekte des Modellierens.

Epistemologischer Aspekt

Die Reflexion über die eigene naturwissenschaftliche Modellierung und Modellnutzung stellt den epistemologischen Aspekt der Modellkompetenz dar. Dieses metakognitive Wissen wird von Schwarz et al. (2009) als Teil eines übergreifenden Naturwissenschaftsverständnisses angesehen. Auch Leisner (2005a) beschreibt den epistemologischen Aspekt der Modellkompetenz als eine dem Wissenschaftsverständnis untergeordnete Teildimension, die sie als Modellverständnis bezeichnet. Meisert (2008) definiert Modellverständnis als das Verstehen der Rolle von Modellen und Modellarbeit und schreibt dieser Teilkompetenz eine übergeordnete Bedeutung zu. Ferner wird diese Dimension als wichtiger Bestandteil eines angemessenen Verständnisses von Nature of Science angesehen (Gilbert, 1991). Im Hinblick auf Nature of Science fungieren Modelle als variable Wissensverzeichnisse, die zu neuen Forschungsfragen anregen (Chiu & Lin, 2019). Den einzelnen Aspekten von Modellverständnis wird im Vergleich internationaler Ansätze eine unterschiedlich große Bedeutung zugeteilt (vgl. Tabelle 44). Meisert (2009) unterteilt das Modellverständnis in vier Teildimensionen, die das Verhältnis von Modell und Original genauer betrachten. Darunter fallen die *Modell-Original-Relation*, die *Funktion von Modellen als Mittel der Erkenntnisgewinnung*, der *Entwicklungscharakter von Modellen* und die *Modellierer-Rolle*. Der explizite Vergleich von Modell und Original führt bei Lernenden zu einem verbesserten Wissenschaftsverständnis (Meisert, 2008). Dies liegt darin begründet, dass es für Schüler*innen einfacher ist, zwischen Modellen und dem modellierten Original zu unterscheiden, als zwischen den Theorien der Wissenschaftler und der darin behandelten Realität zu kontrastieren (Grosslight et al., 1991). Meisert (2008) sieht die drei Teilkompetenzen der Modellkompetenz nicht als getrennte oder unabhängige Lernfelder, sondern als sich gegenseitig ergänzende und differenzierende Teilkompetenzen an. Da dem Modellverständnis eine übergeordnete Rolle zukommt, sind die drei Teilkompetenzen nicht als Paralleldimensionen, sondern als hierarchisches Konstrukt gestaltet (ebd.). Das Modellwissen bietet eine Sammlung konkreter Beispiele, um die Konzepte des Modellverständnisses verstehen und aufgrund der Vielfalt an Modellen erweitern zu können (ebd.). Die Modellarbeit bietet Konkretisierungen zum Interpretations- und Entwicklungscharakter von Modellen (ebd.). Ein weit entwickeltes Modellverständnis hingegen wirkt sich differenzierend auf die Teilkompetenz der Modellarbeit sowie auf die Qualität und das Ausmaß des Modellwissens aus (ebd.; s. Tabelle 33). Wie auch Leisner (2005a) sieht Kircher (1995) das Modellverständnis als eine Teildimension des Wissenschaftsverständnisses an. Wissenschaftsverständnis (= „Understanding the Nature of Science“) umfasst nach Sodian et al. (2002) Wissen über die Entstehung von Erkenntnis und Wissen, die Charakteristika von Experimenten und Theorien und die Zusammenhänge dieser naturwissenschaftlichen Konzepte untereinander. Nach einer Analyse naturwissenschaftlicher, angloamerikanischer Curricula von McComas, Clough &

Almaroza, (1998) wird Wissenschaftsverständnis verstanden als das Wissen um die naturwissenschaftliche Methodik, die Bedeutung von experimentellen Ergebnissen, die Funktion von Theorien im (natur)wissenschaftlichen Erkenntnisprozess, die rationale Begründung und den tentativen Charakter wissenschaftlichen Wissens sowie das evolutionäre wie auch revolutionäre Wesen der Wissenschaften. Auf Basis der Unterteilung in wissenschaftstheoretische, erkenntnistheoretische und wissenschaftsethische Aspekte lässt sich das Wissenschaftsverständnis weiter differenzieren und operationalisieren (Kircher & Dittmer, 2004; Sodian et al., 2002). Die erkenntnistheoretische Dimension beschäftigt sich mit der Existenz und der Erkennbarkeit der Realität sowie mit der Interpretation der Wahrnehmung dieser (Günther, 2006). Der wissenschaftstheoretische Teilbereich umfasst die Natur von Theorien, Hypothesen, Experimenten, Modellen etc. und wie sich diese Teilbereiche zu wissenschaftlichen Methoden zusammensetzen (ebd.). Der wissenschaftsethische Aspekt behandelt ethische Fragen nach der Wertneutralität von wissenschaftlichen Erkenntnissen und den moralischen Grenzen zum Einsatz von Technologien im Kontext der Gesellschaft (ebd.). Kircher & Dittmer (2004) formulieren vier Leitfragen, die das Modellverständnis als Teilaspekte eines Wissenschaftsverständnisses charakterisieren. Als erkenntnistheoretischen Aspekt stellen Kircher & Dittmer (2004) die Frage nach dem Zeitpunkt, ab dem in den Naturwissenschaften ein Modell entwickelt werden muss (1) und damit nach den Grenzen der Methoden zur Erkenntnisgewinnung in den Naturwissenschaften. Nach Meyling (1990) kann diese erste Frage ebenfalls in Teilen als wissenschaftstheoretischer Aspekt angesehen werden. Die Fragen nach den Eigenschaften, die ein naturwissenschaftliches Modell auszeichnen (2) oder wie durch Modellbildung Erkenntnisse über reale Objekte und Phänomene gewonnen werden können (3), werden dem wissenschaftstheoretischen Bereich zugeordnet. Die Modellkritik sowie das Reflektieren über die eigene Modellanwendung (4) bilden den wissenschaftstheoretischen Teil des Modellverständnisses.

Ferner wird Wissenschaftsverständnis als die didaktische Reduktion (Carey et al., 1989) der Konstrukte Natur der Naturwissenschaften (NOS) und Verständnis der Natur der naturwissenschaftlichen Forschung (NOSI) definiert (Sodian et al., 2002; Günther et al., 2003; Birkholz & Elster, 2014). NOS behandelt ausschließlich wissenschaftstheoretische Aspekte wie das Wissen um die Gültigkeit und die Qualitäten wissenschaftlichen Wissens (Smith et al. 2000). Darüber hinaus bezieht NOSI als Verständnis des Forschungsprozesses auch erkenntnistheoretische Aspekte wie das Wissen um das Vorgehen und die Schwierigkeiten der Erkenntnisgenerierung mit ein.

McComas et al. (1998) extrapolieren aus acht internationalen Wissenschaftsstandards einen Konsens über 14 Ziele eines angemessenen Wissenschaftsverständnisses. Leisner (2005a) hebt davon folgende Ziele hervor, die konkret durch eine gezielte Förderung von Modellkompetenz erreicht werden können:

- 1) „Das Wissen in den Naturwissenschaften hat trotz seiner Beständigkeit einen vorläufigen Charakter.
- 2) Das Wissen in den Naturwissenschaften beruht überwiegend auf Beobachtungen, experimentellen Belegen sowie rationalen Überlegungen und Skepsis.
- 3) Es gibt keine allgemeingültige Standardmethode, um Naturwissenschaften zu betreiben und daher auch keine universelle Anleitung für naturwissenschaftliche Forschung.
- 4) Naturwissenschaften sind ein Ansatz, um natürliche Phänomene erklärbar zu machen.

[...]

- 10) Naturwissenschaftler sind kreativ.“ (McComas, Clough & Almazroa, 1998, S.6f)

Diese Ziele spiegeln sich explizit in der Entwicklung von Modellen und deren Relevanz im naturwissenschaftlichen Erkenntnisprozess wider (ebd.). Der tentative Charakter der Modelle (Sodian et al., 2002) drückt sich in der Vorläufigkeit des aus den Modellen resultierenden naturwissenschaftlichen Wissens aus (Ziel 1). Im Zuge des naturwissenschaftlichen Erkenntnisprozesses werden Modelle als Bestandteil von rationalen Argumentationen bei der Hypothesenbildung sowie bei Beobachtungen und Deutungen zu experimentellen Belegen herangezogen (Leisner, 2005a). Die Skepsis gegenüber diesen Deduktionen (Ziel 2) findet dabei insbesondere in der Modellkritik Berücksichtigung, die den Modellcharakter und die Grenzen der Modelle betont. Das Zurückgreifen auf die Modellmethode im naturwissenschaftlichen Erkenntnisprozess stellt eine Alternative zu den klassischen experimentellen Methoden eines forschend-entwickelnden Unterrichts dar. Dadurch wird verdeutlicht, dass es keine Standardmethode in den Naturwissenschaften gibt (Ziel 3). Stoßen Naturwissenschaftler an die Grenzen des Nachweisbaren, helfen sie sich mit (Denk-) Modellen aus, um Phänomene erklärbar zu machen. Das Wissen über die Ko-Existenz mehrerer Denkmodelle und darüber, dass neben Modellen noch andere Ansätze existieren, liefert einen erheblichen Beitrag zum Wissenschaftsverständnis (Ziel 4). Die dabei entwickelten Denkmodelle stellen kreative Ansätze dar (Ziel 10), die von den Lernenden im Umgang mit der Modellmethode ebenfalls Kreativität erfordern. Auch wenn einige Studien auf einen Zusammenhang zwischen Fachwissen und Wissenschaftsverständnis deuten (z.B. Halloun, 2001), ist der Mehrwert des Wissenschaftsverständnisses für den Lernprozess bislang wenig empirisch belegt (Leisner-Bodenthin, 2006).

In ihrem Kompetenzmodell zur Modellkompetenz differenzieren Upmeier zu Belzen & Krüger (2010) das Fachwissen als eine zur Modellkompetenz abgegrenzte Inhaltsdimension (s. Tabelle 33). Modellkompetenz wird hier zunächst ausschließlich innerhalb des Kompetenzbereichs Erkenntnisgewinnung gedacht und als eine Metaebene zur Reflexion über Modelle verstanden (ebd.). Das Rahmenkonzept nach Mayer (2007, S.178) zur Erkenntnisgewinnung definiert die drei zentralen Dimensionen „*Charakteristika der Naturwissenschaften*“ (Wissenschaftsverständnis), „*wissenschaftliche Erkenntnismethoden*“ (Wissenschaftliches Denken) und „*praktische Arbeitstechniken*“ (Manuelle Fertigkeiten). Daran angelehnt operationalisieren Upmeier zu Belzen & Krüger (2010, S.49) Modellkompetenz mittels der Dimension „*Kenntnisse über Modelle*“ äquivalent zum Bereich *Wissenschaftsverständnis* und über die Dimension „*Modellbildung*“ entsprechend den Bereichen *Wissenschaftliches Denken* und *Manuelle Fertigkeiten*. Die Dimension „*Kenntnisse über Modelle*“ umfasst die ontologischen und epistemologischen Positionen zu Modellen (Günther, 2006) und liegt daher im Wissenschaftsverständnis begründet (Upmeier zu Belzen & Krüger, 2010; vgl. Tabelle 33). Dieses individuelle, konzeptuelle Wissen (Steiner, 2007) wird weiterhin in die Teilkompetenzen „*Eigenschaften von Modellen*“ und „*Alternative Modelle*“ unterteilt (Upmeier zu Belzen & Krüger, 2010, S.49). Die Teilkompetenz *Eigenschaften von Modellen* umfasst Abschätzungen zu Ähnlichkeit und Abstraktionsgrad zwischen Modell(-objekt) und Ausgangsobjekt (z.B. Justi & Gilbert, 2003). Die Teildimension „*Alternative Modelle*“ behandelt Vorstellungen zur Frage, warum unterschiedliche Modelle zu einem Ausgangsobjekt existieren können (Upmeier zu Belzen & Krüger, 2010). Eine angemessene Kenntnis über Modelle liegt vor, wenn:

- Ein Modell als gebildete, gedankliche oder gegenständliche Rekonstruktion eines Ausgangsobjekts infolge eines durch ein Subjekt definierten Herstellungsprozesses verstanden wird.
- Modelle als Möglichkeitsformen verstanden werden, die aufgrund des Herstellungsprozesses oder der besonderen, vom Subjekt individuell bestimmten, Anwendungssituation zu alternativen Rekonstruktionen führen.

Modellierungskompetenz

Die Kompetenzmodelle zur Modellkompetenz lassen sich in Ansätze unterscheiden, die sich auf die Nutzung der Modelle beschränken (Treagust et al. 2002; Leisner-Bodenthin, 2006; Oh & Oh, 2011; vgl. Tabelle 44) und solche, die den Prozess der Modellierung mit einbeziehen (Grosslight et al., 1991; Justi & Gilbert, 2003; Crawford & Cullin, 2005; Meisert, 2008; Schwarz et al., 2009; Upmeyer zu Belzen und Krüger, 2010). Um der Modellierungstätigkeit gerecht zu werden, versuchen Krüger & Upmeyer zu Belzen (2021) durch den Wechsel des Begriffs Modellkompetenz zu Modellierkompetenz gegenüber der traditionell überwiegend ontologischen Auffassung von Modellen stärker zu betonen. Dazu betrachten sie die Aspekte von Modellen „Eigenschaften von Modellen“, „Alternative Modelle“, „Zweck von Modellen“, „Testen von Modellen“ und „Ändern von Modellen“ allgemein im Kontext eines Modellierungsprozesses. So definieren sie Modellierkompetenz als „die Fähigkeiten, beim Herstellen von Modellen einen theoriegeleiteten oder kreativen Erkenntnisprozess zu initiieren, bei der Anwendung von Modellen zweckbezogen Erkenntnisse zu gewinnen, über Modelle mit Bezug auf ihren Zweck zu urteilen und über den Prozess der Erkenntnisgewinnung durch Modelle und das Modellieren zu reflektieren“ (Krüger & Upmeyer zu Belzen, 2021, S.134). Neben diesen Dimensionen operationalisieren Grosslight et al. (1991), Crawford & Cullin (2005) und Schwarz et al. (2009) Modellierungskompetenz zusätzlich noch über eine Dimension zum Entwerfen und Erstellen von Modellen. Die Operationalisierung von Justi & Gilbert (2003) hingegen behandelt den Modellierungsprozess als das Erstellen neuer Modelle im Zuge einer erweiterten Testung von Modellen (Accreditation of models), infolge einer unzureichenden Eignung der getesteten Modelle. Für den Kontext chemische Reaktionen operationalisieren Oliva et al. (2015) Modellierkompetenz neben der epistemologischen Dimension „Nature of Models“ über die Dimension „Learning and Working with Models“. Diese zweite Dimension schließt dabei sowohl deklarative als prozedurale Komponenten mit ein. Das Zusammenfassen dieser beiden Aspekte der Modellkompetenz stellt ein Alleinstellungsmerkmal des Kompetenzmodells von Oliva et al. (2015) dar.

Tabelle 3: Positionierung der Dimensionen der nationalen und internationalen Kompetenzmodelle zur Modellkompetenz in Bezug auf Fachwissen und Erkenntnisgewinnung. Hellgrau unterlegt: Kompetenzmodell auf das sich als theoretische Grundlage in dieser Arbeit gestützt wird. Dunkelgrau unterlegt: Modellverständnis als epistemologische Dimension.

	Naturwissenschaftliche Bildung				
Autor	Fachwissen	Erkenntnisgewinnung			
Leisner-Bodenthin (2006)	Deklaratives Wissen		Prozedurales Wissen		Selbstständigkeitsgrad
	Modelleigenschaften/ Inhalt des Modells	Modellverständnis Wissenschaftsverständnis	Anwenden des Modellverständnisses	Anwenden des Modells „Modellmethode“	-
Hodson (1992, 2014)	Learning science	Learning about science		Doing science	-
Henze et al. (2007)	learn scientific models	nature of models		act of modelling	-
Mayer (2007)	-	Wissenschaftsverständnis		Manuelle Fertigkeiten (inquiry methods)	Wissenschaftliche Erkenntnismethoden
Meisert (2008, 2009)	Modellwissen	Modellverständnis		Modellarbeit	-
Schwarz et al. (2009)	subject matter expertise	epistemological understanding		practice of scientific knowledge	-
Upmeyer zu Belzen & Krüger (2010)	-	Kenntnisse über Modelle		Modellbildung	
	-	Eigenschaften von Modellen	Alternative Modelle	Manuelle Fertigkeiten	Wissenschaftliches Denken
Digel (2020)	deklarativ	epistemologisch		prozedural	

Tabelle 4: Aspekte von Modellen im Kontext von Modellverständnis im internationalen Vergleich (erweitert nach Krell et al., 2014)

Quelle	Charakteristika von Modellen	Mehrere Modelle	Zweck von Modellen	Testen von Modellen	Vorläufigkeit	Modelliertätigkeit
Grosslight et al. (1991)	Kinds of Models	Multiple models	Purpose of models	(purpose of models)	Changing a model	Designing and creating models
Treagust et al. (2002)	Models as exact replicas	Models as multiple representations	Uses of scientific models; models as explanatory tools	-	Changing nature of models	-
Justi & Gilbert (2003)	Nature / entities of models	Relative uniqueness	Use; prediction	Accreditation of models	Change over time	Prediction (Accreditation of Models)
Crawford & Cullin (2005)	-	Multiple models to explain and predict	Purpose of models	Validating/testing models	Changing a model	Designing and creating models
Leisner-Bodenthin (2006)	Modelle sind Hypothesen	-	Zweckmäßigkeit von Modellen	Veranschaulichen, Erklären, Vorhersagen	Vorläufigkeit von Modellen	-
Meisert (2009)	Modell-Original-Relation	-	Modelle als Mittel der Erkenntnisgewinnung	-	Entwicklungsscharakter von Modellen	Modellierer-Rolle
Schwarz et al. (2009)	Models are generative tools	-	Using models to explain and to predict	Evaluating models	Revising models; models change to capture improved understanding	Constructing models
Upmeier zu Belzen & Krüger (2010)/ Gogolin und Krüger (2018)	Nature of Models	Multiple models	Purpose of models	Testing models	Changing models	-
Oh & Oh (2011)	Meanings of a model	Multiplicity of scientific models	Purposes of models	Use of models in the science classroom	Change in scientific models	-

Der Modellkompetenzbegriff in der vorliegenden Arbeit

Das Verständnis der Begriffe Modellkompetenz und Modellverständnis in der vorliegenden Arbeit Modellkompetenzbegriff entspricht der Operationalisierung nach Leisner (2005a, 2005b) bzw. Leisner-Bodenthin (2006). Leisner (2005a) schließt den Modellierungsprozess nicht in ihr Kompetenzmodell mit ein. Die Aspekte der Erkenntnisgewinnung mit Modellen beziehen sich ausschließlich auf die Modellanwendung/-nutzung. Weiterhin bietet ihr Ansatz eine stärker ausdifferenzierte Beschreibung der theoriebezogenen - nicht prozeduralen - Dimensionen einer Modellkompetenz für den Umgang mit Erkenntnismodellen (Leisner-Bodenthin, 2006). Leisner (2005a) operationalisiert die Modellkompetenz auf Basis eines erkenntnistheoretisch hypothetisch-realistischen Modellbegriffs. Laut Leisner-Bodenthin (2006) unterteilt sich die Modellkompetenz in deklaratives Wissen, prozedurales Wissen und den Selbstständigkeitsgrad. Das deklarative Wissen umfasst ein angemessenes Modellverständnis sowie das Wissen über die Modelleigenschaften im Sinne des Inhalts eines konkreten (Denk-)Modells (Leisner, 2005a). Leisner-Bodenthin (2006) operationalisiert die Dimension Modellverständnis weiterhin als Wissen über sieben Subkategorien:

- MV 1: Die Definition des Modellbegriffs
- MV 2: Die Unterschiede zwischen der Erfahrungs- und der Modellwelt
- MV 3: Unterschiedliche Modelltypen sowie eine implizite Kenntnis über Struktur-, Funktions- und Konstruktmodelle (Eschenhagen et al., 2008)
- MV 4: Die Unterschiede zwischen naturwissenschaftlichen Modellen und Modellen im Alltag
- MV 5: Den Zweck von Modellen
- MV 6: Die Modellbildung
- MV 7: Die Veränderbarkeit von Modellen und die Rolle multipler Modelle im naturwissenschaftlichen Erkenntnisprozess

Justi & Gilbert (2003) beschreiben ein adäquates Modellverständnis als ein „valides Verständnis der Natur von Modellen und Modellierungen“ (vgl. Justi & Gilbert, 2003, S.1370). Laut Justi & Gilbert (2003) weisen Schüler*innen ein angemessenes Modellverständnis auf, wenn sie über folgende Konzepte verfügen:

- Modelle sind keine einzigartigen Ausschnitte eines realen Objekts, Prozesses oder einer Theorie.
- Modelle sind veränderbar.

- Modelle werden zur besseren Veranschaulichung genutzt. Auch werden Voraussagen über das Verhalten und die Eigenschaften von Objekten und Systemen ermöglicht.
- Modelle können sich erst dann etablieren, wenn sie innerhalb eines Expert*innenkreises der jeweiligen Community anerkannt werden.

Unter Berücksichtigung der Operationalisierung des Modellverständnisses nach Justi & Gilbert (2003) sowie des Stufenmodells zum Wissenschaftsverständnis nach Carey et al. (1989) definiert Leisner (2005a) ein angemessenes Modellverständnis als das Wissen über folgende Punkte:

- (Naturwissenschaftliche) Modelle sind von einem Modellierer konstruiert worden und nicht entdeckt oder zusammengetragen, insbesondere wenn der originale Sachverhalt nicht mehr direkt erfahrbar ist. Modelle erfüllen dann den Zweck Phänomene zu veranschaulichen, zu erklären und vorherzusagen.
- Im Zuge der Konstruktion müssen Vereinfachungen (Lehnert & Köhler, 2012), Spekulationen und Abstraktionen vorgenommen werden.
- Modelle sind stets zweckmäßig. Da der Zweck vom Konstruierenden festgelegt wird, kann auch nur auf Basis dieses Zwecks eine Beurteilung nach der Richtigkeit eines Modells erfolgen.
- Modelle haben einen temporären und hypothetischen Charakter. Daraus resultiert, dass Modelle veränderbar sind.
- Modelle sind abhängig von der Kritik der Forschungscommunity.
- Lernende können zwischen dem Modellierten und dem realen Phänomen unterscheiden.

Leisner (2005a) operationalisiert die Dimension Modelleigenschaften (= „Inhalt des Modells“) analog zur Teilkompetenz Modellwissen nach Meisert (2008). Die Dimension Modelleigenschaften umfasst das Fachwissen bezüglich des Modellgegenstands (Leisner, 2005a). Als deklarative Komponente der Dimension Modelleigenschaften versteht Leisner (2005b) konkrete Annahmen, Idealisierungen und Inhalte naturwissenschaftlicher Modelle. Die Auswahl, das Anwenden und (Be-)Werten naturwissenschaftlicher Modelle im Problemlöseprozess werden zusammenfassend als Modellmethode bezeichnet und dem prozeduralen Wissen zugeordnet (ebd.). Das Anwenden des Modellverständnisses ordnet Leisner (2005a) ebenfalls dem prozeduralen Wissen zu. Leisner (2005a) versteht darunter Fähigkeiten wie die bewusste Differenzierung zwischen Phänomen und Modell, sowie zwischen naturwissenschaftlichen

Modellen und Alltagsmodellen. Die Fähigkeit über naturwissenschaftliche Modelle und den Prozess des Modellierens zu reflektieren, wird ebenfalls als Anwenden des Modellverständnisses dem Prozeduralen Wissen zugeordnet (ebd.).

Unter dem Selbstständigkeitsgrad versteht Leisner (2005a) die Befähigung zur selbstständigen Umsetzung der Dimensionen *Deklaratives* und *Prozedurales Wissen* als aktive Handlungen im naturwissenschaftlichen Unterricht. Dabei sollen die Lernenden in der Lage sein das Wissen und die Fähigkeiten in diversen, unbekannten Situationen angemessen und ergebnisorientiert anzuwenden. Die Aspekte des Modellseins nach Mahr (2008a, 2015) können genutzt werden, um Beschreibungen von Ausprägungen einer Modellkompetenz auf Basis unterschiedlicher Entwicklungsniveaus theoretisch zu begründen (Upmeyer zu Belzen & Krüger, 2010). Die Niveaustufen basieren auf dem Grad der Fähigkeit, diese Aspekte in Problemkontexten beim Modellieren und der Modellnutzung zu berücksichtigen (ebd.). Dabei spielt die Beziehung zwischen dem modellierenden oder nutzenden Subjekt und dem Modell selbst eine bedeutsame Rolle (Kircher, 2015). Das Subjekt entscheidet über die fokussierten Ähnlichkeitsbeziehungen und den Zweck dieser Fokussierung (Giere, 2010).

Leisner-Bodenthin (2006) differenziert in ihrem Kompetenzmodell zwischen einer Fachwissens- und einer Erkenntnisgewinnungskomponente. Gleichzeitig betont dieser Ansatz als einziger den prozeduralen Charakter der Anwendung des Modellverständnisses und des Inhalts des Modells. Dies sind zwei der Schlüsselkompetenzen zum erfolgreichen Problemlösen im Kontext physikalisch-chemischer Phänomene und deren Deutung (Gruber & Stamouli, 2015). Das Anwenden des Inhalts eines Modells ermöglicht es erst, Phänomene auf der Teilchenebene zu deuten. Die anschließende Prüfung der gewonnenen Erkenntnis auf Übertragbarkeit auf andere Phänomene oder gar auf Generalisierbarkeit, auch in Form einer Modellkritik, bedarf der konkreten Anwendung des Modellverständnisses (Leisner 2005a). Im Rahmen dieser Arbeit werden Aspekte der Modellbildung und Modellierung im konzeptionellen, praktischen Teil der Studie III von den Lernenden nicht bearbeitet. Folglich basieren sämtliche weiteren Annahmen im Rahmen der Konzeption, Testung, Datenerhebung und -auswertung auf dem Kompetenzmodell nach Leisner (2005a, 2005b) bzw. Leisner-Bodenthin (2006). Lernende zeigen unabhängig davon, ob sie als Modellnutzer oder Modellierer tätig sind, keine Unterschiede in der Entwicklung ihrer epistemologischen Modellkompetenz (Mierdel & Bogner, 2019). Demnach stellt das Auslassen der Modellierungstätigkeiten eine Vereinfachung des Lernprozesses, jedoch keinen Nachteil für die Entwicklung der Lernenden dar.

Schüler*innenvorstellungen zu Teilchen und dem inneren Aufbau der Materie

Alternative Vorstellungen zum Teilchenmodell treten bei Lernenden jeden Alters auf und zeigen in Abhängigkeit der Ursachen in unterschiedliche Ausprägungen. Obwohl Schüler*innen vielfach mit Modellen zum inneren Aufbau der Materie und mit Teilchenvorstellungen in Berührung kommen, benutzen Sie diese Modelle selten, um Phänomene zu erklären (Fischler & Lichtfeldt, 1997). Besonders bei Alltagsphänomenen nehmen Lernende Erklärungsansätze auf Basis des Diskontinuums nicht als relevant war (ebd.). Fischler und Lichtfeld (1997) heben hervor, dass es im normalen Alltag der Schüler*innen es auch nicht notwendig scheint, Konstrukte dieser Art als Erklärung heranzuziehen, da ihre alternativen Vorstellungen zur Vorhersage von Phänomenen ausreichen. Weiterhin herrscht international betrachtet ein allgemeiner Konsens dazu, auf welche Faktoren die Probleme alternativer Vorstellungen beim Modelllernen zurückzuführen sind (Riaz, 2004):

- Lehrkräfte geben den Schüler*innen, vor allem in den früheren Lernphasen wie der Grundschule, inadäquate Erklärungen des Teilchenkonzepts. Diese sind weiterhin häufig mit einer mangelhaften didaktischen Aufbereitung verbunden.
- Es existieren überwiegend zweifelhafte Erklärungen und Darstellungen des Teilchenkonzepts in Schulbüchern etc.
- Praktischen Aktivitäten zum Teilchenkonzept finden in der Unterrichtspraxis kaum Anwendung.

Schüler*innenvorstellungen spielen eine entscheidende Rolle für die Bildung adäquater, mentaler Modelle (s. *Abschnitt 2.1*) und damit für den Lernerfolg. Für erfolgreiches Modelllernen ist es folglich notwendig die vorliegenden Schüler*innenvorstellungen zu den behandelten Kontexten und zum betrachteten Denkmodell zu berücksichtigen. Alternative Vorstellungen zu Löse- und Diffusionsprozessen basieren zumeist auf abweichenden Konzepten zum Wechsel der Aggregatzustände. Auch gibt es Überschneidungen mit den Alternativkonzepten der Schüler*innen zum Vorgang der chemischen Reaktion. Im Folgenden werden die häufigsten alternativen Vorstellungen der Schüler*innen zu den Themen der Lerneinheit ausgeführt und Implikationen bezüglich der Aufbereitung der Lerninhalte dargestellt. Auf eine vollständige Darstellung der Kategorien existierender Schüler*innenvorstellungen wird aufgrund fehlender Relevanz für die untersuchten Konstrukte und Interventionsinhalte verzichtet.

Wechsel der Aggregatzustände

Die alternativen Vorstellungen der Schüler*innen zum Wechsel der Aggregatzustände lassen sich nach Andersson (1990) in die Kategorien Verschwinden, Verlagerung, Wandlung, Transmutation und chemische Interaktion unterteilen. Rund 60% der jüngeren Schüler*innen im Alter von bis zu sechs Jahren vertreten die Ansicht, dass Flüssigkeiten beim Verdunsten verschwinden (Andersson, 1990; Barker, 2000). Innerhalb dieser Vorstellung können Stoffe aufhören zu existieren und verlieren beim Verdunsten ihre Masse (Barker, 2000). Werden Stoffe gelöst, verschwinden sie zwar, einige ihrer Eigenschaften wie Geschmack, Geruch oder Farbe bleiben dennoch erhalten (ebd.). Diese Vorstellung tritt in den ersten beiden Schuljahren jedoch rapide in den Hintergrund (Andersson, 1990). Im Alter von sieben bis acht Jahren gehen 65% der Schüler*innen davon aus, dass Stoffe nicht verschwinden, sondern sich an einen anderen Ort verlagern (ebd.). Diese Schüler*innen haben das Konzept der Erhaltung der Masse bei Flüssigkeiten bereits verinnerlicht. Übertragen auf das Verhalten von Gasen, speziell Luft können sie dieses Konzept hingegen noch nicht anwenden (ebd.). Ein mentales Konzept zum Vorgang des Verdampfens liegt noch nicht vor. Die Schüler*innen sind bis zum Alter von elf Jahren auch überwiegend der Überzeugung, dass die Blasen in kochendem Wasser aus Luft bestehen (Osborne & Cosgrove, 1983). Erst wenn die Schüler*innen im Alter von zwölf Jahren Luft als Materie verstehen, erklären 60% von ihnen das Verdunsten von Wasser als die Umwandlung in Wasserdampf. Neben dieser Ansicht haben 13% der Schüler*innen die Vorstellungen, dass sich das Wasser in Luft umwandelt. Dieses Konstrukt der Transmutation bekommt bis zu einem Alter von 16 Jahren mit einer Abundanz von 27% noch mehr Gewichtung (Osborne & Cosgrove, 1983). Parallel dazu entwickelt sich bei einem doppelt so großen Teil der Schüler*innen (von 23% auf 48%) die Vorstellung, dass sich die Wassermoleküle in die beiden Elemente Wasserstoff und Sauerstoff aufteilen (ebd.). Ähnliche Vorstellungen besitzen Schüler*innen im Alter von 13-16 Jahren beim Schmelzvorgang von Salzen. Diese alternativen Konzepte der chemischen Interaktion, lassen darauf schließen, dass eine Einführung der Vorstellung von Molekülen erst dann Sinn macht, wenn das Verhalten der Stoffe mit dem Teilchenmodell passend erklärt werden kann. Dies hebt die Bedeutung des Teilchenmodells für das Verständnis primärer Konzepte der Chemie hervor. Ebenfalls ist zu erwarten, dass der Lösungsprozess von Schüler*innen ebenfalls als eine chemische Interaktion der beiden beteiligten Stoffe angesehen wird. Weitere Vorstellungen von Schüler*innen aus dieser Kategorie beschreiben den Schmelzvorgang als die Entstehung von Wasser (Osborne & Cosgrove, 1983). Dieses Gedankenkonstrukt findet sowohl für den Kontext von Eis und Wasser (ebd.) als auch für das Schmelzen von Salzen Anwendung (Karplus, 1996). Die Schüler*innen denken, dass sich innerhalb des Salzkristalls Wasser befindet, das beim Schmelzen austritt und eine Salzlösung herstellt (Renström, Andersson, & Marton, 1990). Ferner kommt es vor, dass Schüler*innen die beiden Begriffe „Wasser“ und „Flüssigkeit“ als Synonyme verwenden

(Andersson, 1990). Prieto, Watson & Dillon (1992) weisen dabei daraufhin, dass Schüler*innen die Begriffe mit einer Intention verwenden, die von der wissenschaftlichen Sichtweise abweicht.

Innerer Aufbau der Materie

In der populärsten Vorstellung der Schüler*innen zum Teilchenmodell werden Atome nicht als ein primärer Baustein angesehen, der permanent und in jeder Phase der Materie vorliegt. Die Schüler*innen verstehen Atome vielmehr als das finale Produkt eines repetitiven Teilungsprozesses von Stoffen (Pfundt, 1981; Andersson, 1990). Da dieser Teilungsprozess von den Schüler*innen als willkürlich angesehen wird, variiert auch die Form der Teilchen eines Elements in der Vorstellung der Schüler*innen (Pfundt, 1981). Da Schüler*innen die Teilchen als Endprodukt der Teilung eines Stoffs ansehen, schreiben sie den Teilchen makroskopische Eigenschaften wie Farbe, Temperatur, Oberflächenbeschaffenheit und Kontinuumseigenschaften zu (Pfundt, 1981; Driver et al. 1985; Renström et al. 1990; Fischler & Lichtfeldt, 1997; Mikelskis-Seifert, 2002). Eine weitere Vorstellung, die aus makroskopischen Beobachtungen von den Schüler*innen auf den Mikrokosmos übertragen wird, ist das Konzept der Reibung. Aufgrund von Reibung zwischen den Teilchen kommt die Brownsche Bewegung irgendwann zum Erliegen (Fischler & Lichtfeldt, 1997). Die Vorstellungen der Lernenden zu kleinsten Teilchen lassen sich laut Andersson (1990) ebenfalls in die Kategorien Verschwinden, Verlagerung, Wandlung, Transmutation und chemische Interaktion unterteilen. So haben Lernende die Vorstellung, dass Atome sich in „Nichts“ auflösen, wenn sie beispielsweise oxidieren. Zur Kategorie der Verlagerung zählen Vorstellungen, die chemische Prozesse mit einer Adhäsion von Atomen aus dem Raum an Oberflächen erklären, anstelle von Wechselwirkungen der Elektronenhüllen und Kerne der einzelnen Atome untereinander. Auch die Vorstellung, dass neu entstandene Moleküle als ein Gemenge aus den Molekülen der Edukte zu verstehen sind, wird der Kategorie der Verlagerung zugeordnet. Unter die Kategorie Modifikationen werden Vorstellungen gefasst, laut denen sich nur die Eigenschaften der Teilchen verändern. Die eigentliche Identität der Teilchen bleibt jedoch dieselbe. Als Transmutationskonzept versteht Andersson Vorstellungen, die eine Veränderung der Anzahl der Elektronen oder Nukleonen bei gleichzeitigem Konservieren der eigentlichen Art von Atomen vertreten.

Kategorisiert man Schüler*innenvorstellungen zum Aufbau von Materie, zeigen sich im Denken von Schüler*innen Parallelen zum historischen Verlauf des Erkenntnisprozesses in der Fachwissenschaft Chemie (Barke et al., 2015). In einem der häufigsten Präkonzepte deuten Lernende den Raum zwischen den einzelnen Atomen nicht als Vakuum (Pfundt, 1981; Barke & Harsch, 2001b Barke, 2011). Schüler*innen im Alter von dreizehn bis sechzehn Jahren vertreten die Vorstellung von Materie in Form eines Stoffs als kontinuierliche Masse, in der

einzelne Teilchen des Stoffes gleichmäßig verteilt vorkommen (Renström, Andersson, & Marton, 1990). Es gibt verschiedene Varianten von Kontinuumsvorstellungen. Bei manchen bestehen das Kontinuum und die darin gelagerten Teilchen aus demselben Stoff (ebd.). Andere wiederum ordnen dem Kontinuum und den Teilchen verschiedene Stoffe zu (Mikelskis-Seifert, 2002). Bei dieser Art von Konstrukten vertreten Schüler*innen häufig die Vorstellung, dass die Räume zwischen den Teilchen mit Luft gefüllt sind (Fischler & Lichtfeldt, 1997). Teilchenbewegung liegt bei dieser Kategorie von Präkonzepten nur vereinzelt bei Schüler*innen vor (Andersson, 1990). Ähnlich zu dieser Kategorie ist die Kategorie der Teilchenkonglomerate (ebd.). In dieser Vorstellung sind die Teilchen eines Stoffes so eng gepackt, dass zwischen ihnen kein freier Raum mehr ist (Barke & Harsch, 2001b, Barke, 2011). Die Möglichkeit der Teilchen sich zu bewegen, wird hier nur sehr selten von den Schüler*innen angenommen (Andersson, 1990). Eine erweiterte Schüler*innenvorstellung beinhaltet das Konzept eines Diskontinuums an Teilchen, deren Eigenschaften jedoch denen von realen, makroskopischen Körpern entsprechen (ebd.). Diese Schüler*innen verfügen über kein bewusstes Denkmodell, das sie im Zusammenspiel mit Beobachtungen versuchen weiterzuentwickeln (ebd.). Vielmehr sind sie versucht ihre mentalen Modelle in jeder Situation so anzupassen, dass eine für sie zufriedenstellende Erklärung zustande kommt (ebd.).

2.4 Modelltypabhängige Lernendenvariablen

Kognitive Belastung

- Die Cognitive Load Theorie (CLT) betrachtet auf Aufgaben während komplexen Lernsituationen. Die Kapazität des Arbeitsgedächtnisses ist begrenzt und darf durch die drei Arten von CL nicht überschritten werden:
 - Intrinsic CL resultiert aus der inhärenten Komplexität der Lernaufgabe.
 - Extraneous CL wird primär durch die didaktische Aufbereitung des Lernmaterials bestimmt.
 - Germane CL ist die effektive kognitive Belastung, die für das Lernen aufgebracht wird.
- Die Cognitive Theory of Multimedia Learning liefert u. a. die Designkriterien Multimediaprinzip, Kohärenz, Räumliche Kontiguität, Segmentierung, Signalisierung und Personifizierung mit den Zielen die sachfremde Verarbeitung zu reduzieren, die kognitive Verarbeitung zu organisieren und ein tieferes Verständnis zu fördern.

Affektive Konstrukte

- Laut der Person-Gegenstands-Theorie setzt sich Interesse aus einer affektiven Komponente zur Klärung der emotionalen Valenz und einer kognitiven Komponente zur Klärung der epistemischen Valenz sowie einer Wertschätzungskomponente zusammen.
- Man unterscheidet zwischen aktuellem-situationalem und dispositionalem-individuellem Interesse: Catch-Faktoren wecken kurzzeitiges situationales Interesse und Hold-Faktoren schaffen durch Sinnhaftigkeit langfristige Aufmerksamkeit, die in individuelles Interesse als eine dauerhafte Neigung übergehen kann.
- Die Selbstbestimmungstheorie nennt drei Basic Needs: Kompetenzerleben sowie die Bedürfnisse nach Autonomie und sozialer Eingebundenheit.
- Neuheitseffekte in Vergleichsstudien begünstigen Interesse aufgrund des Innovationscharakters der Interventionen, nehmen aber mit zunehmender Gewöhnung ab.
- Die Expectancy-Value Theory besagt, dass Erwartungen an die Intervention im Sinne einer selbsterfüllenden Prophezeiung eintreffen.
- Das akademische Selbstkonzept ist das Wissen über die Ausprägung der eigenen Fähigkeiten und gliedert sich akademisches und nichtakademisches Selbstkonzept. Aufgrund seiner affektiven Dimension wirkt sich das Konstrukt auf das psychische Wohlbefinden sowie auf die Leistungsbereitschaft und -fähigkeit aus. Das Fähigkeitsselbstkonzept moderiert den signifikanten Zusammenhang zwischen Interesse und Lernleistung.

System Usability

- Usability ist die Gebrauchstauglichkeit eines Produkts oder Systems und setzt sich aus der Effektivität zur Lösung einer Aufgabe, der Effizienz bei der Aufgabenbearbeitung und der Zufriedenstellung der Nutzer zusammen.
- Man unterscheidet Content-Usability als didaktische Passung von System-Usability als Fehlerrobustheit, Angemessenheit der Aufgaben, Steuerbarkeit, Erwartungskonformität oder Individualisierbarkeit. Beide Arten bedingen und beeinflussen sich gegenseitig.
- Usability hat sowohl auf die Lernerfahrung der Lernenden als auch auf das Lernergebnis einen signifikanten Einfluss.

Kognitive Fähigkeiten

- Die Informationsverarbeitungsgeschwindigkeit und die Arbeitsgedächtniskapazität sind grundlegenden Faktoren von Intelligenz.
- Neben Intelligenz gilt Vorwissen als einer der wichtigsten Prädiktoren für schulischen Erfolg.

Die Art, wie Informationen präsentiert werden, beeinflusst kognitive, motivationale und selbstregulatorische Variablen der Lernenden. Folglich hat der Modelltyp durch die unterschiedliche Gestaltung und Präsentation der Lerninhalte ebenso einen Einfluss auf Lernendenvariablen. Im folgenden Abschnitt werden die theoretischen Grundlagen der vom Modelltyp abhängigen Lernendenvariablen aufgeführt und deren Bedeutung für die vorliegende Arbeit aufgezeigt.

Kognitive Belastung

Dieser Abschnitt befasst sich mit dem Konstrukt kognitive Belastung und der Theorie zur Informationsverarbeitung und Wissensaufbau unter Einbezug der Cognitive Load Theorie sowie der Erweiterung dieser um Designkriterien für die Erstellung von (digitalen) Lehr-Lernmedien.

Das Konstrukt kognitive Belastung im Rahmen der Cognitive Load Theory

Kognitive Belastung ist ein mehrdimensionales Konstrukt, das die Belastung aufgrund einer bestimmten Aufgabe für das kognitive System des Lernenden darstellt (Paas & Van Merriënboer, 1994). Das allgemeine Modell zur kognitiven Belastung nach Paas & Van Merriënboer (1994) umfasst sowohl eine kausale als auch eine Bewertungsdimension. Die kausale Dimension umfasst die Wechselwirkungen zwischen Aufgabenmerkmalen und den Lernendeneigenschaften. Aufgabenmerkmale, die für den Cognitive Load (CL) als relevant identifiziert wurden, sind das Aufgabenformat, die Komplexität der Aufgabe, die Verwendung von Multimedia, der Zeitdruck und die Geschwindigkeit der Aufgabeninstruktion. Als relevante Lernendeneigenschaften gelten das Kompetenzniveau, das Alter und die Fähigkeit zur räumlichen Vorstellung (ebd.). Die Bewertungsdimension spiegelt die messbaren Konzepte von mentaler Belastung, mentaler Anstrengung und der erbrachten Leistung wider (Sweller et al., 1998). Man unterscheidet zwischen zwei Arten von kognitiver Belastung, die während der Ausführung einer Aufgabe auftreten: „task-based“ und „learner-based“ (ebd.). Der „task-based“ Anteil, der aus der eigentlichen Aufgabenstellung resultiert, wird als mentale Belastung (mental load) bezeichnet (ebd.). Der „learner-based“ Anteil hingegen resultiert aus der Investition an kognitiven Kapazitäten, um die Aufgabenstellung zu erfüllen und wird als Denkanstrengung bezeichnet (ebd.). Die Denkanstrengung wird wiederum in eine kognitive und eine motivationale Komponente gegliedert, da die investierten kognitiven Kapazitäten nicht nur durch die Anforderungen der Aufgabe, sondern auch durch die Bereitschaft zur Denkanstrengung begründet sind (Paas et al., 2005). Die mentale Belastung basiert auf der Wechselwirkung von Aufgabenmerkmalen und den Merkmalen der Fachinhalte. Dies gilt als Prädiktor für die erwarteten kognitiven Kapazitätsanforderungen und kann für eine Vorabschätzung der kognitiven Belastung herangezogen werden (Paas & Van Merriënboer, 1994). Mentale Anstrengung bezieht sich auf die tatsächlich beanspruchte, kognitive Kapazität, die zum Erfüllen der gestellten Anforderungen einer Aufgabe aufgewendet werden muss. Individuelle Merkmale, die das Lernen stark vorhersagen, wie kognitive Fähigkeiten, Lerntypen, Selbstregulation, Motivation und Affekt, werden im Rahmen der Cognitive Load Theory (CLT) nicht berücksichtigt (Moreno, 2005). Das einzige individuelle Merkmal, das explizit im theoretischen Rahmen von CL betrachtet wird, ist Vorwissen (Kalyuga, Chandler & Sweller, 1998). Dies grenzt das Konstrukt der kognitiven Belastung vom Konstrukt Workload ab. Beide Konstrukte berücksichtigen zunächst die

Anforderungen, die Aufgaben an Personen stellen (Moreno und Park, 2010). Die psychologischen Auswirkungen auf die Belastungswahrnehmung, ausgehend von den Überzeugungen, Erwartungen und Zielen des Einzelnen, werden in der CLT jedoch nicht berücksichtigt (Moreno und Park, 2010). Für Vergleichsstudien zu Lernmaterialien gilt die kognitive Belastung als ebenso relevant wie die erbrachten kognitiven Leistungen der Schüler*innen, da die kognitive Belastung Auskunft über die Denkanstrengung (mental effort) der erbrachten Leistung gibt (Paas & van Gog, 2006). Daher wird in der vorliegenden Studie der Einfluss des Modelltyps auf die kognitive Belastung erhoben.

Die CLT befasst sich hauptsächlich mit der kognitiven Belastung, die bei Lernenden während komplexen Lernsituationen auftritt (Paas, Renkl, & Sweller, 2004). Das Ziel der CLT ist es Empfehlungen auf Basis der menschlichen, kognitiven Struktur für Handlungsaufforderungen zu geben, die es den Lernenden ermöglichen die Belastung des Arbeitsgedächtnisses gering zu halten und somit besser zu lernen (Sweller, van Meriënboer, & Paas, 1998). Die CLT verfolgte dabei zunächst das primäre Ziel die Informationen so zu strukturieren, dass externe kognitive Belastungen möglichst gering ausfallen und unerfahrene Lernende möglichst viele Ressourcen zum Lernen aufbringen können (Moreno & Park, 2010). Auch die Entwicklung von Unterrichtsmethoden zur effizienten Nutzung der begrenzten, kognitiven Verarbeitungskapazität, um neues Wissen und Fähigkeiten mit vorhandenen zu verknüpfen, sind Gegenstand der CLT (Paas et al., 2003).

In Lernsituationen müssen Informationen im Arbeitsgedächtnis verarbeitet und im Langzeitgedächtnis gespeichert werden. Eine Grundannahme der CLT ist die begrenzte Kapazität des Arbeitsgedächtnisses zur Verarbeitung von Informationen während des Prozesses des Wissenserwerbs (Paas et al., 2010). Die zweite Annahme besteht darin, dass das Langzeitgedächtnis unbegrenzten Informationsspeicher bietet (Sweller et al., 1998). In Anlehnung an die Schema-Theorie wird Wissen als Informationseinheiten organisiert und in Schemata angelegt (Rumelhart, 1981). Dieser Prozess wird verstanden als Organisieren der lerngegenstandsbezogenen Informationen in größere Informationseinheiten (Kalyuga, 2010). Im Anschluss erfolgt der Transfer dieser Informationseinheiten ins Langzeitgedächtnis. Bei erneuter Konfrontation in lerngegenstandsbezogenen Kontexten können Lernende auf diese bereits vorhandenen Schemas zurückgreifen (Kalyuga, 2010). Der automatisierte Zugriff auf die vorhandenen Schemas ermöglicht es die Belastung des Arbeitsgedächtnis im Problemlöseprozess zu reduzieren (Moreno & Park, 2010). Für die Schemabildung gelten folgende Prinzipien (Sweller et al., 1990): (1) Die Schemabildung ist ein Baustein einer qualitativen Lernleistung. (2) Sie erfordert eine auf Problemstellungen und die damit verbundenen Lösungsschritte gerichtete Aufmerksamkeit. (3) Der Lernprozess wird begünstigt, wenn Lernende sich um den Erwerb von Schemas bemühen. (4) Um eine schwere kognitive Belastung und somit die Beeinträchtigung des

Lernens zu vermeiden, gilt es andere kognitive Aktivitäten einzuschränken. Die CLT unterscheidet drei verschiedene Arten von kognitiver Belastung, die *intrinsic cognitive load* (ICL), *extraneous cognitive load* (ECL) und *germane cognitive load* (GCL) genannt werden. Um Lernen zu ermöglichen, darf laut Sweller (2010) die Summe der drei Arten kognitiver Belastung die Kapazität des Arbeitsgedächtnisses nicht überschreiten. Dies gewährleistet ausreichend freie kognitive Kapazitäten zur Bewältigung der eigentlichen Aufgabe und wirkt somit einem Cognitive Overload entgegen (Sweller, 2010). Als eine Folge von kognitiver Überlastung neigen Lernende zu stark vereinfachten Erklärungen im Bereich des deklarativen Wissens (Ziemek et al., 2005).

Intrinsic Cognitive Load

Intrinsic Cognitive Load resultiert aus der inhärenten Komplexität der Lernaufgabe. Diese Art der Belastung hängt von den beiden Faktoren der Elementinteraktivität und dem Vorwissen der Lernenden ab (Moreno und Park, 2010). Die Elementinteraktivität korrespondiert mit der Anzahl an Elementen, die von den Lernenden während des Bearbeitens der Lernaufgabe gleichzeitig bewältigt werden müssen. Eine Aufgabe hat eine geringe Elementinteraktivität, wenn die einzelnen Elemente nacheinander bearbeitet werden können. Dies ist insbesondere dann gewährleistet, wenn sich die einzelnen Aufgabenelemente aufeinander beziehen (Sweller, 2010). Aufgaben mit hoher Elementinteraktivität umfassen Elemente die stark miteinander verknüpft sind und daher eine gleichzeitige Bearbeitung erfordern (ebd.). Das Vorwissen spielt dann eine große Rolle, wenn neue Informationen mit bestehenden Schemas verknüpft werden können. Besteht ein umfangreiches und gut strukturiertes Vorwissen müssen die Lernenden weniger Elemente ohne Bezüge zu bereits bestehendem Wissen verarbeiten (ebd.). Daraus ergeben sich zwei effektive Maßnahmen, um die intrinsische kognitive Belastung zu reduzieren. Zunächst besteht die Möglichkeit die Elemente zu segmentieren, um die Elementinteraktivität zu reduzieren (Mayer & Moreno, 2010). Dabei werden die Informationen nacheinander präsentiert, um Lernenden mit geringerem Vorwissen dabei zu unterstützen die anfallenden Informationen zu organisieren. Dadurch erlangen die Lernenden mit jedem Schritt mehr Vorwissen, wodurch es ihnen leichter fällt, neues Wissen aus den kommenden Schritten mit den vorherigen zu verknüpfen (ebd.). Die Segmentierung ist Teil des Designs des Arbeitsmaterials und beeinflusst daher auch immer den ECL. Eine weitere Möglichkeit die intrinsische Belastung zu reduzieren, besteht darin, den Lernenden das nötige Wissen bereits zuvor zur Verfügung zu stellen (Mayer & Pilegard, 2005). Diese „künstliche“ Reduzierung von ICL in Form einer Reduzierung des Schwierigkeitsgrades wird jedoch mit einer reduzierten geforderten Lernleistung gleichgestellt (van Meriënboer & Sweller, 2005). Generell besteht ein Problem darin, die intrinsische kognitive Belastung zu reduzieren, ohne das eigentliche Lernziel der Aufgabe zu verändern. Ist Wissenserwerb das Ziel einer Aufgabe, wird durch ein Voranstellen

relevanter Informationen der eigentliche Zweck als Lernaufgabe nicht mehr erfüllt. Die Aufgabe erfordert in diesem Fall nur noch das Reproduzieren oder den Transfer des vorhandenen Wissens.

Extraneous Cognitive Load

Der eigentliche Lernprozess mit dem Lerngegenstand führt bei den Lernenden ebenfalls zu kognitiver Belastung (Sweller, 2010). Als Lernprozess wird die aktive und konstruktive Auseinandersetzung mit dem Lerngegenstand verstanden. Voraussetzung dafür ist eine möglichst hohe, freie Kapazität an kognitiven Ressourcen, die in den Lernprozess investiert werden können. Gilt dies als erfüllt, kann ein Wissenszuwachs in Form des Erwerbs und der Automatisierung von neuen und bestehenden Schemas erfolgen (Sweller, 2010). ECL wird primär durch die didaktische Aufbereitung des Lernmaterials bestimmt (Sweller, 1994). Immer wenn der Lernende mentale Ressourcen für Prozesse aufbringen muss, die nicht dem Bearbeiten der Aufgabe selbst dienen, erhöht sich der ECL (ebd.). Beispiele hierfür wären das Suchen oder Ausblenden von Informationen aus dem Lernmaterial. Um den ECL bei der Nutzung von Anschauungsmodellen möglichst gering zu halten, sind daher die Übersichtlichkeit und die Redundanzfreiheit der Modelle von Bedeutung. Wenn das Modell nur Informationen enthält, die zum Lernziel der Aufgabe passen, werden keine kognitiven Ressourcen für redundante Inhalte aufgebracht. Basierend auf Erkenntnissen in der medienpsychologischen Forschung ergeben sich eine Vielzahl von Möglichkeiten, um den ECL zu senken (Klepsch et al., 2017). Die Designprinzipien multimedialer Informationspräsentation nach Mayer (2009) liefern hierfür eine gute Grundlage (Klepsch et al., 2017). Diese Prinzipien zur Reduzierung des ECL beschreibt Mayer (2009) in der kognitiven Theorie des multimedialen Lernens und werden im Abschnitt *Cognitive Theory of Multimedia Learning* dieser Arbeit näher erläutert.

Germane Cognitive Load

Der Germane Cognitive Load wird auch effektive kognitive Belastung genannt und grenzt sich damit von der ineffektiven kognitiven Belastung ab (Paas, Renkl & Sweller, 2003). GCL resultiert aus Aktivitäten, die für das Lernen aufgebracht werden (ebd.). Diese Aktivitäten erleichtern dabei das Lernen oder tragen zu einer Transferleistung bei, um ein korrektes mentales Modell zu generieren (ebd.). Daher zeigt ein hoher GCL-Wert an, dass die Lernenden engagiert sind und ihre noch vorhandenen mentalen Ressourcen für den Lernprozess und die Beschäftigung mit der Aufgabe aufbringen. Weiterhin steht den Lernenden nur dann freie kognitive Kapazität zum Lernen zur Verfügung, wenn die Belastung durch den ECL nicht bereits das Arbeitsgedächtnis vollständig beansprucht (Klepsch et al., 2017). Dadurch ist eine Erhöhung des GCL auch nur dann möglich, wenn die mentale Kapazität nicht bereits durch ineffektive oder inhaltsbezogene kognitive Belastung vollkommen ausgereizt ist. Strategien zur Erhöhung der

effektiven kognitiven Belastung, sind zum Beispiel die Erhöhung Variabilität von Aufgabenbeispielen (Paas & Van Merriënboer, 1994) oder deren kontextuelle Interferenz (van Merriënboer et al., 2002). Der GCL steht in direktem Zusammenhang zur aktuellen Motivation der Lernenden (Paas, Renkl & Sweller, 2003). Dabei wird zum einen angenommen, dass eine höhere aktuelle Motivation sich positiv auf die investierte Denkanstrengung während der Aufgabenbearbeitung auswirkt (ebd.). Auf der anderen Seite gehen Vertreterinnen und Vertreter der „Expectancy-Value Theory of Motivation“ davon aus, dass eine höhere effektive kognitive Belastung zu einer erhöhten aktuellen Motivation führt (Feldon et al., 2019).

Cognitive Theory of Multimedia Learning

Das Ziel der Cognitive Theory of Multimedia Learning (CTML) ist es Designkriterien für die Erstellung von (digitalen) Lehr-Lernmedien zu generieren und deren Einbettung in die Lernumgebung lernförderlich zu gestalten (Mayer, 2009, 2014). Die CTML liefert hierbei einen theoretischen Rahmen zur Gestaltung von Lernmaterialien, die simultan Bild und Text zur Darstellung der Inhalte nutzen (Mayer, 2009, 2014). Mayer geht bei der CTML von drei Grundannahmen aus:

1. Das Arbeitsgedächtnis verfügt nur über begrenzte Kapazität.
2. Dem Arbeitsgedächtnis stehen zwei Verarbeitungskanäle zur Informationsverarbeitung zur Verfügung.
3. Der Wissensaufbau kann nur durch eine aktive Informationsverarbeitung erfolgen.

Das Verstehen eines Lerninhalts wird hier aufgefasst als die Bildung eines kohärenten mentalen Modells, das die relevanten Informationen des Lerninhalts und deren Beziehungen untereinander abbildet (Mayer, 2009, 2014). Zur Bildung eines solchen kohärenten mentalen Modells, müssen die Lernenden zunächst die relevanten Bild- und Textelemente selektieren. Diese Elemente müssen im Anschluss innerhalb eines verbalen sowie eines bildhaften mentalen Modells organisiert werden. Schließlich werden beide mentalen Modelle miteinander und zusammen mit dem Vorwissen aus dem Langzeitgedächtnis in ein kohärentes mentales Modell integriert. Dieser letzte Schritt gilt dabei als schwierigste, jedoch auch gleichzeitig entscheidendste Leistung beim Lernprozess (ebd.). Zur möglichst lernförderlichen Gestaltung der Verarbeitungsprozesse wurden Designkriterien basierend auf den theoretischen Annahmen zur Informationsverarbeitung entwickelt und in mehreren Studien empirisch belegt (Mayer, 2009, 2014). Mayer (2009) formuliert für diese Kriterien drei konkrete Ziele:

1. Die Vermeidung sachfremder kognitiver Verarbeitungsprozesse (reduce extraneous processing).

2. Die Bewältigung und Organisation grundlegender kognitiver Verarbeitungsprozesse (manage essential processing).
3. Die Unterstützung der Verarbeitungsprozesse, die einem tieferen Verständnis dienen (foster generative processing).

Die entwickelten Designkriterien sind noch nicht abschließend dahingehend überprüft, ob sie einen Einfluss auf bloß ein Ziel oder gar auf mehrere Ziele simultan haben (ebd.) Im Folgenden werden lediglich die Kriterien nach Mayer (2009) aufgeführt, die für die Konzeption der Modelle und der Lernmaterialien relevant sind und im Laufe des Konstruktionsprozesses berücksichtigt wurden. Zusätzliche Designkriterien sind bei Mayer (2009, 2014) zu finden.

Multimediaprinzip

Im Vergleich zum Lernen, das alleinig anhand von Texten erfolgt, ist das Lernen mit Bild und Text gleichzeitig deutlich lernwirksamer ($d = 1.39$). Aufgrund der simultanen Bildung eines verbalen und eines visuellen mentalen Modells werden Verarbeitungsprozesse, die ein tieferes Verständnis aufbauen gefördert. Das Multimedia-Prinzip gilt als besonders bedeutsam für Lernende mit geringem Vorwissen.

Kohärenzprinzip

Inhalte von Lernmaterialien, die für die Lernenden interessant aber für das Lernziel irrelevant sind (Seductive Details), können den Lernerfolg negativ beeinflussen. Der Lernprozess gestaltet sich erfolgreicher, wenn irrelevante Inhalte, wie Informationen in Form von Text oder Abbildungen, nicht in das Lernmaterial integriert sind ($d = 1.66$). Darüber hinaus erhöht eine zusätzliche Reduktion der Inhalte auf das Nötigste die Lerneffektivität weiter. Dadurch werden unnötige Verarbeitungsprozesse reduziert. Die negative Auswirkung der „Seductive Details“ kann verschiedene Ursachen haben. Zum einen werden die Lernenden abgelenkt und es werden kognitive Kapazitäten von irrelevanten Inhalten genutzt. Zum anderen kann dadurch irrelevantes Vorwissen aktiviert werden, was den Aufbau eines adäquaten mentalen Modells stören oder zur Einbettung der Informationen in unpassende Schemata führen kann.

Räumliches Kontiguitätsprinzip

Abbildungen und die dazugehörigen Texte sollten stets räumlich nah zueinander präsentiert werden. Durch die räumliche Nähe werden vermeidbare Suchprozesse umgangen und die Verknüpfung zwischen den verbalen und bildlichen Informationen erleichtert. Der Einsatz von Lernmaterialien, die nach dem räumlichen Kontiguitätsprinzip gestaltet wurden, führten zu einem signifikant höheren Wissenszuwachs als Lernmaterialien, die dieses Kriterium verletzten ($d = 1.09$). Das räumliche Kontiguitätsprinzip gilt besonders für komplexe Lernmaterialien als relevant.

Segmentierungsprinzip

Zur Verarbeitung komplexer Lerninhalte empfiehlt es sich diese in kleinere Sinnabschnitte zu unterteilen. Dies ermöglicht es den Lernenden in ihrem persönlichen Tempo die einzelnen Inhalte durcharbeiten. Die schrittweise Segmentierung bietet dadurch die Möglichkeit am Ende eines Segments bereits einzelne Verarbeitungsschritte abzuschließen, bevor bereits neue Informationen kognitive Kapazitäten beanspruchen. Darüber hinaus erleichtert die Einteilung in Sinneinheiten, den Aufbau sinnvoller Schemata sowie deren Integrierung in neue oder bestehende mentale Modelle (Mayer & Pilegard, 2005). Mayer & Pilegard (2014) konnten die erhöhte Lernwirksamkeit von segmentierten Lernmaterialien im Vergleich zu unsegmentierten Lernmaterialien bereits empirisch belegen ($d = 0.79$).

Signalisierungsprinzip

Um die kognitive Last durch Suchprozesse weiter zu reduzieren, empfiehlt es sich die Aufmerksamkeit der Lernenden auf zentrale Begrifflichkeiten und Bildelemente zu lenken. Bezogen auf Text kann Signalisierung in Form von Zwischenüberschriften, Hervorheben einzelner Wörter und die Verwendung strukturierender Begriffe erfolgen. Für die visuelle Signalisierung gelten Pfeile oder markante Farben als hilfreich. Die höhere Lernwirksamkeit des Einsatzes von Stilmitteln zur Signalisierung konnte mit einem mittleren Effekt ($d = 0.52$) empirisch bestätigt werden. Insbesondere Novizen profitieren von der Unterstützung durch Signalisierung. Eine besondere Form der visuellen Signalisierung stellt die Farbkodierung dar. Zusammengehörige Elemente werden dabei in den gleichen Farben gekennzeichnet. Diese Art der Signalisierung eignet sich besonders für komplexeres Lernmaterial. Eine Farbkodierung des Lernmaterials führt bei Lernenden zu einer geringeren kognitiven Belastung sowie zu einem erhöhten Lernerfolg (J.-H. Wu & Yuan, 2003). Signalisierung unterstützt die Lernenden nicht nur in der Selektion der Lerninhalte, speziell die Farbkodierung erleichtert auch die Organisation und Integration der Informationen in vorhandene Schemas (van Gog, 2014).

Personifizierungsprinzip

Lernende lernen erfolgreicher, wenn Texte im Stil gesprochener Sprache anstelle von formalen Schreibstilen geschrieben sind. Methoden zur Umsetzung dieses Prinzips, sind die direkte Ansprache der Lernenden als Personen oder das Nutzen von Erwähnungen mit Bezug auf den Autor als konkrete Persönlichkeit. Durch Erwecken des Eindrucks, dass die Lernenden direkt angesprochen würden, steigt ihre Motivation und sie sind eher bemüht die Lerninhalte zu verstehen. Personifizierte Lernmaterialien zeigen sich mit einem mittleren Effekt ($d = 0.79$) lernwirksamer als solche, die in einem formalen Schreibstil formuliert wurden (Mayer, 2014a).

Affektive Konstrukte

Neben den leistungsbezogenen Kriterien des Kompetenzerwerbs gelten affektive und motivationale Aspekte der Kompetenzgewinnung wie das Selbstkonzept, Lernfreude oder das Interesse als weitere wichtige Zielgrößen guter Bildung (Jansen et al., 2013; Weinert & Helmke, 1996). Im Folgenden werden die Konstrukte (situationales) Interesse und Fähigkeitsselbstkonzept näher erläutert.

Interesse

In der didaktischen Forschung hat das Konstrukt Interesse in den letzten Jahren eine zentrale Rolle eingenommen. Interesse gilt als wichtiger Faktor für erfolgreiches Lernen und ist daher ein häufig betrachtetes Konstrukt fachdidaktischer Forschung (Blankenburg & Scheersoi, 2018). Das Feld der Interessensforschung brachte verschiedene theoretische Auffassungen des Konstrukts Interesse hervor, die wiederum zu einer Vielzahl von Operationalisierungen des Interessenskonstrukts geführt haben (Krapp & Prenzel, 2011). Die Person-Gegenstands-Theorie nach Krapp (1992) ist im pädagogisch-psychologischen Feld anerkannt und gilt als die am weitesten verbreitete Interessentheorie, die sich explizit auf Lernprozesse bezieht (Blankenburg & Scheersoi, 2018). Interesse wird hier als das dynamische Verhältnis zwischen einer Person und einem Erfahrungsbereich oder Wissensgebiet definiert (Krapp, 1992). Dabei ist Interesse nicht mit dem Spaß am Lernen gleichzusetzen (Krapp & Prenzel, 2011). Das Konstrukt ist vielmehr inhaltspezifisch operationalisiert und wird als Interesse einer Person an einer Handlung oder Sache gesehen (Gardner, 1998). Interesse wird als die langfristige Grundlage von aktueller Motivation betrachtet (Schiefele, 2009). Ferner wird Interesse im Rahmen der Person-Gegenstands-Theorie als ein multidimensionales Konstrukt aus affektiven, wertbezogenen und kognitiven Merkmalskomponenten betrachtet (Hidi et al., 2004; Krapp, 1992; Krapp & Prenzel, 2011). Die affektive Komponente bezieht sich auf die emotionale Valenz, die eine Person mit einem bestimmten Gegenstand verbindet. Die wertbezogene Komponente hingegen bezieht sich auf die persönliche Wertschätzung eines Gegenstandes für die Person. Die kognitive Komponente betrachtet die epistemische Valenz bezogen auf das ausdifferenzierte Wissen, das eine Person über einen bestimmten Gegenstand besitzen möchte. Durch die multidimensionale Natur von Interesse kann es von anderen motivationalen Konstrukten wie Selbstkonzept, Bedürfnissen, Zielorientierungen oder Motiven abgegrenzt werden.

Weiterhin lässt sich Interesse in zwei Formen unterteilen (Krapp, 1992; Schiefele, 2008) - eine aktuelle, situationale Form und eine dispositionale, individuelle Form. Das situationale Interesse bezieht sich auf die Interessantheit, die durch spezifische Situationen ausgelöst wird (Schiefele, 2008). Bezogen auf Unterricht können zum Beispiel die besonderen

Charakteristika einer Lernumgebung solche Situationen bieten (Blankenburg & Scheersoi, 2018). Das dispositionale, individuelle Interesse hingegen stellt ein stabiles Persönlichkeitsmerkmal in Form eines andauernden Zustands dar (Blankenburg & Scheersoi, 2018). Für Medienvergleichsstudien ist vor allem der Einfluss auf das situationale Interesse bedeutsam, da das Medium und die Lernumgebung nur einen geringen Einfluss auf das vorhandene individuelle Interesse haben (Willems, 2018). Jedoch besteht die Annahme, dass das individuelle Interesse durch das situationale beeinflussbar ist (Krapp, 2002a). Zudem wird die dispositionale Form als notwendige Voraussetzung für den Wissenserwerb (Edelmann, 2003; Willems, 2018) durch das situationale Interesse der Lernenden begünstigt (Schiefele, 2008). So geht auch der Bildung von individuellem, dispositionalem Interesse die Generierung von situationalem, aktuellem Interesse vorweg (Krapp, 2002b). Der Prozess der Interessensentwicklung wird zunächst überwiegend durch die emotionale Valenz und affektive Komponenten des Interesses bestimmt, während im späteren Verlauf Aspekte der wertbezogenen und epistemischen Valenz sowie kognitive Faktoren des Interesses in den Fokus rücken (Renninger & Hidi, 2011). Das daraus resultierende individuelle Interesse stellt ein zeitlich stabiles und situationsunabhängiges Persönlichkeitsmerkmal dar (Hidi & Renninger, 2006). Hidi und Renninger (2006) beschreiben in ihrem Vier-Phasen-Model die Entwicklung von Interesse über folgende psychologische Interessenszustände:

- (1) Das ausgelöste situationale Interesse („Triggered Situational Interest“) resultiert aus kurzfristigen Veränderungen in der affektiven und kognitiven Verarbeitung. Das betrachtete kurzfristige Interesse wird meist durch äußere *Catch-Faktoren* des Interessensgegenstands geweckt, wie etwa unpassende oder überraschende Informationen, persönliche Relevanz oder Identifikation mit dem Thema (Garner et al., 1992; Mitchell, 1993; Renninger & Hidi, 2002; Sadoski, 2001).
- (2) Das aufrechterhaltene situationale Interesses folgt auf die erste Phase und zeichnet sich durch fokussierte und anhaltende Aufmerksamkeit gegenüber dem Interessensgegenstand aus. Diese Aufmerksamkeit tritt über eine längere Zeit erneut auf und hält wiederum länger an. Voraussetzung für das Aufrechterhalten sind *Hold-Faktoren* wie die Sinnhaftigkeit der Aufgaben und/oder aktive Beteiligung der Lernenden (Harackiewicz et al., 2000; Mitchell, 1993).
- (3) Das beginnende individuelle Interesse stellt neben einem psychologischen Interessenszustand auch die Anfangsphase einer relativ dauerhaften Neigung dar. Diese Phase ist durch positive Gefühle, grundlegendes Wissen und assoziierte Werte gekennzeichnet. Diese Faktoren erleichtern es dem Lernenden sich auch ohne externe

Anregung im Laufe der Zeit immer wieder mit dem Interessensgegenstand auseinanderzusetzen (Renninger, 2000).

- (4) Ein ausgeprägtes individuelles Interesse bezieht sich sowohl auf die psychologische Interessenslage als auch auf die relativ dauerhafte Bereitschaft, sich selbstbestimmt im Laufe der Zeit wieder mit betroffenen Inhaltsklassen zu beschäftigen. Diese Phase umfasst zunächst alle Charakteristika des beginnenden individuellen Interesses. Darüber hinaus zeichnet es sich dadurch aus, dass mehr positive Gefühle, umfangreicheres Wissen sowie mehr Wertassoziationen in Bezug auf den Interessensgegenstand als auf andere Themen und Aktivitäten vorliegen (Renninger, 2000). Die Lernenden befassen sich selbst dann mit dem Interessensgegenstand, wenn dies mit Anstrengung und Herausforderungen verbunden ist.

Um die Entwicklung von Interesse zu beschreiben, orientiert sich die Person-Gegenstands-Theorie an den Kategorien der Selbstbestimmungstheorie der Motivation von Deci und Ryan (1993), wobei zwei Ebenen der Handlungssteuerung unterschieden werden (Krapp, 2002a). Die erste Ebene basiert auf kognitiv-rationalen Prozessen und bezieht sich vor allem auf das Problem der Intentionsbildung bzw. der bewussten Auswahl von Lernzielen (ebd.). Es gilt die Prämisse, dass eine Person nur dann langfristiges Interesse entwickeln kann, wenn der Interessensgegenstand basierend auf kognitiv-rationalen Überlegungen als ausreichend bedeutsam bewertet wird (Krapp, 1998). Dieser Aspekt steht im Mittelpunkt der eher kognitiv orientierten Motivationsforschung (Krapp, 2002a), bezieht jedoch im Gegensatz zu kognitiven Motivationstheorien, wie zum Beispiel der Theorie der Selbstwirksamkeitserwartung nach Bandura (1977), auch emotionale Aspekte mit ein (Krapp & Ryan, 2002). Die zweite Komponente bezieht sich auf die Qualität der subjektiven Erfahrung und des unmittelbaren emotionalen Feedbacks einer Handlung (Krapp, 2002a). Während die kognitiv-rationale Komponente der Handlungssteuerung typischerweise auf der Ebene der bewusst-reflexiven Informationsverarbeitung stattfindet, funktioniert die affektiv-emotionale Komponente sehr häufig auf einer unbewussten Ebene der Informationsverarbeitung. Um Interesse zu entwickeln, müssen dabei die positiven emotionalen Erlebnisse mit dem Lerngegenstand überwiegen (Krapp & Ryan, 2002). Als notwendige Voraussetzung hierfür gilt die Befriedigung der Grundbedürfnisse für menschliches Wohlbefinden („basic needs“) nach Erleben von Autonomie, Kompetenz und sozialer Eingebundenheit (Deci & Ryan, 1993; Ryan & Deci, 2002).

Die Zusammenhänge zwischen Interessensentwicklung und der intrinsischen Motivation sowie der wahrgenommenen psychologischen Grundbedürfnisse gelten als empirisch belegt (Krapp, 2005). Die Selbstbestimmungstheorie der Motivation nach Deci und Ryan erklärt diese Zusammenhänge wie folgt (Deci & Ryan, 1993; Ryan & Deci, 2002): Die Basis stellen drei

Grundbedürfnisse dar, deren Befriedigung als Voraussetzung intrinsische Motivation gilt. Das Bedürfnis nach Kompetenzerleben bezieht sich auf die Neigung den spezifischen Anforderungen der jeweiligen Situationen gerecht zu werden und Ergebnisse zu erzielen. Das Autonomiebedürfnis basiert auf dem Wunsch Entscheidungen eigenständig zu treffen ohne sich kontrolliert zu fühlen. Das Bedürfnis nach sozialer Eingebundenheit ist mit dem Wunsch nach Anerkennung von Personen im jeweiligen sozialen Umfeld assoziiert.

Unterrichtsbedingungen wie Gruppenarbeit, Rätsel, Medieneinsatz etc. zeigen einen empirisch belegten Einfluss auf das situationale Interesse (Cordova & Lepper, 1996; Mitchell, 1993). Werden Lernumgebungen als sinnvoll empfunden und persönlich einbeziehende Aktivitäten geboten, kann dies zur Aufrechterhaltung des situationalen Interesses beitragen (Mitchell, 1993; Renninger et al., 2004). Dabei haben sich sowohl kooperative Gruppenarbeit als auch Einzelarbeit in projektbasierten Settings als lernförderlich erweisen (Renninger et al., 2004). Auch allgemeine Unterrichtsbedingungen oder das Lernumfeld können die Weiterentwicklung von aufkommendem, individuellem Interesse ermöglichen (Krapp & Lewalter, 2001; Renninger et al., 2004). Ein gut entwickeltes, individuelles Interesse unterstützt die Lernenden dabei sowohl den Kontext als auch den Inhalt einer Aufgabe im Prozess der Problemlösung und des Textverständnisses zu berücksichtigen (Renninger et al., 2002). Insbesondere bei Vergleichsstudien mit unterschiedlichen Medien sollte das Auftreten des Neuheitseffekts (auch Neuigkeitseffekt) berücksichtigt werden (Theyßen, 2014). Dieser besagt, dass der Einfluss neuer Medien auf die Motivation, das situationale Interesse und das Lernverhalten von Schüler*innen aufgrund der Neuheit kurzfristig begünstigen kann (Urhahne et al., 2000). Auch kann der Neuheitseffekt infolge neuer Konzeptionen mit Eventcharakter im Zuge der Intervention auftreten (Theyßen, 2014). Im Vergleich zu traditionellem Unterricht können solche Events einen positiven Einfluss auf die Motivation und das Interesse der Lernenden haben. Dieser Einfluss nimmt jedoch mit der Gewöhnung an die Intervention ab (Urhahne et al., 2000). Eine Möglichkeit dem Neuheitseffekt entgegenzuwirken sind Langzeitstudien, in denen zu Beginn eine Eingewöhnungsphase der eigentlichen Intervention vorgeschaltet wird (Theyßen, 2014). Der interessensteigernde Effekt von digitalen Medien wie AR-Modellen (s. Abschnitt 3) wird häufig auf den Neuheitseffekt zurückgeführt, wobei gleichzeitig über eine Abnahme dieses Effekts mit zunehmender Häufigkeit der Nutzung berichtet wird (Di Serio et al., 2013). Einen weiteren Erklärungsansatz für die motivationalen Vorteile von digitalen Medien liefert die Expectancy-Value Theory (Wigfield & Eccles, 2000). Laut dieser wirken sich sowohl Erwartungen an das genutzte Medium und den erwarteten Erfolg im Umgang mit dem Medium als auch die Zuschreibung verschiedener Wertungen des Mediums auf die Motivation während der Nutzung des Medium aus (ebd.). Kinder und Jugendliche nutzen digitale Medien im Alltag meist zu Unterhaltungszwecken (Medienpädagogischer Forschungsverbund Südwest, 2022). Im

Hinblick auf die „Expectancy-Value Theory“ lässt sich annehmen, dass die Lernenden beim Gebrauch von Tablets einen Unterhaltungswert erwarten und dem Medium damit einen höheren „intrinsic value“ zuschreiben (Eccles & Wigfield, 2020). Schüler*innen empfinden die Nutzung von Tablets im Lernprozess als unterhaltsam und angenehm (Dündar & Akçayır, 2014) und haben die Vorstellung, dass Lernen mit AR motivierender ist (Cheng, 2018). Dadurch steigt auch im Sinne einer selbsterfüllenden Prophezeiung die intrinsische Motivation (Deci & Ryan, 1985; Harter, 1981) der Schüler*innen ein Tablet in einer Lernsituation zu nutzen (Deci & Ryan, 1985; Eccles & Wigfield, 2020). Nach der Self-Determination-Theory (Deci & Ryan, 1985) haben Lernende im Umgang mit digitalen Medien ein höheres Kompetenzerleben (Hillmayr et al. 2020). Nehmen die Lernenden gleichzeitig Autonomieerleben wahr, führt dies zu mehr intrinsischer Motivation (Ryan & Deci, 2002). So wiesen die Lernenden in der Untersuchung von Sofianidis (2022) bei der Nutzung AR beim selbstgesteuerten Lernen auch über Neuheitseffekt hinweg eine gesteigerte Motivation auf. Die Lernenden gaben als Vorteil von AR unter anderem an, dass die Darstellungen beliebig oft wiederholbar seien und man somit besser angepasst an das eigene Lerntempo und Lerntyp arbeiten könne (ebd.). Da die Art des Mediums einen Einfluss auf motivationale Lernendenvariablen zeigt, wird in der vorliegenden Studie der Einfluss des Modelltyps auf das situationale Interesse untersucht.

Fähigkeitsselbstkonzept

Das Interessenskonstrukt ist eng an das akademische Selbstkonzept gekoppelt (Daniels, 2008). Konkret wird unter dem akademischen Selbstkonzept das Wissen über die Ausprägung der eigenen Fähigkeiten verstanden (Thomsen et al., 2018). Thomsen et al. (2018) grenzen in ihrer Definition das Selbstkonzept vom Selbstwert ab: Als kognitiv-deskriptiver Teil des Selbst bezieht sich das Selbstkonzept auf das selbstbezogene Wissen eines Individuums. Somit besteht es aus einer Vielzahl an Selbstbeschreibungen. Als affektiv-evaluative Komponente des Selbst bildet der Selbstwert die Bewertung dieser Selbstaspekte ab. Laut dem Selbstkonzeptmodell nach Shavelson et al. (1976) gliedert sich das generelle Selbstkonzept zunächst in ein akademisches und ein nichtakademisches Selbstkonzept. Diese Art der Unterscheidung spiegelt den schulischen Kontext der Selbstkonzeptentwicklung und damit auch den Forschungskontext dieses Ansatzes wider (Thomsen et al., 2018). Das nichtakademische Selbstkonzept wird wiederum in soziales, emotionales und körperliches Selbstkonzept unterteilt (ebd.). Das akademische Selbstkonzept (auch Fähigkeitsselbstkonzept) wird in ein mathematisch-naturwissenschaftliches (=MINT-) und ein verbales Fähigkeitsselbstkonzept gegliedert (Shavelson et al., 1976). Diese lassen sich wiederum in einzelne Schulfächer, Fachinhalte und letztlich konkrete Aufgabentypen unterteilen (ebd.). Für den schulischen Kontext ist insbesondere das akademische Selbstkonzept von Interesse (Roos & Schöler, 2013). Mittels empirischer Untersuchungen wurden die Strukturen der einzelnen Faktoren zum akademischen Selbstkonzept

im Modell nach Shavelson et al. (1976) überprüft. Zum einen konnte die Multidimensionalität des Konstrukts bestätigt werden (Marsh, 1990). Konkret zeigte sich das akademische Selbstverständnis als fachspezifisch, stark domänenspezifisch (MINT vs. sprachlich) und hierarchisch organisiert, wobei das allgemeine akademische Selbstverständnis an der Spitze der Hierarchie steht (Brunner et al., 2010). Darüber hinaus erwiesen sich das sprachliche und das MINT-Selbstkonzept als nahezu nicht miteinander korrelierend (Marsh, 1990; Brunner et al. 2010). Die Bedeutung des Fähigkeitsselbstkonzepts für den schulischen Unterricht zeigt sich in mehreren Zusammenhängen. Je nach Ausprägung kann das Fähigkeitsselbstkonzept über verschiedene Mechanismen Lernprozesse begünstigen oder erschweren (Dickhäuser, 2006). Aufgrund seiner affektiven Dimension wirkt sich das Konstrukt auf das psychische Wohlbefinden sowie auf die Leistungsbereitschaft und -fähigkeit aus (Stiensmeier-Pelster & Schöne, 2008). Als Vorstellung über das Niveau der persönlichen Fähigkeiten steht das Fähigkeitsselbstkonzept in direktem Zusammenhang mit einem Referenzrahmen (Dickhäuser, 2006). In Abhängigkeit von der wahrgenommenen Selbsteinschätzung und der Bezugsgruppe kommt es zu Kontrast- oder Assimilationseffekten (ebd.). Als Kontrast- oder auch „big-fish-little-pond“-Effekt wird der Anstieg des Fähigkeitsselbstkonzepts infolge des Vergleichs mit leistungsschwächeren Schüler*innen bezeichnet (Marsh, 1987). Ein Assimilationseffekt (auch „basking in reflected glory“ genannt) liegt vor, wenn eine Verbesserung des Fähigkeitsselbstkonzepts aufgrund von Assoziation mit leistungsstärkeren Schüler*innen erfolgt (Marsh, 1990). Beide Effekte konnten bereits in einer Längsschnittstudie mit 1290 Schüler*innen der gymnasialen Oberstufe empirisch belegt werden (Köller et al., 2006). In diesem Kontext moderiert das Fähigkeitsselbstkonzept den signifikanten Zusammenhang zwischen Interesse und Lernleistung (Köller et al., 2000). Die positiven Effekte des schulischen Selbstkonzepts und des Interesses auf den Wissenszuwachs konnten in diesem Zusammenhang ebenfalls empirisch bestätigt werden (Köller et al., 2006). Aufgrund zwischen Fähigkeitsselbstkonzept und Interesse bzw. Lernleistung wird das Fähigkeitsselbstkonzept in der vorliegenden Studie als Kontrollvariable erhoben.

System Usability

Unter Usability wird Gebrauchstauglichkeit eines Produkts oder Systems verstanden. Dabei werden drei Maße definiert: Die Effektivität zur Lösung einer Aufgabe, die Effizienz bei der Aufgabenbearbeitung und die Zufriedenstellung der Nutzer zur Erreichung bestimmter Ziele im jeweiligen Kontext (DIN EN ISO 9241-11:2018-03, 2018). Da sich Usability auf die Interaktion zwischen Nutzer und System bezieht, ist sie eher als ein Interaktionsmerkmal anstatt einer Systemeigenschaft im eigentlichen Sinne zu betrachten (Brooke, 1996). Im Kontext von Lehr-Lernmedien unterscheidet man zudem zwischen inhaltsbezogener Content-Usability und systembezogener System-Usability (Nikmehr & Doroodchi, 2008). Content-Usability richtet sich

dabei primär auf Merkmale der angemessenen didaktischen Aufbereitung der Lerninhalte sowie die adäquate technische Implementierung (ebd.). Auch die Konformität zum Lerntyp des Nutzers und seinen persönlichen Präferenzen sowie die Interaktivität des Lernmediums selbst werden der Content Usability zugeordnet (ebd.). Die System-Usability umfasst wiederum Merkmale wie Fehlerrobustheit, Angemessenheit der Lernaufgaben, Steuerbarkeit, Erwartungskonformität oder Individualisierbarkeit (DIN EN ISO 9241-110:2020-05, 2020). Eine strikte Trennung der beiden Konzepte kann jedoch nur schwer erfolgen. Beispielsweise wird die Interaktivität eines Lernmediums anhand didaktischer Überlegungen begründet und weist damit einen eindeutigen Lernbezug auf (Höffler, 2008). Auf der anderen Seite kann Interaktivität nicht ohne die Berücksichtigung von Aspekten der Interfacegestaltung geplant werden. Folglich können die Kategorien nicht als sich gegenseitig ausschließende Faktoren, sondern vielmehr als Merkmalsmengen mit gewissen Überschneidungen betrachtet werden (Karapanos et al., 2018). Eine Usability-Evaluierung stellt somit eine Bewertung der Gesamteffektivität eines Systems im Lernprozess dar (Thuseethan et al., 2014).

Die wahrgenommene Usability (digitaler) Lernmedien beeinflusst den Lernprozess stark (Karapanos et al., 2018; Mayes & Fowler, 1999; Niegemann et al., 2008). Sie hat dabei sowohl auf das Lernergebnis als auch auf die erlebte Lernerfahrung einen signifikanten Einfluss (Karapanos et al., 2018; Vlachogianni & Tselios, 2022). Der Einfluss der Usability auf das Lernen wird dabei über zwei theoretische Zugänge erklärt. Zunächst hat die Usability eines Lernmediums einen direkten Einfluss auf den Cognitive Load während des Lernprozesses (Tracy & Albers, 2006). Wird das System als schwierig zu bedienen empfunden, verbringen die Lernenden mehr Zeit damit, die Nutzung des Systems als etwas über die Inhalte zu lernen (Ardito et al., 2006). Eine unzureichende Usability führt zu einer Steigerung der extrinsischen kognitiven Belastung (ebd.) und verhindert oder erschwert somit das Lernen (Meiselwitz & Sadera, 2008). Darüber hinaus hat die Benutzerfreundlichkeit eines Bildungsmediums einen Einfluss auf die affektiven Variablen des Lernprozesses (Karapanos et al., 2018; Tselios et al., 2008). So kann eine ungenügende Usability beim Benutzer zu einer verschlechterten Lernerfahrung (Vlachogianni & Tselios, 2022) oder gar zu Frustration im Umgang mit dem Lerngegenstand führen (Bruun et al., 2016). Dies kann zunächst zu einem geringeren situationalen Interesse während der Lernsituation führen (Karapanos et al., 2018). Darüber hinaus kann die Akzeptanz des Bildungsmediums auf Seiten der Lernenden durch eine unpassende Anwenderfreundlichkeit verhindert werden (Tselios et al., 2008). Aufgrund der Unterschiede in der Handhabung der unterschiedlichen Modelltypen wird in der vorliegenden Studie der Einfluss des Modelltyps auf die System Usability als Kontrollvariable mit erhoben.

Kognitive Fähigkeiten

In der differentiellen Psychologie werden die kognitiven Leistungs- und Denkfähigkeiten von Individuen als Teilkonstrukte der Intelligenz aufgefasst. Obwohl der Begriff Intelligenz kontrovers diskutiert wird, nehmen die meisten psychologischen Theorien heute eine Interpretation an, die Intelligenz definiert als die Fähigkeit Probleme zu lösen, aus Erfahrungen zu lernen und sich an neue Situationen anzupassen (Myers, 2014). Die Informationsverarbeitungsgeschwindigkeit und die Arbeitsgedächtniskapazität gelten dabei als wesentlich bedeutsamere kognitive Fähigkeiten als komplexere Konstrukte wie zum Beispiel das logische Schlussfolgern (Lehrl et al., 2016). Empirische Untersuchungen zur Entwicklung von Intelligenz im Verlauf der Lebenszeit weisen darauf hin, dass ein Rückgang der kognitiven Verarbeitungsgeschwindigkeit mit der Abnahme der fluiden Intelligenz (nach Horn & Cattell, 1966) korreliert. Es bestehen jedoch keine Zusammenhänge zwischen Verarbeitungsgeschwindigkeit und kristalliner Intelligenz (Finkel et al., 2007). Auch wenn dies als Bestätigung der Relevanz von Verarbeitungsgeschwindigkeit angesehen wird, gilt die mentale Geschwindigkeit als ein deutlich komplexeres und noch nicht vollends beschriebenes Konstrukt (Nettelbeck, 2011). Empirische Untersuchungen von Süß et al. (2002) zur Arbeitsgedächtniskapazität ergaben, dass diese als zentrale Komponente der Intelligenz den einzigen limitierenden Faktor für das logische Schlussfolgern abbildet. Ackerman et al. (2005) stellten diese Annahme jedoch in Frage, da sie in ihrer Metaanalyse lediglich eine mittlere, messfehlerbereinigte Korrelation zwischen Intelligenz und Arbeitsgedächtnis ($r = 0.48$) feststellen konnten. Doch auch diese Ergebnisse gelten als nicht unumstritten (z.B. Kane et al., 2005) und werden weiterhin diskutiert (Burgoyne et al., 2023).

Intelligenz gilt als einer der wichtigsten Prädiktoren für schulischen Erfolg (Deary et al., 2007). Neben der Intelligenz hängt der Erfolg eines Lernprozesses und die Entwicklung von Wissen und Kompetenzen beim Lernenden noch von weiteren individuellen Lernendenvariablen ab. Insbesondere Vorwissen wird als entscheidender Faktor neben der Intelligenz für die Interpretation von individuellen Unterschieden im Lernerfolg betrachtet (Stern, 2001). Vorwissen schließt dabei sowohl die Fachinhalte als auch die metakognitiven Lern- und Problemlösestrategien mit ein (Gruber & Stamouli, 2015). Daher werden in der vorliegenden Medienvergleichsstudie sowohl die kognitiven Fähigkeiten als auch das Vorwissen der Proband*innen als Kontrollvariablen erhoben werden, um ihren Einfluss auf den gemessenen Lernerfolg in Abhängigkeit des Modelltyps zu berücksichtigen.

3. Forschungsstand zu Modelltypen in Medienvergleichsstudien

- Medienvergleichsstudien zu Modellen haben das Problem eines unfairen Vergleichs. Bisher existieren keine Studien, die eine gleichwertige Modellqualität gewährleisten.
- Metaanalysen zum Vergleich von Animationen mit anderen Modelltypen, zeigen Vorteile von AR in Bezug auf Fachwissenszuwachs, Motivation und Entwicklung konzeptueller Kompetenzen auf:
 - In den Naturwissenschaften fallen die Effekte geringer aus als im Sprachenlernen.
 - Einige Metaanalysen weisen jedoch einen Publication Bias auf.
 - Die Metaanalysen basieren auf dem Vergleich AR vs. Non-AR. Die Betrachtung der in den Vergleichsgruppen genutzte Medien als Moderatorvariablen bleibt aus.
- Vergleichsstudien zur Lernwirksamkeit unterschiedlicher Modelltypen im Chemieunterricht zeigen ein uneinheitliches Bild.
- Bislang gibt es keine empirischen Untersuchungen zum Einfluss des Modelltyps auf das Modellverständnis.

3.1 Medienvergleichsstudien

Die Anzahl an Medienvergleichsstudien in den naturwissenschaftsdidaktischen Disziplinen steigt in den letzten Jahren immer weiter an (Buchner & Kerres, 2023; Garzón et al., 2019). Die Kritik von Morrison, Tversky und Bétrancourt, dass der überwiegende Teil dieser Studien eine mangelnde Vergleichbarkeit der Versuchsgruppen aufweist (Morrison et al., 2000; Tversky et al., 2002), ist auch heute noch aktuell (Buchner & Kerres, 2023). Neben den miteinander verglichenen Medien unterscheiden sich die Versuchsgruppen noch in weiteren Variablen. Diese zusätzlichen Unterschiede beziehen sich sowohl auf den Umfang der in den Modellen dargestellten Inhalte (Morrison, Tversky & Bétrancourt, 2000; z.B. bei Ardac & Akaygun, 2004) als auch auf den Grad der Interaktivität der Modelle (Tversky et al., 2002). Weiterhin kam es bei einer Vielzahl der Studien zu Nebeneffekten anderer unabhängiger Variablen, die nicht mit einem geeigneten Studiendesign voneinander abgegrenzt wurden. Beispielsweise wurde neben einem Wechsel des Mediums auch die Darstellung der Räumlichkeit in Versuchs- und Kontrollgruppe unterschiedlich abgebildet (z.B. Urhahne, Nick & Schanze, 2009; Özmen, 2011b) oder die genutzte Sozialform variiert (Barak & Dori, 2005; Q. Liu et al., 2022). In ihrem Systematic Review konstatieren Buchner & Kerres (2023), dass für den Großteil aktueller AR-Vergleichsstudien auch weiterhin das Problem jenes unfairen Vergleichs besteht. Ebenfalls gilt als problematisch dass die Kontrollgruppen zumeist weniger aufwendig durchgeplante, didaktische Konzeptionen aufweisen als die Experimentalgruppen (Buchner & Kerres, 2023; Tversky et al., 2002). Bei Medienvergleichen mit AR liegt zudem gerade im Hinblick auf den mäßigen Beitrag zur Theoriebildung (Buchner & Kerres, 2023) das Problem eines Publication Bias vor (Cao & Yu, 2023; Garzón & Acevedo, 2019). So weisen Garzón & Acevedo (2019) in ihrer Metastudie darauf hin, dass Studienergebnisse, die keinen signifikanten Unterschied zwischen den Vergleichsgruppen aufzeigen, weniger häufig publiziert werden. Im

Folgendes wird der aktuelle Forschungsstand zu Medienvergleichsstudien im Bildungskontext mit Fokus auf den naturwissenschaftlichen Unterricht dargestellt. Aus der aufgeführten Kritik am Forschungsfeld der Medienvergleichsstudien ergibt sich, dass die aktuelle Studienlage jedoch nur als eingeschränkt interpretier- und generalisierbar angesehen werden kann.

Als eine moderne Repräsentationsform digitaler Animationen rücken AR-Anwendungen aktuell immer weiter in den Fokus der Bildungsforschung (Buchner & Kerres, 2023). Die in den Metastudien zu Medienvergleichen mit AR gefundenen Overall-Effekte bezüglich einer gesteigerten Lernwirksamkeit von AR weisen eine Spanne von kleinen bis großen ($d = 0.49 - 0.85$) Effektstärken auf (Cao & Yu, 2023; Chang et al., 2022; Garzón et al., 2019; Garzón & Acevedo, 2019; Kalemkuş & Kalemkuş, 2022; Tekedere & Göke, 2016; Xu et al., 2022). Da in diesen Analysen weder das Auftreten weiterer lernförderlicher Effekte (z.B. der Split-Attention-Effekt) noch die Art der AR-Anwendung (z.B. markerbasiert oder standortbasiert) oder der dargestellten Objekte (z.B. Text oder Bild) für die Wahl der Stichprobe berücksichtigt wurden, sind die Ergebnisse nicht uneingeschränkt als generalisierbar zu betrachten. Zudem wurde der Einfluss weiterer Wirkgrößen wie Selbstwirksamkeit, Motivation oder Selbstregulation nicht berücksichtigt (Tschiersch et al., 2021). Eine domänenübergreifende, gesteigerte Lernwirksamkeit von AR-Anwendungen im Vergleich zu Non-AR ($d = 0.68$) konnten Tekedere und Göke (2016) in einer Metaanalyse über 15 von 171 Medienvergleichsstudien feststellen. Garzón et al. (2019) stellten in einer vergleichbar groß angelegten Metaanalyse mit insgesamt 27 unabhängigen Stichproben ebenfalls eine höhere Lerneffektivität für AR im Vergleich zu Non-AR Lernumgebungen fest ($d = 0.64$). Unter Berücksichtigung einer Gesamtzahl von 62 unabhängigen Stichproben konnte dieser Effekt ($d = 0.71$) bestätigt werden (Garzón & Acevedo, 2019). Der größte bisher berichtete Effekt ($d = 0.85$) ist in der Metaanalyse von Cao und Yu (2023) mit einer Gesamtstichprobe von 2926 Proband*innen zu finden. Gleichzeitig berichten die Autoren jedoch von einem Publication Bias in Bezug auf den Vergleich des Lernerfolgs (Cao & Yu, 2023). Auch wenn der Großteil der in die Metaanalysen eingeflossenen Studien aus dem naturwissenschaftsdidaktischen Bereich stammen, gilt es für die Interpretation der Ergebnisse den Einfluss der Domäne als Moderatorvariable berücksichtigen. So zeigt die Metaanalyse von Chang et al. (2022), dass die mittlere Effektgröße (d) für Studien über AR zum Sprachenlernen um 0.56 Punkte höher ist als die mittlere Effektstärke (d) bei AR-Studien zum naturwissenschaftlichen Lernen.

Xu et al. (2022) untersuchten in einer Metaanalyse die Lernwirksamkeit von AR im Vergleich zu Interventionen ohne AR-Einsatz im Kontext naturwissenschaftlicher Lernsettings. Für die Stichprobe von 35 Studien mit 39 gefundenen Effekten ergab sich ein mittlerer Effekt zugunsten von AR ($d = 0.74$). Die Analyse auf Moderatorvariablen zeigte marginale Effekte für Einfluss des AR-Typs (markerbasiert, markerlos oder standortbasiert) und der Fachrichtung,

während Bildungsstufen, Anzeigegeräte, Interventionsdauer, Gruppengröße und verwendete Lernstrategien keinen signifikanten Einfluss aufwiesen (Xu et al., 2022). Diese Metaanalyse ist dahingehend als limitiert zu betrachten, dass die Art der Intervention bzw. die Modellierungstätigkeiten in den Vergleichsgruppen weder bei der Auswahl der Stichprobe der Metaanalyse noch als Moderatorvariable berücksichtigt wurden. Folglich lässt diese Analyse keine Aussagen über den konkreten Vergleich von AR mit einem spezifischen anderen Medium (z.B. Illustration oder Desktop-Anwendung) zu. Für den Vergleich von Desktop-Anwendungen und AR-Anwendungen im naturwissenschaftlichen Unterricht zeigt sich in der Metaanalyse von (Kalemkuş & Kalemkuş, 2022) einer mittlerer Overall Effekt der Lernwirksamkeit zugunsten von AR ($d = 0.64$).

3.2 Der Einfluss des Modelltyps auf den Lernprozess

Die Studienlage zum Einfluss des Modelltyps¹ (Animation, Illustration, haptisches Modell) auf den Lernerfolg weist stark unterschiedliche Ergebnisse auf (Bétrancourt & Tversky, 2000). In der Mehrheit der Vergleichsstudien konnte jedoch eine höhere Lernwirksamkeit beim Umgang mit digitalen Lernanimationen im Vergleich zur Verwendung illustrativer Modelle nachgewiesen werden (Höffler & Leutner, 2007). In der Metaanalyse von Höffler & Leutner (2007) konnte mit insgesamt 76 paarweisen Vergleichen dynamischer und statischer Visualisierungen ein mittlerer Gesamtvorteil der Lehranimationen gegenüber statischen Illustrationen ($d = 0.37$) festgestellt werden. Moderatoranalysen deuten auf noch größere Effektgrößen hin, wenn die Animation eher repräsentativ als dekorativ ist ($d = 0.40$) und deklaratives Wissen ($d = 0.44$) anstelle von problemlösendem Wissen ($d = 0.24$) abgefragt wurde. In der Metaanalyse von (Höffler, 2008) zeigten computerbasierte Animationen in MINT-Kontexten eine bessere Lernwirksamkeit als statische Bilder ($d = 0.40$). Dieser Effekt fällt jedoch geringer aus, wenn eine größere Anzahl an Standbildern in den Illustrationen dargestellt wird. So zeigte sich für Animationen ein höherer Lernzuwachs im Vergleich zu Illustrationen, die aus vier Abbildungen bestehen ($d = 0.55$), jedoch kein signifikanter Unterschied zwischen Bildreihen aus elf Bildern und Animationen. Eine wichtige Moderator-Variable zeichnet sich hierbei durch die Domäne ab. So zeigen sich Animationen zu chemischen Fachinhalten ($d = 0.77$) deutlich lernwirksamer als der Einsatz von Animationen in den Fächern Biologie ($d = 0.29$), Physik ($d = 0.23$) oder Mathematik ($d = 0.22$).

¹Bétrancourt & Tversky verwenden im Originaltext den Begriff „forme de représentation“. Da der Begriff Repräsentationsform im deutschsprachigen Raum auch Unterschiede zwischen Abbildungen wie Diagrammen oder piktoralen Darstellungen abdeckt, wird im Rahmen dieser Arbeit der Begriff Modelltyp verwendet.

Die Lernwirksamkeit von unterschiedlichen Modelltypen im Zusammenhang mit räumlichem Vorstellungsvermögen in chemischen Kontexten wurde bereits umfassend untersucht (z.B. Abraham et al., 2010; Barrett et al., 2015; Behmke et al., 2018; Stull & Hegarty, 2016). In Bezug auf den Fachwissenserwerb zum Aufbau und der Struktur organischer Verbindungen zeigen sich AR-Modelle in einigen Studien als ebenso lernwirksam wie ihre haptischen Äquivalente (Barrett et al., 2015; Behmke et al., 2018; Stull & Hegarty, 2016). Jedoch zeigt sich in der Vergleichsstudie von Abraham et al. (2010), dass Computeranimationen im Vergleich zu haptischen Modellen und 2D-Illustrationen zu einem erhöhten Fachwissenszuwachs zur Stereochemie und einem gesteigerten räumlichen Denken führen. Auch Mazzuco et al. (2022) kommen in ihrem Systematic Review zum Schluss, dass AR Vorteile für den Wissenserwerb in dieser spezifischen Domäne bietet. In der Vergleichsstudie von Elford et al. (2022) konnten diese Unterschiede zwischen AR und 2D-Abbildungen jedoch nicht bestätigt werden.

Effekte auf den Fachwissenserwerb zur Deutung von Experimenten auf Teilchenebene sind bislang wenig untersucht. Während einige Studien AR als erfolgsversprechende Alternative zu vergleichbaren Computeranimationen herausstellten (z.B. Ibáñez et al., 2014), zeigen aktuellere Untersuchungen diesbezüglich keine signifikanten Unterschiede (Peeters et al., 2023; Probst et al., 2022). Die Ergebnisse von Peeters et al. (2023) sowie von Keller, Habig & Rumann (2020) weisen jedoch darauf hin, dass die kognitive Belastung von AR-Animationen und somit der Lernerfolg in diesem Kontext variieren kann. Die stark unterschiedlichen Ergebnisse der Medienvergleiche zum Lernerfolg unter Berücksichtigung der kognitiven Belastung (vgl. Tabelle 55, z.B. Tversky et al., 2002; Höffler & Schwarz, 2011) lassen darauf schließen, dass das Konstrukt Cognitive Load (CL) als abhängige Variable eher ungeeignet ist. Vielmehr sollte die kognitive Belastung als Moderator- bzw. Kontrollvariable überprüft werden und für alle Vergleichsgruppen auf gleichem Niveau liegen. So können Illustrationen, die im Hinblick auf den CL optimiert wurden, auch dann zu mehr Lernzuwachs als Animationen führen, wenn diese nach dem Multimedia-Prinzip den auditiven und optischen Kanal ansprechen (Mayer et al., 2005). Auch der Autonomiegrad der Lernumgebung wirkt sich auf den CL unterschiedlicher Modelltypen aus. So brachten Animationen in selbstgesteuerten Lernsettings weniger CL als Abbildungen hervor (Höffler & Schwarz, 2011). In systemgesteuerten Lernsettings brachten jedoch die Abbildungen bessere Ausprägungen der kognitiven Belastung hervor (Höffler & Schwarz, 2011). Bétrancourt und Tversky (2000) betonen in diesem Zusammenhang auch die Bedeutung der Usability als Kontrollvariable, um gleiche Lernchancen bei Medienvergleichsstudien zu gewährleisten. Beispielweise zeigen sich in der empirischen Vergleichsstudie von Barrett et al. (2015) haptische Modelle bei gleicher Usability als ebenso lernwirksam wie die konkurrierenden AR-Modelle. Der Einfluss der Usability auf den Lernerfolg und weitere Lernendenvariablen wird im Abschnitt 2.4 unter dem Punkt *System Usability* genauer erläutert.

Unter anderem führt eine schlechte Usability dazu, dass Lernende Frustration und negative Assoziationen gegenüber dem Lerngegenstand entwickeln können (Karapanos et al., 2018). In der Untersuchung von Harp & Mayer (1997) erwiesen sich deskriptive Illustrationen, die ohne Berücksichtigung von Usability-Kriterien entwickelt wurden, weder im Hinblick auf emotionales noch auf kognitives Interesse als vorteilhaft gegenüber dem Arbeiten mit erklärenden Texten. Somit ist sowohl ein direkter Einfluss der genutzten Modelltypen als auch ihrer Usability auf affektive und kognitive Konstrukte zu erwarten. Garzón et al. (2019) stellen in ihrem Systematic Review motivationale Effekte als die am zweithäufigsten berichteten Vorteile von AR heraus. Radu (2014) beschreibt in einem Comparative Systematic Review, dass Schüler*innen vermehrt von Spaß beim Lernen mit AR berichten und häufiger bereit sind, die AR-Erfahrung zu wiederholen. Zudem schreiben Lernende bereits bevor sie AR-Anwendungen im Lernkontext genutzt haben AR motivierende Effekte im Lernprozess zu (Cheng, 2018). Die Metaanalyse von Yu (2023) stützt dieses Bild jedoch nur teilweise. Die Ergebnisse weisen eine signifikant verbesserte Einstellung der Lernenden gegenüber AR bei zeitgleichem Ausbleiben einer gesteigerten Motivation auf (Yu, 2023). Der direkte Einfluss affektiver Wirkgrößen auf den Lernerfolg von Interventionen (Köller et al., 2000; Köller et al., 2006) wird in Abschnitt 2.4 unter dem Punkt „*Affektive Konstrukte*“ ausführlicher erläutert. Daraus erschließt sich die Notwendigkeit der Erhebung affektiver Konstrukte als abhängige bzw. Kontrollvariablen in Medienvergleichen zu erheben.

Bislang existieren keine empirischen Studien, die auf kausale Zusammenhänge zum Einfluss des Modelltyps auf den Erwerb von Modellkompetenz bzw. Modellverständnis schließen lassen. Jedoch liegen empirische Befunde vor, dass dem Modellverständnis verwandte Konstrukte in Abhängigkeit zum Modelltyp stehen. So zeigen Studien zum selbstgesteuerten Lernen, dass Lernende beim Arbeiten mit Illustrationen bessere Problemlösestrategien nutzen als bei der Nutzung von Computersimulationen (z.B. Lewalter, 1997a). Zudem konnte nachgewiesen werden, dass Schüler*innen durch die Nutzung von Computersimulationen im Kontext der Modellmethode einen höheren Zuwachs an Problemlösefähigkeit (Ali et al., 2022; Williamson & Abraham, 1995) und kritischem Denken (Ali et al., 2022) verzeichnen konnten als Schüler*innen, die mit Illustrationen gearbeitet haben. Bevan (2023) berichtet in seinem Review über die Vorteile von AR in Bezug auf das meta-kognitive Konstrukt *Conceptual Knowledge*. Zudem stellte Cheng (2018) in einer Studie mit 267 Schüler*innen der Junior High School fest, dass ihre Einstellungen zum Lernen mit AR positiv mit den epistemischen Überzeugungen in Naturwissenschaften korrelieren. Auch Vergleichsstudien im Kontext von Chemieunterricht zeigen unter anderem zu den drei Ebenen chemischen Wissens nach Johnstone (1999) Vorteile von Animationen auf (z.B. Urhahne et al., 2009; Barak, 2013; s. Tabelle 55). Laut der Definition von Modellverständnis nach Leisner (2005a) setzt sich das Modellverständnis aus

epistemologischen Überzeugungen und Fähigkeiten zusammen. Dazu zählt auch die Unterscheidung zwischen Realität und Modell. In seiner Vergleichsstudie zum Einfluss des Modelltyps auf das chemische Fachwissen zu unterschiedlichen Atommodellen postuliert Saborowski (2000) den Vorteil von Animationen gegenüber Illustrationen aufgrund einer geringeren Verwechslungsgefahr von Modell mit einem vergrößerten Original. Als Ursache hierfür nennt Saborowski einen verstärkten Modellcharakter, da das Modell in der Animation lediglich simuliert werde.

Tabelle 5: Übersicht über den Forschungsstand zu Medienvergleichsstudien im Kontext von naturwissenschaftlichem Lernen (Anim = Animationen, IL = Illustrationen, AR = Augmented Reality, Hap = Haptisch, ECL = Extraneous Cognitive Load, selbstg. = selbstgesteuert, fremdg. = fremdgesteuert, PPP= PowerPointPräsentation, AC = Anorganische Chemie, OC = Organische Chemie)

Quellen	Fachwissen	Metakognitive Variablen	Motivation	Cognitive Load	Usability	Zielgruppe	Domäne
Williamson und Abraham (1995)	Anim > IL	<u>Bildung mentaler Modelle:</u> Anim > IL				Studierende	Aggregatzustände, Diffusion Lösung
Lewalter (1997a)		<u>Problemlösen:</u> IL > Anim		ECL: IL > Anim		Studierende	Astrophysik
Saborowski (2000)	Anim > IL					Klasse 7; 8 & 9	Kugelpackung, Atomhüllenmodell, Orbitalmodell
Bétrancourt und Tversky (2000)	Anim > IL				Anim = IL	Review	Review
Tversky et al. (2002)	IL > Anim (Wenn Anima > IL, dann weil mehr Informationen enthalten)			ECL: Anim > IL			Review
Ardac und Akaygun (2004)	Anim > IL					8. Klasse	Teilchenmodell Reaktionen AC
Mayer et al. (2005)	IL > Anim					Studierende	Alltagsprozesse Makroebene
Barak und Dori (2005)	Anim > IL	<u>Conceptual Understanding:</u> Anim > IL (Ungleicher Vergleich da Projektunterricht)				Studierende	Reaktionen AC
Chen (2006)			Hap > AR			Studierende	OC Aminosäuren
Hornecker und Dünser (2007)			AR > IL			Grundschule	Lesen
Höffler und Leutner (2007)	Anim > IL					Metaanalyse	Metaanalyse
Höffler (2008)	Anim > IL; (Je mehr Bilder, desto					Metaanalyse	Metaanalyse

	geringer der Vorteil Je mehr Interaktion, desto geringer der Vorteil)						
Urhahne et al. (2009)	Anim = IL Anim = IL	<u>Conceptual knowledge:</u> Anim = IL Anim > IL		ECL: Anim = IL		Studierende 10. Klasse	Modifikation des Kohlenstoffs
Korakakis et al. (2009)			Anim > IL	ECL: Anim > IL		8. Klasse	Lösungen & Destillieren
Abraham et al. (2010)		<u>Spatial Ability:</u> PC-Modell > Hap Hap = IL (2D)				Studierende	OC Stereochemie
Özmen, (2011b)	Anim > Traditionell					Klasse 6	Teilchenmodell
Höffler und Schwartz (2011)	selbstg.: Anim > IL			selbstg. ECL: IL > Anim fremdg. ECL: Anim > IL		Studierende	Oberflächen & Waschmittel
Barak (2013)	Anim = IL	<u>Conceptual understanding:</u> Anim > IL	Anim > IL			Grundschule	Gold Eigenschaften und Verwendung
	Anim selbstg. > Anim. fremdg. = IL selbstg.	<u>Conceptual understanding:</u> Anim selbstg. = Anim fremdg. > IL selbstg.				12. Klasse	OC räumliche Struktur
		Conceptual understanding: Anim > IL <u>Repräsentationskompetenz:</u> Anim > IL				Studierende	OC räumliche Struktur
Ibáñez et al. (2014).	AR > Anim					12. Klasse	Elektromagnetis- mus
Barrett et al. (2015)	Hap = AR		AR > Haptisch	AR weniger CL als haptisch	Hap = AR	Studierende	OC räumliche Struktur
Chen et al. (2015)	AR > IL					10. Klasse	Orbitalmodell

Stull und Hegarty (2016)	Spatial Ability: Hap = Anim	Repräsentationskompetenz: Hap = Anim		Anim = Hap		Studierende	OC räumliche Struktur
Behmke et al. (2018)	Hap = AR		AR > Haptisch			Studierende	OC räumliche Struktur
Astuti et al. (2019)	AR > PPP	Problem solving: AR > PPP	AR > PPP				Pflanzengewebe
Keller, Habig und Rumann (2020)				AR weniger CL als Abbildungen		Studierende	OC Stereochemie
Probst et al. (2022)	AR = Tablet > Schulbuch		AR = Tablet > Schulbuch			Klasse 8	Teilchenmodell Diffusion
Xu et al. (2022)	AR > IL		AR > IL				Metaanalyse
Ali et al. (2022)		critical thinking: AR > Non-AR, problem-solving: AR > Non-AR					Review
Tarng et al. (2022)	AR > traditionell		AR > traditionell	AR = traditionell			AC Reaktionen Kalotten-Modelle
Elford et al. (2022)	3D AR = 2D IL			3D AR = 2D IL			OC Stereochemie
Silva et al. (2023)	3D AR > 2D IL		AR > Abbildungen				OC- Hybride des Kohlenstoffs
Lang (2023)	Anim = IL	Modellbildungskompetenz: Anim = IL				Klasse 8	Massernerhaltung
Peeters et al. (2023)	AR = Anim = Video			ECL: AR > Anim & Video			Lösung von sauren Salzen

4. Ableitung der Forschungsfragen und Hypothesen

Studie I: Validierung der Gütekriterien im Zuge der Entwicklung eines Messinstruments zu Bewertung der Modellqualität

In der naturwissenschaftsdidaktischen Literatur existieren bislang keine validierten und quantifizierbaren Gütekriterien, mit deren Hilfe sich Anschauungsmodelle bewerten lassen. Eine Bewertung der Modellqualität ist jedoch notwendig, um einen fairen Vergleich unterschiedlicher Modelltypen zu ermöglichen (Morrison et al., 2000; Przywarra et al., 2021; Przywarra & Risch, 2022; Tversky et al., 2002). Daraus ergibt sich das Desiderat theoriegeleitet normative und quantifizierbare Kriterien zur Bewertung der Modellqualität zu deduzieren. In informativ-scher Primärliteratur lassen sich detaillierte Beschreibungen und Definitionen dieser Kriterien finden. Die ermittelten Gütekriterien spiegeln sich in den Kriterien von Eschenhagen et al. (2008) wider. Um sicherzustellen, dass die Expert*innen die Bewertung der Modelle auf Grundlage dieses fundierten Maßstabs durchführen, ergibt sich folgende Forschungsfrage:

*FF1 Inwiefern stimmen die Aussagen und Einschätzungen von Expert*innen zu den Gütekriterien für Anschauungsmodelle in der Naturwissenschaftsdidaktik mit den Beschreibungen in der Primärliteratur überein?*

Daraus ergibt sich folgende Hypothese, die mittels Expert*inneninterviews und anschließender qualitativer Inhaltsanalyse überprüft wird:

*H1 Die Antworten von Expert*innen zu den entwickelten Items entsprechen inhaltlich den Beschreibungen der Gütekriterien aus der Primärliteratur.*

Die Hypothese H1 basiert auf der Annahme, dass Expert*innen aufgrund ihres Fachwissens und ihrer Erfahrung in der Lage sind, die Gütekriterien für Modelle angemessen zu beurteilen. Durch die Überprüfung dieser Hypothese wird nicht nur die Konsistenz zwischen den Antworten der Expert*innen und den Beschreibungen der Gütekriterien in der Literatur überprüft, es wird auch die Validität der Expert*inneneinschätzungen bestätigt. Wird die Hypothese verifiziert, weist dies daraufhin, dass die Expert*innen eine fundierte und konsistente Vorstellung von den Gütekriterien haben und dass ihre Einschätzungen zur Modellbewertung auf einer fundierten Grundlage beruhen.

Studie II: Gütebestimmung der Qualität der drei Modelltypen HI, AR und IL mittels Expert*innenratings

Auf Grundlage des validierten Messinstruments wird es möglich die Qualität unterschiedlicher Typen von Anschauungsmodellen mittels Expert*innenrating bewerten zu lassen (Przywarra et al., 2021; Przywarra & Risch, 2022). Dies ermöglicht die Sicherstellung eines fairen Vergleichs zwischen den unterschiedlichen Gruppen in Medienvergleichsstudien. Es herrscht ein Mangel an Publikationen zu empirischen Medienvergleichsstudien, die den Einfluss des Modelltyps auf die Lernprozess untersuchen, wobei lediglich das Medium als einzige Modalität variiert und die qualitative Vergleichbarkeit empirisch gesichert wird (Buchner & Kerres, 2023; Morrison et al., 2000; Tversky et al., 2002). Angesichts dieser Forschungslücke ergibt sich das Desiderat, den Einfluss des Modelltyps auf kognitive und affektive Lernendenvariablen unter Einhaltung dieser Vergleichbarkeit zu untersuchen. Mithilfe einer Evaluation durch Expert*innenrating werden die Modelle der drei in vorliegender Arbeit verwendeten Modelltypen Augmented Reality, haptisch-interaktiv und illustrativ hinsichtlich ihrer Qualität bewertet. Um die Gleichwertigkeit der Modelle zu gewährleisten, soll folgende Forschungsfrage beantwortet werden:

FF2 Welche Qualität weisen die Modelle der drei Typen Augmented Reality, haptisch-interaktiv und illustrativ auf?

Die einzelnen Modelle der verschiedenen Modelltypen wurden nach dem aktuellen Forschungsstand bestmöglich und gezielt zum Erreichen zuvor definierter Lernziele konzipiert. Darüber hinaus wurde im Zuge der Entwicklung darauf geachtet, dass die Modelle vergleichbare Interaktivität aufweisen, um den Lernprozess gleichwertig zu unterstützen. Daraus ergibt sich folgende Hypothese, die anhand von Unterschiedsprüfungen zwischen den Mittelwerten der Gesamtgüte der drei Vergleichsgruppen überprüft wird:

H2 Die Gesamtgüte der Modelle unterscheidet sich nicht signifikant in Abhängigkeit des Modelltyps.

Durch die Überprüfung der Hypothese H2 wird nicht nur die Konsistenz zwischen den Zielen der Modellentwicklung und den festgelegten Lernzielen überprüft, sondern auch die Qualität der Umsetzung dieser Ziele in den Modellen bestätigt. Wird die Hypothese verifiziert, deutet dies darauf hin, dass die Modelltypen die gleichen Voraussetzungen bieten, die definierten Lernziele zu erreichen. Dadurch wird ein fairer Vergleich innerhalb einer Interventionsstudie erst möglich.

Studie III: Interventionsstudie zum Einfluss der drei Modelltypen HI, AR und auf kognitiv-affektive Lernendenvariablen

Weiterhin leistet die vorliegende Arbeit zu diesem Forschungsdesiderat einen Beitrag in Form einer Interventionsstudie zum Einfluss unterschiedlicher Modelltypen auf den Zuwachs von Fachwissen, Modellverständnis und das Interesse im chemischen Anfangsunterricht. Daraus leitet sich folgende Forschungsfrage ab:

*FF3 Wie wirken sich die unterschiedlichen Modelltypen auf das situationale Interesse, das Modellverständnis und das Fachwissen von Schüler*innen der Klassenstufe acht aus?*

Auf Basis der Expectancy-Value-Theorie wird bei der Zielgruppe ein hoher zugeschriebener „Intrinsic Value“ für die AR-Modelle erwartet (Eccles & Wigfield, 2020), da Kinder und Jugendliche digitale Medien hauptsächlich zur Unterhaltung nutzen (Medienpädagogischer Forschungsverbund Südwest, 2022). Ferner berichtet Cheng (2018), dass Lernende bereits, bevor sie AR-Anwendungen im Lernkontext genutzt haben, AR motivierende Effekte im Lernprozess zuschreiben. Zudem weisen Schüler*innen bei der Nutzung von AR in selbstgesteuerten Lernsettings aufgrund eines höheren Kompetenzerlebens im Umgang mit digitalen Medien (Hillmayr et al. 2020) auch über den Neuheitseffekt hinweg eine gesteigerte Motivation auf (Sofianidis, 2022). Daraus ergeben sich für die vorliegende Medienvergleichsstudie folgende Hypothesen:

H3 Die Entwicklung des situationalen Interesses der einzelnen Gruppen (AR, HI & IL) unterscheidet sich über den Interventionszeitraum hinweg.

H3.1 Die Lernenden zeigen während der Intervention ein höheres situationales Interesse im Umgang mit den AR-Modellen als im Umgang mit den haptisch-interaktiven Modellen.

H3.2 Die Lernenden zeigen während der Intervention ein höheres situationales Interesse im Umgang mit den AR-Modellen als im Umgang mit den illustrativen Modellen.

Die Überprüfung der Hypothese H3 ermöglicht es festzustellen, ob sich das situationale Interesse der verschiedenen Gruppen im Laufe der Intervention unterscheidet. Dies ist wichtig, um zu verstehen, ob die Intervention einen Einfluss auf das situationale Interesse der Lernenden hat. Zudem ermöglicht dies festzustellen, ob zu den unterschiedlichen Modelltypen ein Neuheitseffekt vorliegt. Hypothese H3.1 und H3.2 ermöglichen eine genauere Differenzierung, indem sie spezifische Vergleiche zwischen den Gruppen betrachten. Diese Vergleiche sind wichtig, um festzustellen, welche Art von Modellen möglicherweise einen stärkeren Einfluss auf das situationale Interesse der Lernenden hat. Insgesamt ermöglicht die Überprüfung dieser Hypothesen eine detaillierte Analyse der Entwicklung des situationalen Interesses der

einzelnen Gruppen im Rahmen der Intervention und kann wichtige Erkenntnisse über die Wirksamkeit der einzelnen Modelltypen liefern.

Eine Vielzahl an empirischen Studien und Metaanalysen zeigt, dass digitale Medien wie Animationen und AR dazu beitragen können, komplexe Konzepte besser zu vermitteln und das Lernen zu erleichtern (vgl. Abschnitt 2.2). Dabei zeigen sich Animationen besonders lernwirksam in Bezug auf deklaratives Wissen (Höffler & Leutner, 2007). Daher ergeben sich für den Einfluss des Modelltyps auf die Entwicklung des Fachwissens folgende Hypothesen:

H4 Das Fachwissen der verschiedenen Gruppen (AR, HI, IL) entwickelt sich über die einzelnen Messzeitpunkte unterschiedlich stark.

*H4.1 Das Fachwissen der Schüler*innen in der Gruppe AR nimmt mehr zu als in der Gruppe HI.*

*H4.2 Das Fachwissen der Schüler*innen in der Gruppe AR nimmt mehr zu als das der IL.*

Bislang existieren keine empirischen Untersuchungen zum Einfluss des Modelltyps auf den Erwerb von Modellverständnis. Verschiedene Modelltypen führen jedoch bei Lernenden zu unterschiedlichen Zuwächsen an epistemologischen Fähigkeiten (Bevan, 2023; Urhahne et al., 2009). Für den Einfluss auf die Entwicklung des Modellverständnisses als epistemologischen Aspekt der Modellkompetenz gilt es zu untersuchen, ob sich diese ebenfalls in Abhängigkeit vom Modelltyp unterscheidet. Ein möglicher Grund für einen Unterschied könnte in einer verbesserten Trennung zwischen Modell und Realität liegen (Saborowski, 2000). Da die Studienlage zur Entwicklung epistemologischer Fähigkeiten Vorteile von Animationen (Urhahne et al., 2009) bzw. AR (Bevan, 2023) gegenüber anderen Modelltypen aufzeigt, ergeben sich folgende Hypothesen zum Einfluss auf das Modellverständnis:

H5 Das Modellverständnis der verschiedenen Gruppen (AR, HI, IL) entwickelt sich über die einzelnen Messzeitpunkte unterschiedlich stark.

*H5.1 Das Modellverständnis der Schüler*innen in der Gruppe AR nimmt mehr zu als in der Gruppe HI.*

*H5.2 Das Modellverständnis der Schüler*innen in der Gruppe AR nimmt mehr zu als in der Gruppe IL.*

Die Überprüfung der Hypothesen H4 und H5 ermöglicht es festzustellen, ob und wie sich das Verständnis der Lernenden über den Verlauf der Intervention verändert. Zudem erlaubt dies eine Prüfung auf Unterschiede zwischen den Gruppen. Die Hypothesen H4.1 und H4.2 sowie H5.1 und H5.2 ermöglichen eine detailliertere Analyse dahingehend, ob eine bestimmte Gruppe möglicherweise ihr Modellverständnis stärker entwickelt als andere.

5. Konzeption der Lernumgebung

In diesem Abschnitt wird die Konzeption der verschiedenen Interventionen zu den drei Modelltypen der Medienvergleichsstudie beleuchtet. Zunächst werden Lernziele formuliert und in ein Curriculum eingeordnet. Daraufhin wird die Konzeption der Lernumgebungen und Experimente auf Basis dieser Lernziele vorgestellt. Dadurch wird die generelle Lernwirksamkeit der Interventionen sichergestellt. Anschließend wird die Konzeption der drei unterschiedlichen Modelltypen präsentiert. Die Formulierung klarer Ziele ermöglicht es dabei, die Wirksamkeit der Interventionen Modelle genauer abzustimmen. Zudem wurden die Experimente und Modelle spezifisch mit dem Ziel der Beantwortung der Forschungsfragen der vorliegenden Arbeit neu konzipiert. Dieses Kapitel legt den Fokus auf die Umsetzung empirisch fundierter Designkriterien zur Gestaltung von Lernumgebungen und Modellen, die spezifisch zum Erreichen der Lernziele in die Konzeption der Interventionen mit eingeflossen sind. Dabei sind sämtliche Kriterien berücksichtigt worden, die in Abschnitt 4.1 aufgeführt werden. Darüber hinaus wird ein Ansatz zur Kontrolle von Störvariablen aufgezeigt und diskutiert, um eine Verzerrung des Medienvergleichs zu vermeiden. Die kriteriengeleitete Konzeption der Modelle bietet somit zusätzlich den Vorteil, die interne Validität der Studie sicherzustellen.

5.1 Kriterien zur Konzeption der Lernmaterialien

- Die Förderung von Modellkompetenz bedarf der Berücksichtigung von Vorwissen und des Einsatzes von Conceptual Change Tasks (CCT).
- Die Reflexion theoretischer Annahmen und die Integration in lebensweltliche Kontexte sind entscheidend, um angemessene fachliche Vorstellungen zu entwickeln und inkorrekte Präkonzepte zu überwinden.
- Eine Reflexion über bestehende Vorstellungen, die Einbindung in Lebenskontexte und die Entwicklung von Problemlösefähigkeiten sind entscheidend, um angemessene Vorstellungen zu entwickeln und inkorrekte Präkonzepte zu überwinden.
- Die Rekonstruktion von Präkonzepten erfordert Unzufriedenheit mit vorhandenem Wissen, Verständlichkeit und Plausibilität neuer Vorstellungen sowie die Fruchtbarkeit neuer Konzepte für die Lösung von Problemen.
- Das Anfertigen von Schüler*innen-Illustrationen zu ihren Teilchenvorstellungen hilft unzureichendes Verständnis von Modelldarstellungen aufgrund mangelnder Repräsentationskompetenz zu überwinden.
- Die Förderung von Teilkompetenzen der Modellkompetenz sollte selektiv erfolgen, wobei die Trennung von Erfahrungswelt und Modellwelt sowie eine abschließende Modellkritik unterstützend sind.

Das Denken in Modellen basiert auf einer Verknüpfung von intuitivem Wissen mit wissenschaftlich fundierten Konzepten (Lehrer et al., 1994). Aus diesem Grund sehen sowohl Terzer und Upmeyer zu Belzen (2007) als auch Leisner (2005a) das Vorwissen als bedeutend für den Aufbau von Modellwissen und Modellkompetenz an. Wenn naive Präkonzepte als intuitive Denkmodelle im Alltag ausreichende Erklärungen liefern und nicht durch wissenschaftlich angemessene Sichtweisen auf Modelle erweitert werden, kann dieses Wissen

kompartimentalisiert und nicht effizient genutzt werden. Daher empfehlen Terzer und Upmeyer zu Belzen (2007) zur Entwicklung der Modellkompetenz einen Conceptual-Change-Ansatz als graduellen Prozess (Clement, 2000).

Zur Gestaltung eines Conceptual Changes werden in den Naturwissenschaftsdidaktiken zwei alternative Ansätze verfolgt. Die erste Variante als Conceptual Changes Tasks (CCT) beginnt mit der Aufgabe, eine Vorhersage zu einem Phänomen zu geben (Chambers & Andre, 1997; Yürük, 2007). Der Zweck dieser Frage besteht darin, die Vorkenntnisse der Schülerinnen und Schüler zu aktivieren. Im nächsten Schritt wird von den Lernenden eine verbreitete, alternative Vorstellung identifiziert und durch die Präsentation von Informationen, die die Inkonsistenz zwischen der alternativen Vorstellung und der wissenschaftlichen Vorstellung aufzeigen, in Frage gestellt. Schließlich werden den Schüler*innen die wissenschaftlich akzeptierten Vorstellungen mithilfe von Beispielen, Abbildungen, Erklärungen und Situationen aus dem Alltag vermittelt (Chambers & Andre, 1997). Die alternative Variante erklärt eine verbreitete alternative Vorstellung und widerlegt sie direkt, durch Präsentation des wissenschaftlich akzeptablen Konzepts, ohne die Schülerinnen und Schüler um eine Vorhersage zur gegebenen Situation zu bitten (Yürük, 2007). Diese Aufgaben werden als "refutational tasks" (RT) bezeichnet. Der Hauptunterschied zwischen CCT und RT besteht in der Aufforderung der Schüler*innen zur Formulierung einer Vorhersage über die Situation. Die Konfrontation von Schüler*innen mit Vorstellungen, die der eigenen Vorstellung widersprechen können, zeigte sich am Beispiel von Nature of Science (NOS)-Vorstellungen als vorteilhaft für die Ausbildung einer NOS-Kompetenz (Hammerich, 1998). Aufgrund der Analogien in der Operationalisierung der Konstrukte NOS und Modellverständnis, wird für eine Entwicklung des Modellverständnisses der CCT-Ansatz als vorteilhaft angesehen.

Zum gleichzeitigen Lernen mit und über Modelle hat sich die Modellmethode etabliert (Kircher, 1995, Brademann, 1997, Leisner, 2005a; Kircher & Dittmer, 2004). Zur Anwendung der Modellmethode für Phänomene, deren Erklärung einen Wechsel auf die submikroskopische Ebene erfordern, gibt es methodische Ansätze mit unterschiedlichem Fokus auf die einzelnen Aspekte der Modellmethode (Kircher & Dittmer, 2004). Die grundlegenden Aspekte dieser Ansätze spiegeln sich jedoch stets in denselben fünf Schritten wider (Leisner, 2005a):

1. Vertraut werden mit dem Phänomen / Beobachtung des Phänomens
2. Erkenntnis der eingeschränkten Aussagekraft direkter Untersuchungen des Originals zur Erklärung des Phänomens
3. Modellauswahl [in Ausnahmefällen Modellentwicklung, z.B. bei Gilbert und Justi (2016)]
4. Modellanwendung zur Problemlösung bzw. Beantwortung der Fragestellung

5. Modellkritik in Form einer Reflexion der Zweckdienlichkeit, der Aussagekraft bzw. der Grenzen des Modells

Durch die Modellmethode werden die Lernenden beim Erwerb adäquater Modellvorstellungen unterstützt, da die enge Rahmung durch die einzelnen Schritte des Konzepts und die Auswahl guter Modelle fehlendes Wissen und Übung ausgleichen (Leisner, 2005a). Dabei ist es über den gesamten Prozess hinweg wichtig, die Erfahrungswelt auf Ebene der Phänomene von der Modellwelt auf Ebene der submikroskopischen Denkmodelle zu trennen (Mikelskis-Seifert & Leisner, 2003; Leisner, 2005a; Haas & Marohn, 2022). Weiterhin wird empfohlen die Trennung der Ebenen durch Piktogramme zu verdeutlichen (Haas & Marohn, 2022).

Caspari et al. (2018) betonen, dass im Kontext der Modellmethode das Anknüpfen an die Präkonzepte der Schüler*innen eine Schlüsselstelle zur Entwicklung der epistemologischen Fähigkeiten der Lernenden darstellt. Die Lehr- und Lernmöglichkeiten werden ebenso wie die Lernbedingungen durch ebendiese Präkonzepte bestimmt (Gropengießer & Kattmann, 2013). Im Zuge der didaktischen Gestaltung ist es von Vorteil, den Lernenden eine fachliche Rahmung zu geben. Der gewählte Rahmen wird dabei durch einen übergreifenden oder vorausgehenden Grundgedanken, wie etwa das Teilchenmodell als Theorie, verkörpert. Durch diese zusätzliche Information steigert sich zunächst die Komplexität des Lerngegenstands (ebd.). Diese Steigerung zeigt sich häufig als lernförderlich und schränkt das Lernen seltener ein (ebd.). Des Weiteren ist die Komplexität nötig, um im Bezugsrahmen von Fachwissenschaft und Lebenswirklichkeit die Konstruktion angemessener, fachlicher Vorstellungen zu fördern und inkorrekten Präkonzepten angemessen zu begegnen. Die Reflexion theoretischer Vorannahmen und gegenläufiger Auffassungen stellt hierbei eine notwendige Bedingung dar. Während die fachliche Rahmung fachbezogene Orientierung bietet und damit das Verstehen ermöglicht, bietet die Eingliederung in außerfachliche Kontexte Anschluss an die Lebenswelt der Schüler*innen (ebd.). Auf der Grundlage der fachlichen Rahmung und der Einbindung in lebensweltliche Kontexte erfolgt die Elementarisierung des Lerngegenstands. Die Ergebnisse der fachlichen Klärung werden dazu auf die wesentlichen Grundideen beschränkt. Der Unterrichtsgegenstand wird so zerlegt, dass Elemente entstehen, die das lebensweltliche Verständnis der Schüler*innen und die fachliche Grundidee in Einklang bringen (ebd.).

Durch die Konfrontation mit der fachlich korrekten Sichtweise werden die Präkonzepte geformt, wenn auch nicht zwangsläufig die fachwissenschaftlich richtige Anschauung adaptiert (Hammann & Asshoff, 2023). Den Lernenden soll dabei jedoch bewusstwerden, dass die neuen fachwissenschaftlichen Konzepte und Inhalte ihre eigenen Überzeugungen womöglich angemessener erklären als ihr bereits vorhandenes Verständnis. Neue Vorstellungen werden primär dann übernommen, wenn sie mit den durch eigene Erfahrungen gebildeten Vorstellungen übereinstimmen (Gropengießer & Kattmann 2013). Unterscheiden sich die vorhandenen

Konstrukte jedoch stark von den neu zu erlernenden, muss eine Vorstellungsänderung erfolgen. Damit es zu einer Veränderung der Vorstellungen von Schüler*innen kommt, müssen nach der Conceptual-Change-Theorie zunächst vier Bedingungen erfüllt sein (Duit, 1995a). Diese Bedingungen sind die Unzufriedenheit mit den existierenden Vorstellungen, die Verständlichkeit und die Plausibilität der neuen Vorstellung sowie die Fruchtbarkeit des neuen Konstrukts. Die Grundvoraussetzung zur Rekonstruktion der Vorstellungen ist dabei die Unzufriedenheit mit der bisherigen Vorstellung. Nur wenn Phänomene das Vertrauen in die alten Denkmuster negieren, sind Lernende bereit, eine neue Vorstellung anzunehmen (Krüger, 2007). Des Weiteren ist die Verständlichkeit von elementarer Bedeutung für einen erfolgreichen Conceptual Change (ebd.). Ist die neue Vorstellung nicht rational nachvollziehbar, können die Zusammenhänge, die mit ihr einhergehen, nicht von den Lernenden erfasst werden. Die Schüler*innen müssen dabei über ein bestimmtes Basiswissen verfügen, um überhaupt einen Zugang zu den neuen Gesichtspunkten zu finden. Die Bedingung der Plausibilität setzt Verständlichkeit voraus, ist aber von dieser abzugrenzen. Anomalien, die mit den existierenden Vorstellungen nicht zu deuten sind, sollten durch die neuen Vorstellungen erklärbar erscheinen (ebd.). Wie plausibel die neue Vorstellung erscheint, hängt vom Grad der Glaubwürdigkeit und der Übereinstimmung mit den vorherrschenden Denkweisen der Lernenden ab (ebd.). Die Akzeptanz der Erklärungen hängt von ihrer Vereinbarkeit mit den epistemologischen Überzeugungen ab (ebd.). Ebenso ist die Unwiderlegbarkeit der vermittelten wissenschaftlichen Konzepte entscheidend für die Übernahme der neuen Vorstellungen (ebd.). Die Fruchtbarkeit der neuen Vorstellungen knüpft an deren Verständlichkeit und Plausibilität an. Es ist wichtig, dass das erneuerte Konzept von den Schüler*innen zur Lösung weiterer, alltagsrelevanter Problemstellungen werden kann. Haben die Lernenden erfahren, dass die neue Vorstellung sowohl verständlich als auch plausibel neue Erfahrungen erklärt, werden sie auch zukünftig versuchen auftretende Phänomene mit diesem neuen Konzept zu erklären (ebd.). Resultieren aus diesen Problemlöseprozessen neue Erkenntnisse, erscheint dem Lernenden die neue Vorstellung als fruchtbar und er wird sie infolgedessen häufiger anwenden (ebd.). Durch diesen Prozess überprüfen die Lernenden die Viabilität ihrer eigenen Denkkonzepte und kognitiven Strukturen. Diese Überprüfung gilt als eine hinreichende Bedingung für die Initiierung einer Weiterentwicklung des Wissenssystems (von Aufschnaiter & von Aufschnaiter, 2005).

Um angemessene mentale Modelle aus (mehreren) externen Ressourcen zu generieren, benötigen Lernende eine gut entwickelte Repräsentationskompetenz. Es kann jedoch nicht davon ausgegangen werden, dass Lernende über diese Kompetenzen verfügen, insbesondere wenn sie Novizen in einem Lernbereich sind (Rau, 2017). Üblicherweise wird von Lernenden erwartet, dass sie mit externen, domänenspezifischen Darstellungen lernen, obwohl sie möglicherweise nicht imstande sind, diese angemessen in ihren Lernprozess zu integrieren. Sie

müssen in der Lage sein, externe Darstellungen zu interpretieren und relevante Informationen daraus auszuwählen und zu organisieren. In der fachdidaktischen Literatur wird dieser Konflikt als "Repräsentations-Dilemma" bezeichnet (ebd.). Das Anfertigen von Schüler*innen-Illustrationen zu ihren Teilchenvorstellungen wirkt sich positiv auf die Generierung mentaler Modelle zum Teilchencharakter der Materie aus (z.B. Sanger, 2000) und wirkt damit diesem Dilemma entgegen. Weiterhin gilt es in diesem Kontext als bedeutsam, dass Schüler*innen mit Modellen aktiv handeln, um diese als epistemische Werkzeuge effektiv nutzen zu können (NRC, 2012). Im Sinne der konstruktivistischen Lerntheorien werden ebenfalls die Aktivität und Eigentätigkeit der Lernenden als Voraussetzung für einen Conceptual Change genannt (Duit, 1995). Für Medienvergleichsstudien ist es daher wichtig, dass alle Modelltypen einen äquivalenten Grad an Interaktivität bieten, um einen fairen Vergleich zu ermöglichen (Tversky et al., 2002).

Eine simultane Förderung mehrerer Teilkompetenzen der Modellkompetenz sehen Fleige et al. (2012) als problematisch an. Um eine Überforderung der Lernenden zu vermeiden, sei eine Auswahl von Teilkompetenzen sinnvoll. Dadurch wird die Anzahl zu reflektierende Aspekte verringert und eine Förderung der Teilkompetenzen möglich. Auf Basis der Operationalisierung von Modellverständnis nach Leisner (2005a) gelten die Teilkompetenzen *Testen von Modellen* und *Eigenschaften von Modellen* (s. Abschnitt 2.3) als gut kombinierbar (Fleige et al., 2012). Das Konzept nach Mikelskis-Seifert und Leisner (2003) beziehungsweise Leisner-Bodenthin (2006) zur Förderung des Modellverständnisses sieht eine Trennung der Inhaltsebenen während der Intervention vor. Sie unterscheiden zwischen der „Erfahrungswelt – Die Wirklichkeit“ und der „Modellwelt – Das Ausgedachte“ (Mikelskis-Seifert & Leisner, 2003, S. 34). Zur Wirklichkeit zählen alle direkten Wahrnehmungen, wie Objekte und ihre physikalischen Eigenschaften. Die Modellwelt beinhaltet alle Annahmen und Vereinfachen sowie die entsprechenden (Denk-)Modelle. Auch Haas und Marohn (2022) nehmen eine Trennung in diese beiden gedanklichen Ebenen vor. Während beide Ansätze es vorsehen, die Ebenen im Arbeitsmaterial mittels Farbcodes voneinander zu trennen, ziehen Haas und Marohn (2022) zusätzlich noch die Verwendung von Piktogrammen zur Symbolisierung der Ebenen vor. Als eine weitere Ebene, die mit dem Testen von Modellen in Verbindung steht, weisen Barke et al. (2015) auf eine abschließende Modellkritik hin. Diese soll auf die Grenzen des Modells, sowie auf redundante Informationen innerhalb des Modells hinweisen (ebd.). Für eine erfolgreiche, abschließende Besprechung der Eigenschaften der Modelle und des Testens der Modelle sehen Fleige et al. (2012) eine schriftliche Sicherung von Merksätzen oder Ergebnissen besonders wichtig an.

Für die Darstellung von Prozessen in statischen, illustrativen Modellen gelten weitere spezielle Ansprüche. Zunächst sollten die Schritte zwischen den Bildern der Bildreihen nicht zu groß sein, damit sie für die Lernenden noch nachvollziehbar sind (Höffler, 2008). Um die

ablaufenden Prozesse verständlich darzustellen, benötigen die Bildreihen der Legebilder Pfeile und Highlightings (Mayer & Gallini, 1990). Dadurch wird es den Lernenden erleichtert den Fokus auf die wichtigsten Aspekte der Abbildungen zu setzen und den Veränderungen von Bild zu Bild zu folgen. Auf der einen Seite verringert dies die kognitive Eigenleistung der Lernenden, die nicht dem Bearbeiten der Aufgabe selbst dienen (s. Abschnitt 2.4 ECL). Auf der anderen Seite bringen diese Zusätze durch Pfeile & Highlightings auch eine Steigerung der kognitiven Belastung mit sich (Lewalter, 1997b).

5.2 Formulierung und Einordnung der Lernziele

Die konzipierte Intervention findet im Anfangsunterricht Chemie der Klassenstufe acht statt. Sie wirkt dort als Bindeglied zwischen dem Themenfeld 1 (TF1) „Chemikers Vorstellung von den Stoffen“ und dem Themenfeld 2 (TF2) „Von der Saline zum Kochsalz“ des Rheinland-Pfälzischen Lehrplans (Ministerium für Bildung, Wissenschaft, Weiterbildung und Kultur, 2014, S. 66–69). Im Vordergrund steht der Beitrag zur Entwicklung des Basiskonzepts Teilchen-Materie-Stoff mit Fokus auf den modellhaften Charakter von Beschreibungen auf der Teilchenebene (ebd.) über drei Doppelstunden hinweg. Die behandelten Themen sind Diffusion, Temperaturabhängigkeit der Diffusionsgeschwindigkeit, Bildung der Hydrathülle bei Salzen und bei anderen Stoffen, Stofftrennung durch Abdampfen sowie die Unterscheidung von Teilchen aufgrund unterschiedlicher Teilchengrößen mithilfe des Teilchensiebens. In einer ersten Version wurden zusätzlich noch ein Experiment und ein Modell zum Phänomen Osmose entwickelt und in Studie II bewertet. Im Zuge der Pilotierung wurden dieser Versuch und das zugehörige Modell aufgrund von Überlänge aus der Intervention gestrichen. Der Versuch zum Teilchensieben wurde in die Herstellung der Gojilösung und dem eigentlichen Teilchensieben aufgeteilt. Die Modelle zur Herstellung der Gojilösung entsprechen dabei den Modellen zur Herstellung der Salzlösung.

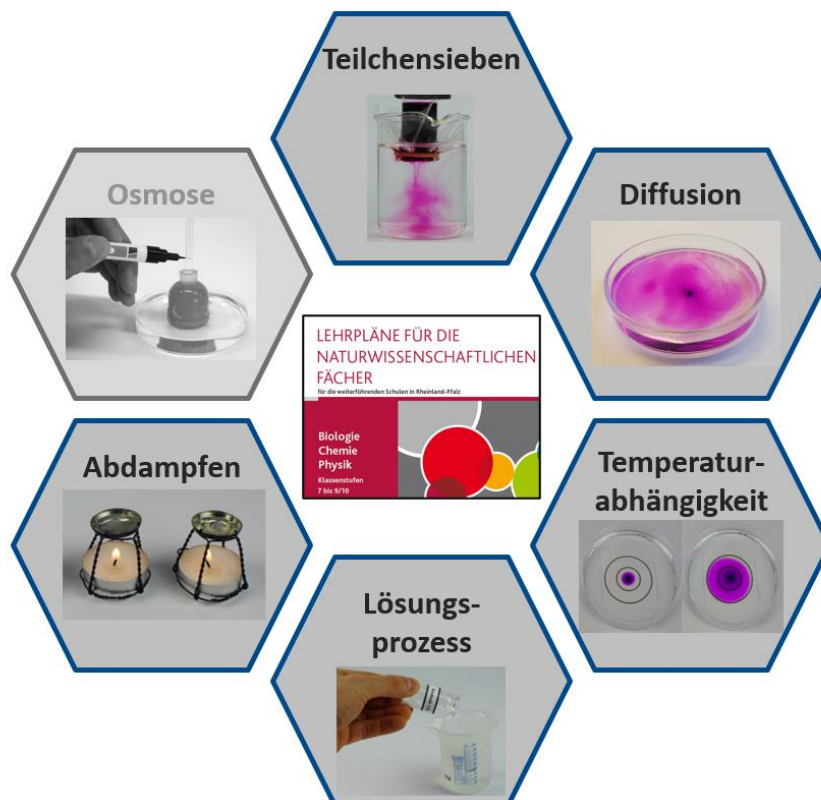


Abbildung 5: Themengebiete der Interventionen aus dem Rheinland-Pfälzischen Lehrplan für Chemie in Klassenstufe 8. (das Thema Osmose wurde aus zeitlichen Gründen von der Intervention ausgeschlossen und ist daher ausgegraut)

Folgende Kompetenzen aus dem Rheinland-Pfälzischen Lehrplan (Ministerium für Bildung, Wissenschaft, Weiterbildung und Kultur, 2014, S. 66–69) werden dabei in den drei Doppelstunden adressiert:

- „Die Schüler*innen führen einfache Untersuchungen zu Stoffeigenschaften durch“ (TF1).
- „Die Schüler*innen nutzen einfache Atomvorstellungen, um Phänomene auf der Teilchenebene zu erklären“ (TF1).
- „Die Schüler*innen führen einfache Versuche zu Stoffeigenschaften von Salzen und zur Stofftrennung durch und protokollieren diese“ (TF2).
- „Die Schüler*innen erklären die Eigenschaften von Salzen durch die Anwendung von Modellen“ (TF2).
- „Die Schüler*innen stellen den Ablauf der Salzgewinnung auf Teilchenebene dar“ (TF2).

Großlernziele zum Modellverständnis:

- Die Schüler*innen sollen die Rolle von Modellen in der Erklärung naturwissenschaftlicher Phänomene analysieren können.
- Die Schüler*innen sollen erkennen können, dass Modelle nie die ganze Wahrheit erklären können, und kritisch diskutieren, wie gut ein Modell ein Phänomen erklären kann.
- Die Schüler*innen sollen bewerten können, wann das einfachste Modell zur Erklärung eines Phänomens ausreicht und wann komplexere Modelle erforderlich sind.

Lernziele zu den Experimenten

Experiment 1: Diffusion

- Die Schüler*innen sollen den Prozess der Lösung von Salz im Wasser beschreiben können.
- Die Schüler*innen sollen Diffusion als den Mechanismus der gleichmäßigen Verteilung gelöster Stoffe im Wasser erklären können.
- Die Schüler*innen sollen die Brownsche Bewegung als Ursache für die eigenständige Bewegung gelöster Teilchen identifizieren können.

Experiment 2: Temperaturabhängigkeit der Diffusion

- Die Schüler*innen sollen den Zusammenhang zwischen Temperatur und Geschwindigkeit der Diffusion erklären können.

Experiment 3: Herstellung einer Salzlösung

- Die Schüler*innen können den Unterschied zwischen einem Reinstoff und einem Stoffgemisch erklären.
- Die Schüler*innen können die Hydrathülle als Resultat der Anziehungskräfte zwischen den Salzteilchen und den Wasserteilchen erklären.

- Die Schüler*innen können den Zusammenhang zwischen den Anziehungskräften zwischen Salz- und Wasserteilchen und der Bildung einer Lösung herstellen.
- Die Schüler*innen können den Begriff "Lösung" erklären und erläutern, warum er ein Beispiel für ein Stoffgemisch ist.

Experiment 4: Abdampfen einer Salzlösung

- Die Schüler*innen können den Vorgang des Abdampfens von Wasser beschreiben und den Unterschied in der Verteilung der Wasserteilchen und der Salzteilchen im Raum erklären.
- Die Schüler*innen können den Zusammenhang zwischen der Siedetemperatur von Kochsalz und Wasser und der Möglichkeit, Kochsalz durch Abdampfen aus Salzwasser zu gewinnen, erklären.

Experiment 5: Herstellung einer Goyilösung

- Die Schüler*innen können den Unterschied zwischen einem Reinstoff und einem Stoffgemisch erklären.
- Die Schüler*innen können die Hydrathülle als Resultat der Anziehungskräfte zwischen den Goyifarbstoffteilchen und den Wasserteilchen erklären.
- Die Schüler*innen können den Zusammenhang zwischen den Anziehungskräften zwischen Goyifarbstoff- und Wasserteilchen und der Bildung einer Lösung herstellen.
- Die Schüler*innen können den Begriff "Lösung" erklären und erläutern, warum er ein Beispiel für ein Stoffgemisch ist.

Experiment 6: Teilchensieben

- Die Schüler*innen können die Struktur einer Membran mit Poren beschreiben.
- Die Schüler*innen können beschreiben, wie man sich auf Basis des undifferenzierten Teilchenmodells die Anordnung der Teilchen einer Membran mit Poren vorstellt.
- Die Schüler*innen können auf Basis des undifferenzierten Teilchenmodells erklären, warum es Stoffe gibt, die aufgrund ihrer Teilchengröße durch die Membran diffundieren können und Stoffe, die dieses Verhalten nicht aufzeigen.

Lernziele zum Einsatz der Modelle

Modell 1: Diffusion

- Die Schüler*innen sollen beobachten können, dass Teilchen sich eigenständig und ungerichtet (diffus) bewegen.
- Die Schüler*innen sollen beobachten können, dass gelöste Stoffe sich gleichmäßig in der Lösung verteilen.

Modell 2: Diffusion bei untersch. Temperaturen

- Die Schüler*innen sollen beobachten können, dass gelöste Teilchen sich in Abhängigkeit von der Temperatur unterschiedlich schnell bewegen und in der Lösung verteilen.

Modell 3: Herstellung einer Salzlösung

- Die Schüler*innen sollen die Bildung der Hydrathülle als kennzeichnend für den Löseprozess erkennen können.

Modell 4: Abdampfen einer Salzlösung

- Die Schüler*innen sollen die Endzustände, dass die Wasserteilchen sich beim Verdampfen im Raum verteilen und die Salzteilchen in der Abdampfschale zurückbleiben, erkennen können.
- Der Beobachtungsauftrag beim Modell Abdampfen besteht darin festzustellen, welche Teilchen von den Senfkörnern beziehungsweise den Styroporkugel repräsentiert werden und wie sich die Teilchen beim Abdampfvorgang verhalten.

Experiment 5: Herstellung einer Gajilösung

- Die Schüler*innen sollen die Bildung der Hydrathülle als kennzeichnend für den Löseprozess erkennen können.

Modell 6: Teilchensieben

Die Schüler*innen sollen erkennen können, dass...

- ... Membranen nur für bestimmte Teilchen durchlässig sind.
- ... unterschiedliche Teilchen unterschiedliche Größen haben.

5.3 Konzeption des Aufgabenmaterials

Im Material zu den Interventionen wird durchgehend zwischen der „Ausgedachten Modellwelt der Teilchen“ und der „Realen Welt der Experimente & Phänomene“ unterschieden. Neben den Verweisen in den Texten wird die Unterscheidung durch drei optische Marker hervorgehoben (s. Abbildung 6 und Abbildung 7):

1. Überschriften in den Kopfzeilen der Arbeitsblätter:
 - a. Ausgedachte Modellwelt der Teilchen
 - b. Reale Welt der Experimente & Phänomene
2. Farbliche Kodierung:
 - a. Grüne Kästen um die Texte, Hintergrundfarbe der Seitenzahlen und Piktogramme für die Modellwelt
 - b. Blaue Kästen um die Texte, Hintergrundfarbe der Seitenzahlen und Piktogramme für die reale Welt
 - c. Orange Kästen um die Texte und Piktogramme für die Modellkritik
3. Piktogramme:
 - a. Grüne Wolke signalisiert die ausgedachte Modellwelt
 - b. Blauer Erlenmeyerkolben signalisiert die reale Welt der Experimente und Phänomene
 - c. Orange Glühbirne signalisiert die Modellkritik

Auswertung Versuch 2: Hat die Temperatur einen Einfluss auf das Verhalten von wasserlöslichen Stoffen, wenn man diese zu Wasser gibt?	
Beobachtung Stoffebene:	
• Je _____ die Temperatur, desto _____ verteilt sich der gelöste Stoff in der Lösung.	
Deutung Teilchenebene:	
• Die gelösten Teilchen bewegen sich bei _____ Temperaturen _____. • Dadurch verteilen sich die gelösten Teilchen schneller unter den _____.	
Modellkritik	
• Der Untergrund müsste aus Teilchen bestehen. • Die Teilchen müssten dichter gepackt sein. • Zwischen den einzelnen Teilchen dürfte absolut _____ sein = _____	

Abbildung 6: Ergebnissicherung nach dem Experiment zur Temperaturabhängigkeit der Diffusion. Es liegen optische Marker zur Symbolisierung der drei Ebenen - Blauer Kasten und Erlenmeyerkolben: „Reale Welt der Experimente und Phänomene“; Grüner Kasten und Wolke: „Ausgedachte Modellwelt“; Oranger Kasten und Glühbirne „Modellkritik“

5.4 Ablauf der Erarbeitungsphase

Schritt 1: Einstieg und Positionierung

Damit der Einstieg in jede Doppelstunde an die Vorerfahrungen und -kenntnisse der Schüler*innen anknüpft, wird ein alltagsnaher Kontext gewählt. So wird beispielsweise zum Einstieg in das Thema Brownsche Bewegung, die Bewegung des Wassers nach dem Umrühren als Kontext gewählt (s. Abbildung 7). Im Sinne der Conceptual Change Theorie sollen sich die Schüler*innen zu einer gegebenen Aussage positionieren und sich somit über ihre eigenen Präkonzepte bewusstwerden (s. Abbildung 7).

Schritt 2: Experiment 1

Das erste Experiment bietet den Lernenden die Möglichkeit, ihre eigenen Präkonzepte überprüfen. Weichen ihre Präkonzepte von der wissenschaftlichen Erklärung ab, wird bei den Lernenden eine kognitive Dissonanz erzeugt. Dies führt folglich zur Unzufriedenheit mit den eigenen, bereits existierenden Vorstellungen und damit zur Voraussetzung für die Einleitung eines Conceptual Changes.

Schritt 3: Deutung des Experiments durch Nutzung eines Modells

Um die entstandene kognitive Dissonanz aufzulösen, deuten die Schüler*innen das Experiment mithilfe eines Modells. Durch die vom Modell gelieferten Erklärungen wird das zuvor gezeigte Phänomen für die Lernenden verständlich. Dadurch wird eine weitere Bedingung für den Conceptual Change ermöglicht.

Schritt 4: Erneute Positionierung

Durch eine erneute Positionierung in Form der Anfertigung einer Zeichnung auf Teilchenebene wird die eigene Vorstellung überdacht und die Erweiterung des eigenen mentalen Modells angeregt (z.B. Sanger, 2000). Dadurch werden die neuen Vorstellungen für die Lernenden plausibel und der Conceptual Change weiter vorangetrieben.

Schritt 5: Sicherung der Ergebnisse

Die Ergebnissicherung erfolgt in Form eines Lückentexts mit anschließender Besprechung der Ergebnisse im Plenum. Dies ermöglicht die weitere Festigung der neuen Vorstellungen, da die vorherigen Schritte miteinander in Beziehung gesetzt werden und offene Fragen, sowie Verständnisprobleme im Plenumsgespräch erörtert werden können.

Schritte 6-10: Transfer auf ein neues Phänomen

In der zweiten Hälfte der Doppelstunde wird das bisher Gelernte auf ein neues Phänomen übertragen. In der ersten Doppelstunde ist dies das Phänomen der Temperaturabhängigkeit der Diffusion. Es werden die Schritte 1-5 in einem neuen alltagsbezogenen Kontext – z.B. das Lösen von Zucker im Vergleich von heißem zu kaltem Tee – wiederholt durchlaufen. Der Transfer auf einen neuen Sachverhalt ermöglicht es den Schüler*innen die neu erworbenen Vorstellungen als fruchtbar wahrzunehmen. Dadurch kann der Conceptual Change abgeschlossen werden und die neuen Vorstellungen langfristig in die mentalen Modelle der Lernenden integriert werden.

2

Ausgedachte Modellwelt der Teilchen



Was passiert mit den Teilchen eines wasserlöslichen Stoffs, wenn man diesen zu Wasser gibt?

Aufgabe 1:

Um sich das Verhalten von Stoffen erklären zu können, stellt man sich die Stoffe aus vielen kleinsten Teilchen zusammengesetzt vor. Rührt man ein Glas Wasser um, bewegt sich das Wasser darin. Die Bewegungen im Wasser werden als ein Verschieben der Wasserteilchen entlang der Rührbewegung gedeutet. Diese Strömung kann man auch mit Sensoren messen. Wartet man jedoch eine Weile, kann man auch mit den Sensoren keine Bewegung mehr feststellen. Also bewegen die Teilchen sich auch nicht mehr.

Was hältst du von dieser Aussage? Stelle **eine begründete Vermutung** auf und notiere diese im vorgesehenen Feld! Verwende dabei den Begriff Teilchen.

Abbildung 7: Beispiel zum Einstieg in die erste Doppelstunde zum Thema Brownsche Bewegung. Zu sehen sind der grüne Kasten, die grün hinterlegte Seitenzahl und die grüne Wolke als optische Marker für die Modellebene.

Die konzipierten Lernmaterialien zu allen drei Vergleichsgruppen befinden sich in Anhang 1 bis 9 zum Download.

5.5 Konzeption der Experimente

Im Mittelpunkt der Konzeption der Experimente stehen die zuvor formulierten Lernziele, die durch die Durchführung der Experimente erreicht werden sollen. Dabei wurden die Experimente so konzipiert, dass sie sowohl passend zum Forschungsdesign dieser Arbeit sind als auch bildungspraktische Implikationen und Materialien für weiterführende Studien liefern. Alle Experimente sind nach dem Low-Cost-Prinzip konzipiert, um die Replizierbarkeit in Einzelarbeit für größere Gruppen zu ermöglichen (vgl. Tabelle 6). Die grundlegenden Anforderungen an Schulexperimente nach Bader und Lühken (2018) wurden dabei zur Entwicklung der Experimente beachtet und umgesetzt:

Fachliche Korrektheit: Die Experimente stellen den zu vermittelnden Inhalt korrekt dar und zeigen die beabsichtigten Lernziele klar auf. Dies ist entscheidend, um die wissenschaftliche Genauigkeit zu gewährleisten und Missverständnisse zu vermeiden.

Deutliche Effekte: Aufgrund des Zeitpunktes im chemischen Anfangsunterricht wurden die Experimente so konzipiert, dass die Effekte des Experiments schnell und klar erkennbar sind. Dies erleichtert das Verständnis und erhöht die Motivation und das Interesse an naturwissenschaftlichen Themen.

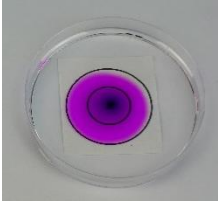
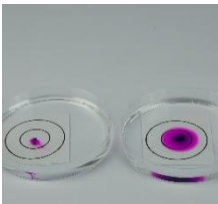
Erfolgreiche Durchführung: Alle Arbeitsschritte wurden so konzipiert, dass für jeden Schüler und jede Schülerin erfolgreich durchführbar sind. Dazu wurden die Experimente zunächst mit Studierenden präpilotiert und anschließend mit Schüler*innen pilotiert (s. Abschnitt 8.1). Zusätzlich wurden während der Intervention Hilfskräfte zur Hilfestellung bereitgestellt. Weiterhin wurden die Experimente auf Einhaltung des **zeitlichen Rahmens** überprüft. Dies umfasst sowohl die Durchführung selbst als auch die Vor- und Nachbereitungszeit. Zudem wurde die Durchführung der Experimente auf die **räumlichen Gegebenheiten und Ausstattungen** angepasst.

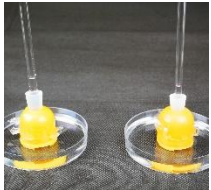
Angemessenheit an die Altersstufe: Es wurde darauf geachtet, dass die Experimente dem kognitiven Niveau der Schüler*innen entsprechen und im Zuge der Pilotierung überprüft (s. Abschnitt 8.1). Diese Überprüfung umfasst die Komplexität des Inhalts, die Art der verwendeten Materialien sowie den Umfang der Durchführung.

Berücksichtigung des Vorwissens: Um einen effektiven Erkenntnisgewinn zu ermöglichen, wurden die Experimente entsprechend der vorhandenen Kenntnisse der Schüler*innen entwickelt. Dazu wurden die bereits bearbeiteten Themenpunkte aus den rheinland-pfälzischen Lehrplänen sondiert. Dadurch wird sichergestellt, dass die Schüler*innen in der Lage sind, die neuen Informationen sinnvoll zu verarbeiten und ihre vorhandenen mentalen Modelle zu integrieren.

Arbeitssicherheit und Gefährdungsbeurteilung: Die verwendeten Chemikalien, deren eingesetzte Mengen und Konzentrationen sowie der Versuchsaufbau und die Sicherheitsvorkehrungen (z.B. Schutzbrillen) entsprechen den Sicherheitsstandards der Deutschen Gesetzlichen Unfallversicherung und wurden mittels der Online-Plattform DEGINTU (<https://degintu.dguv.de/>) überprüft. Beim Experiment zur Diffusion wurde aufgrund der erfolgten Substitutionsprüfung Kaliumpermanganat durch mit Lebensmittelfarbe eingefärbte Zuckerkristalle ersetzt.

Tabelle 6: Übersicht der für die Intervention entwickelten Versuche

Kurzbeschreibung des Versuchs	Didaktische Rekonstruktion des Versuchs	Versuchsaufbau
Diffusion Beobachtung der Diffusion von Kaliumpermanganat bei Raumtemperatur	Im Versuch zur Diffusion wird ein farbiger Kristall in eine mit destilliertem Wasser gefüllte Petrischale gegeben. Die Petrischale ist auf einer Unterlage mit konzentrischen Kreisen positioniert, um die Gleichmäßigkeit der radialen Verteilung zu verdeutlichen. Aufgrund einer Änderung des Tätigkeitsverbots für Schülerinnen und Schüler im Umgang mit farbigen Schwermetallsalzen wie Kaliumpermanganat oder Ammoniumchromat, wurde für diesen Versuch Hagelzucker mit der Lebensmittelfarbe Betanin purpurn eingefärbt.	
Temperaturabhängigkeit der Diffusion Beobachtung der Diffusion von Kaliumpermanganat bei 1°C und bei 60°C	Der Versuchsaufbau orientiert sich an dem des vorangegangenen Experiments, wird jedoch um eine zweite Petrischale erweitert. In beiden Petrischalen befindet sich destilliertes Wasser mit unterschiedlichen Temperaturen. Zur Demonstration des Temperatureinflusses auf die Diffusionsgeschwindigkeit wird simultan in beide Schalen ein eingefärbter Hagelzuckerkristall eingebracht. Um Verletzungen durch Verbrühen oder dem Beschädigen der Petrischalen durch zu große Hitze zu vermeiden, wurde sich für den Ansatz mit erhöhter Temperatur für 60°C entschieden. Um den Effekt noch deutlich und erfolgssicher für jeden Schüler und jede Schülerin zu gewährleisten, wurde für die niedrigere Temperatur ein Wert von 1°C gewählt.	
Herstellen einer Salzlösung Lösen von Salz in destilliertem Wasser	In diesem Versuch wird eine Kochsalzlösung der Massenkonzentration $w = 3\%$ aus Kochsalz und destilliertem Wasser hergestellt. Um die Zeit für das Herstellen und das Abdampfen (im nächsten Versuch) zu verkürzen, wird mit einer Wassertemperatur von 60°C gearbeitet. Zusätzlich ist die benötigte Menge an Kochsalz bereits abgewogen und in ein Rollrandglas abgefüllt. Im Sinne der Arbeitssicherheit werden für diesen Versuch PET-Bechergläser verwendet. Diese Wahl des Materials minimiert das Risiko von Verbrühungen durch das erwärmte Wasser und trägt somit zur Sicherheit der Durchführenden bei.	

Kurzbeschreibung des Versuchs	Didaktische Rekonstruktion des Versuchs	Versuchsaufbau
Abdampfen einer Salzlösung Stofftrennung von Salz und Wasser aufgrund unterschiedlicher Siedepunkte	Zur Stofftrennung von Kochsalz und Wasser aufgrund unterschiedlicher Siedepunkte wird eine Kochsalzlösung ($w = 3\%$) in der Kappe einer Agraße mit der Agraße als Stativ über einer Teelichtflamme abgedampft. Als Kontrollprobe wird mit destilliertem Wasser auf die gleiche Weise verfahren. Zur Verdeutlichung der Erhaltung der Masse, wird der Deckel einer Petrischale als Kondensationsoberfläche über die Abdampfschalen gehalten.	
Herstellen einer Lösung von Goji-Farbstoff Extrahieren von Cyanidin mit destilliertem Wasser aus schwarzen Gojibeeren	In diesem Versuch werden schwarze Gojibeeren in einer Reibschale zerrieben. Als Lösungsmittel dient destilliertes Wasser mit einem pH-Wert von 6,3. Diese spezifische pH-Einstellung ist essenziell, um einen Farbton zu erzielen, der dem einer 0,01 molaren Kaliumpermanganat-Lösung entspricht. Die präzise Abstimmung des Farbtons ist von entscheidender Bedeutung für den nachfolgenden Versuch, in dem die phänotypisch ähnlichen Lösungen aufgrund der unterschiedlichen Eigenschaften ihrer gelösten Teilchen differenziert werden sollen.	
Teilchensieben Vergleich der Diffusion von Kaliumpermanganat und Cyanidin durch eine semipermeable Membran	In diesem Experiment werden zwei optisch äquivalente Lösungen - eine 0,01 molare Kaliumpermanganat-Lösung und eine Cyanidin-Lösung aus dem vorherigen Versuch - mittels einer semipermeablen Membran hinsichtlich ihrer Diffusionseigenschaften untersucht und verglichen. Das primäre didaktische Ziel dieses Versuchs ist es, den Lernenden zu demonstrieren, dass Substanzen trotz phänotypischer Ähnlichkeit unterschiedliche Teilcheneigenschaften aufweisen können. Um möglichst geringe Stoffmengen einzusetzen und die Materialkosten gering zu halten, werden Ü-Ei-Kapseln als Kammern verwendet. Dies hat zudem den Vorteil, dass der Aufbau für Lernende einfach zu handhaben ist und der Versuch somit eine hohe Erfolgswahrscheinlichkeit aufweist.	
Osmose (nicht Teil der Interventionsstudie) Vergleich des osmotischen Drucks einer Salzlösung zu destilliertem Wasser	Es werden eine gesättigte Kochsalzlösung und destilliertes Wasser in jeweils eine Ü-Ei-Kapsel mit Steigrohr gefüllt, die mit einer semipermeablen Membran verschlossen sind. Anschließend werden die Kapseln mit der Membran in destilliertes Wasser gestellt und die osmotischen Prozesse beobachtet.	

Die Anleitungen zu jedem Versuch sind in Materialien und Durchführung unterteilt. Neben der Materialliste ist eine Abbildung des Versuchsaufbaus dargestellt (s. Abbildung 8). Die Durchführung ist wiederum in vier Unterpunkte gegliedert:

1. Vorbereiten des Versuchsaufbaus
2. Durchführung des Versuchs/ Darstellung des Phänomens
3. Formulierung der Beobachtung
4. Entsorgung

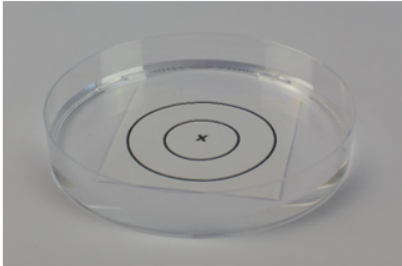
Materialien:	
• 1 Röhrchen mit violetten Salzkristallen • Leitungswasser mit Zimmertemperatur • Petrischale • Unterlage mit Kreisen • 1x Becherglas (250 mL) • 1 Pinzette	
Durchführung:	
Schritt 1: Präparieren der Petrischale 1. Stelle die Petrischale mit der Öffnung nach oben auf die Unterlage mit den Kreisen. <i>Achtung: Das Kreuz muss in der Mitte sein</i> 2. Fülle so viel Wasser in die Petrischale, dass der Boden gerade so bedeckt ist.	
Schritt 2: Diffusion 1. Nimm nun eines der kleinen Plastikröhrchen und gib einen violetten Salzkristall auf eine trockene Stelle auf dem Tisch. 2. Nimm nun den Kristall mit der Pinzette und lege ihn an die Stelle, an der sich das Kreuz befindet. <i>Wenn du das Kreuz nicht genau getroffen hast, kannst du Petrischale festhalten und die Unterlage vorsichtig an die richtige Stelle schieben.</i> Jetzt darfst du Schale und das Wasser nicht mehr berühren oder bewegen! 3. Beobachte was nun passiert. 4. <i>Nach einer Minute ist der Versuch beendet.</i>	
Beobachtung:	
Entsorgung: Lösungen in den Abfallbehälter auf dem Lehrerpult. Die restlichen Kristalle zurück in das Plastikröhrchen.	

Abbildung 8: Versuchsanleitung zum Versuch Diffusion

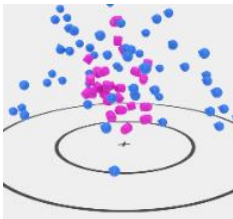
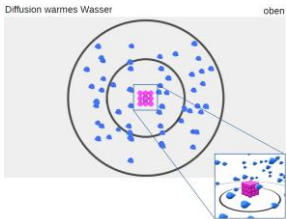
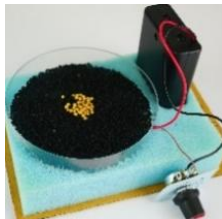
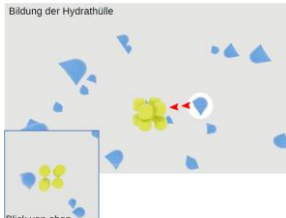
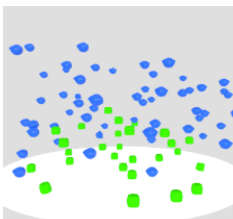
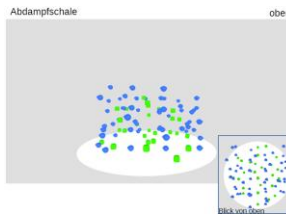
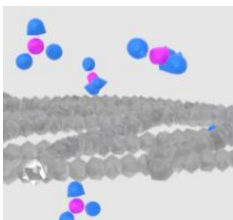
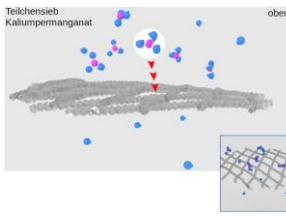
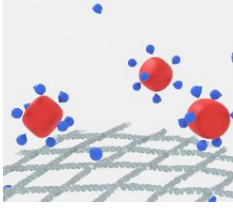
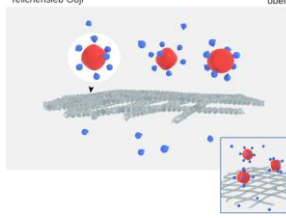
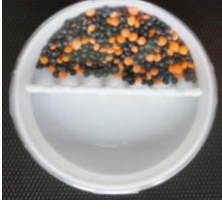
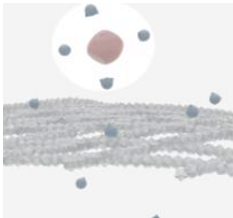
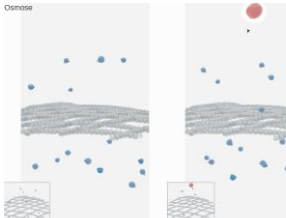
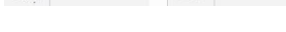
5.6 Konzeption der Modelle

Im Folgenden wird die Konzeption der Modelle am Beispiel der drei pilotierten und nach den Expert*innenkommentaren überarbeiteten Modelle zum Phänomen der Bildung der Hydrathülle erläutert. Die restlichen, optimierten Modelle sind in den Materialien zu den jeweiligen Interventionstagen in den Anhängen eins bis neun zu finden. Die zuvor formulierten Lernziele für die Arbeit mit den Modellen nehmen bei der Konzeption dieser ebenfalls eine zentrale Rolle ein. Die Lernziele fungieren dabei sowohl als Leitlinie als auch als Mindest- und Höchst-Maßstab für die Qualität der einzelnen Modelle der drei Modelltypen. Folglich werden die Modelle so konzipiert, dass keines mehr oder weniger Informationen bereitstellt, als zum Erreichen der Lernziele benötigt wird. Dadurch soll gewährleistet werden, dass die Modelltypen in gleichem Maße die Schüler*innen dabei unterstützen die geforderten Lernziele zu erreichen. In Tabelle 7 ist ein Überblick über alle, in der vorliegenden Arbeit entwickelten Modelle aufgeführt.

Weiterhin wurde bei der Erstellung der Modellanleitungen darauf geachtet, dass diese für alle drei Modelltypen gleich lang ausfallen. Dies ermöglicht die Kontrolle des Anteils der Beschreibungen am ECL. Um Gruppenunterschiede beim ECL auf den Einfluss des Modelltyps einzugrenzen, gilt es hierbei noch die Vergleichbarkeit der Modellanleitungen mittels Textoberflächenanalyse zu überprüfen (s. Abschnitt 8.1).

Als abschließendes Kriterium zur Konzeption der Modelle wurde die gleichwertige Interaktivität der Modelltypen berücksichtigt, um einen fairen Vergleich zu ermöglichen (Tversky et al., 2002). Die AR-Modelle bieten Interaktivität durch die Möglichkeit die Animation zu starten, anzuhalten und zu stoppen sowie den Betrachtungswinkel durch Bewegen des Tablets oder des optischen Markers zu variieren. Die HI-Modelle bieten die Möglichkeiten zur Interaktion in Form des Startens und Stoppens der Simulation, sowie zur Berührung und dadurch zum Verändern des Blickwinkels. Zudem interagieren die Lernenden mit dem Modell während des Aufbaus. Um eine Interaktivität bei den IL-Modellen zu erzeugen, die über das Anfassen der Papierbögen hinausgeht, wurde sich für die Variante einer ungeordneten Bildfolge entschieden. Die Lernenden interagieren mit dem Modell, indem sie die Bilder der Serie in die korrekte Reihenfolge bringen. Die Handhabung der einzelnen Modelle der unterschiedlichen Modelltypen kann als Video unter dem Link im digitalen Anhang 14 betrachtet werden.

Tabelle 7: Übersicht zu den Modellen der drei Modelltypen zu den Phänomenen der Intervention

Phänomen	AR	Illustrativ	Haptisch-interaktiv	
Diffusion				In eine Petrischale mit Vibrationsmotor werden Schwarzkümmel und Senfkörner gegeben.
Bildung der Hydrathülle				Klettkugeln werden mit Plüschkugeln in einem Pappbecher gemischt.
Abdampfen				Senfkörner und Styroporkugeln werden in einem doppelten Becher (Kleiner Becher innen zum Start; großer Becher außen) mit Netzboden und -deckel über einem Ventilator getrennt.
Teilchensieben				Eine Mischung aus schwarzen Linsen und Kichererbsen wird in einem 3D-gedruckten Modell mit eingedruckter, semi-permeabler Membran in Teilchendarstellung „gesiebt“.
KMnO₄				
Gojifarbstoff				Eine Mischung aus roten und schwarzen Linsen wird ebenso gesiebt.
Osmose (aufgrund eines zu großen Zeitaufwands von der Intervention ausgeschlossen)				Eine Mischung aus schwarzen Linsen und Kichererbsen wird in einem 3D-gedruckten Modell mit eingedruckter semi-permeabler Membran in Teilchendarstellung gerüttelt.

Anmerkungen: AR = Augmented Reality; Das AR-Modell wird durch Screenshots dargestellt. Die Abbildungen der illustrativen Modelle zeigen jeweils eines von acht Serienbildern mit dem zusätzlichen Blickwinkel

5.6.1 Augmented Reality Modelle

Die in dieser Arbeit entwickelten AR-Modelle wurden als markerbasierte Android Apps konzipiert. Die Darstellung erfolgt in Form einer dreidimensionalen Animation der Prozesse auf Teilchenebene (s. Abbildung 9). Dabei werden die Salzteilchen als gelbe Zylinder und die Wasserteilchen als blaue Kugeln mit je zwei kegelförmigen Ausstülpungen dargestellt. Zu besserer Übersichtlichkeit wird zusätzlich eine weiße Fläche als Hintergrund abgebildet. Zur Projektion wird ein optischer Marker und ein zehn Zoll Handheld Tablet eingesetzt (hier Augmented Solutions-Karte, s. Abbildung 9). Wird der Marker von der Kamera erfasst, wird das Modell in Teilchendarstellung auf dem Bildschirm des Tablets sichtbar. Die App verfügt über eine farbkoodierte, grafische Benutzeroberfläche. Durch Betätigen des grünen Play-Buttons wird die Animation gestartet und durch Betätigen des weißen Pause-Buttons pausiert. Betätigt man den Stopp/Reset-Button wird das Modell auf Anfang zurückgesetzt. Zur Erstellung der 3D-Modelle wurde die Software Blender (v 2.80) verwendet. Die Umsetzung der Plattform erfolgte mittels Unity (v 2019.2.3f1).

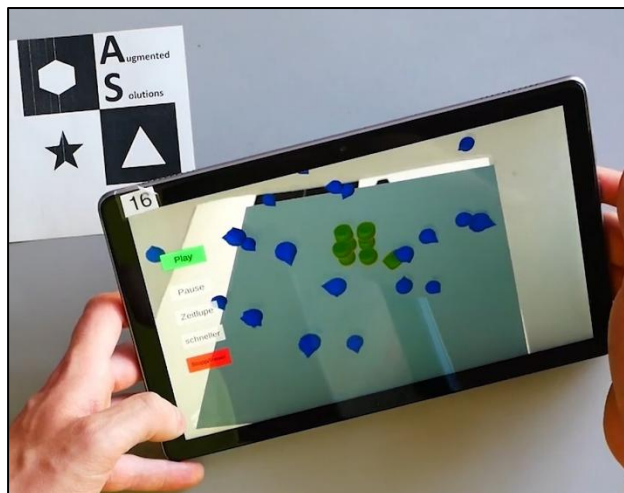


Abbildung 9: Nutzung des AR-Modells zur Bildung der Hydrathülle - oben links: Augmented Solutions-Karte als optischer Marker; Mitte: Tablet mit dem AR-Modell zur Teilchendarstellung der Bildung der Hydrathülle

Die Arbeitsaufträge zu jedem AR-Modell teilen sich in Materialien und eine Durchführung auf. Unter den Materialien sind die benötigten Marker und das Tablet mit einer Abbildung des Icons der benötigten App dargestellt. Die Durchführung ist wiederum dreigeteilt:

1. Vorbereiten des Modells
2. Nachbilden des Phänomens
3. Beobachtungsaufträge

Die Beobachtungsaufträge sind in Form von Fragen zu einfach beobachtbaren Ereignissen und richten sich inhaltlich nach den zuvor formulierten Lernzielen (s. Abbildung 10).

Materialien:	
• Aufbau aus dem vorherigen Versuch • Augmented Solutions Karte (im braunen Umschlag) • Tablet mit der App „Salzlösung 1“	
Durchführung:	
Schritt 1: Vorbereiten des Modells 1. Lege die Karte mit dem Augmented Solutions Aufdruck vor dich auf den Tisch. 2. Nimm nun das Tablet in die Hand und starte die App „Salzlösung“. Schritt 2: Simulieren der Teilchenbewegung und des Lösens in Wasser • Wenn du nun das Modell benutzt, sollst du herausfinden wie sich die Teilchenarten beim Lösevorgang anordnen. • Gelbe Teilchen = Salzteilchen Blaue Teilchen = Wasserteilchen • Wenn du nun die Kamera des Tablets auf den Aufkleber „Augmented Solutions“ hältst, erscheint das Modell. • Durch Drücken des Play-Buttons startet die Animation. • Mit dem Pause-Button hält die aktuelle Animation an. Du kannst sie durch erneutes Drücken auf den Play Button wieder starten. • Mit den Buttons „schneller“ oder „langsamer“ lässt sich die Geschwindigkeit der Animation steuern. • Durch den Stopp/Reset-Button wird das Modell auf Anfang zurückgesetzt. Du kannst auch hier durch den Play-Button wieder die Animation starten. Beobachtung • Welche Teilchen befinden sich nun wo? Wie sind die Teilchen zueinander geordnet?	

Abbildung 10: Arbeitsmaterial und -auftrag zum AR-Modell zur Bildung der Hydrathülle

Die APK-Dateien der entwickelten Apps und die zugehörigen Marker wurden zum Download auf der Plattform [zenodo.org](https://doi.org/10.5281/zenodo.14203449) unter <https://doi.org/10.5281/zenodo.14203449> veröffentlicht.

5.6.2 Illustrative Modelle

Die in dieser Arbeit entwickelten illustrativen Modelle wurden als Legebilder konzipiert. Die Darstellung erfolgt in Form von dreidimensionalen Abbildungen der Prozesse auf Teilchenebene (s. Abbildung 11). Ein illustratives Modell besteht aus vier Legebildern, die von den nutzenden Personen in die korrekte Reihenfolge gebracht werden müssen. Jedes Legebild setzt sich aus jeweils zwei aufeinanderfolgenden Screenshots der AR-Modelle zusammen. Zur Fokussierung auf die ablaufenden Prozesse, wurden relevante Stellen in den Bildern mit weißen Highlightings hervorgehoben. Die Bewegung der hervorgehobenen Teilchen wird durch Pfeile symbolisiert. Zur Verdeutlichung des Ablaufs der Prozesse im dreidimensionalen Raum, wurde ein zusätzlicher Bildausschnitt aus einem anderen Blickwinkel eingefügt.

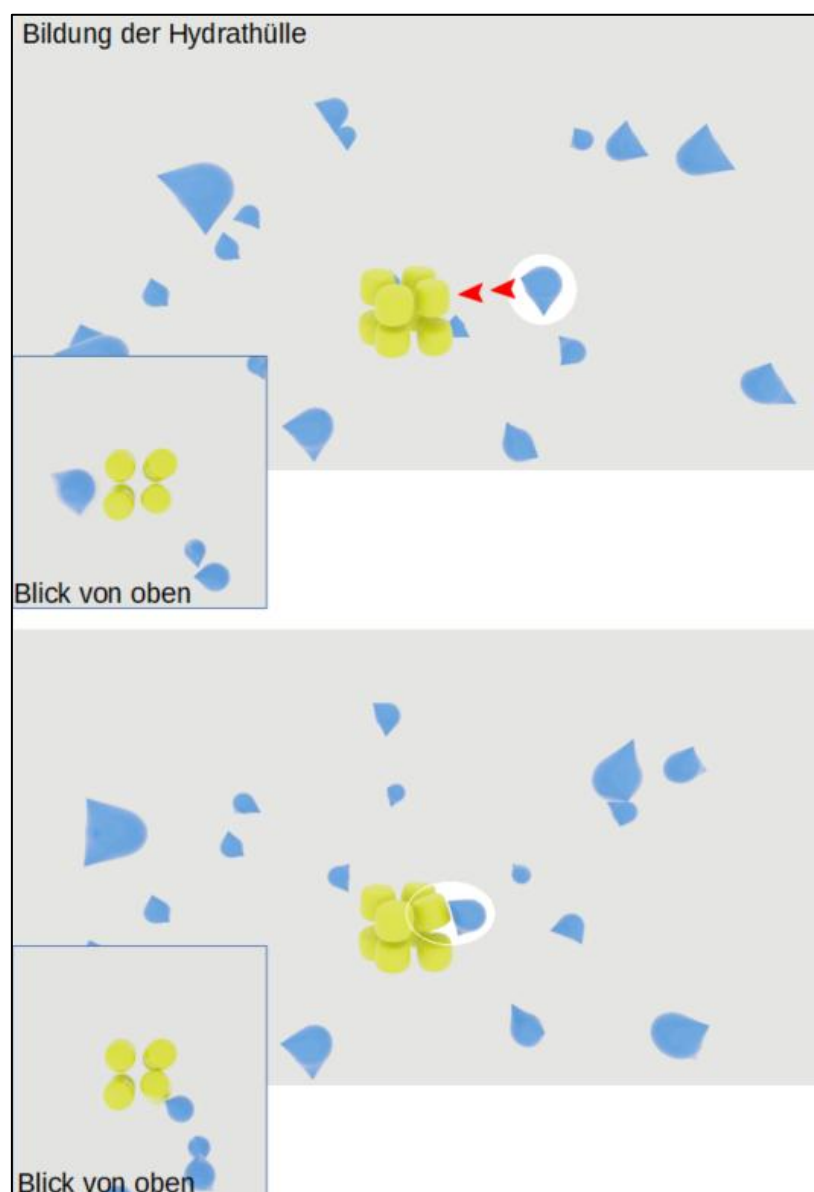


Abbildung 11: Eines von vier Legebildern zur Darstellung der Bildung der Hydrathülle. Zu erkennen sind die blauen Wasserteilchen, die gelben Salzteilchen sowie die roten Pfeile zur Symbolisierung der Teilchenbewegung und weißen Highlightings der relevanten Stellen.

Die Arbeitsaufträge zu jedem IL-Modell teilen sich in Materialien und eine Durchführung auf. Unter den Materialien sind die Legebilder aufgezählt und mit einer Abbildung aus der jeweiligen Bildfolge repräsentiert. Die Durchführung ist in zwei Schritte gegliedert:

1. Vorbereiten des Modells
2. Nachbilden des Phänomens und Beobachtungsaufträge

Die Beobachtungsaufträge sind in Form von Fragen zu einfach beobachtbaren Ereignissen und richten sich inhaltlich nach den zuvor formulierten Lernzielen (s. Abbildung 12).

Materialien:	
• 4 Legebilder zum Teilchenmodell zum Löseprozess	
Durchführung:	
Schritt 1: Vorbereiten des Modells 1. Nimm die 4 Legebilder zum Löseprozess und bringe sie in die richtige Reihenfolge. Tipp: Du kannst kontrollieren, ob die Reihenfolge stimmt, indem du die Nummerierung auf der Rückseite überprüfst. Schritt 2: Betrachten der Teilchenbewegung und des Lösens in Wasser 1. Wenn du nun das Modell benutzt, sollst du herausfinden wie sich die Teilchenarten beim Lösevorgang anordnen. 2. Schaue dir nun die Bildreihe genau an. Und beantworte Fragen a und b. a. Wo befinden sich welche Teilchen auf dem letzten Bild? b. Wie sind die Teilchen zueinander geordnet?	

Abbildung 12: Arbeitsmaterial und -auftrag zum illustrativen Modell zur Bildung der Hydrathülle

5.6.3 Haptisch-interaktive Modelle

Die in dieser Arbeit entwickelten haptisch-interaktiven Modelle wurden wie die Experimente ebenfalls nach dem Low-Cost-Prinzip konzipiert. Die Darstellung erfolgt in Form von dreidimensionalen, haptischen Modellen. Im haptisch-interaktiven Modell zur Bildung der Hydrathülle werden die Wasserteilchen durch farbige Klettkugeln und die Salzteilchen durch farbige Plüschkugeln dargestellt (s. Abbildung 13). Die Darstellung der Prozesse auf Teilchenebene erfolgt dabei durch die Einwirkung der Nutzer. Im Fall des Modells zur Hydrathülle werden die Wasser- und die Salzteilchen zusammen in einen Pappbecher gegeben und durch Rotieren des Bechers die Teilchenbewegung simuliert.

Die Arbeitsaufträge zu jedem haptisch-interaktiven Modell teilen sich in Materialien und eine Durchführung auf. Unter den Materialien sind die Komponenten der Modelle aufgezählt und mit einer Abbildung des jeweiligen Modells dargestellt. Die Durchführung ist in zwei Schritte untergliedert:

1. Vorbereiten des Modells
2. Nachbilden des Phänomens und Beobachtungsaufträge

Die Beobachtungsaufträge sind in Form von Fragen zu einfach beobachtbaren Ereignissen und richten sich inhaltlich nach den zuvor formulierten Lernzielen (s. Abbildung 13).

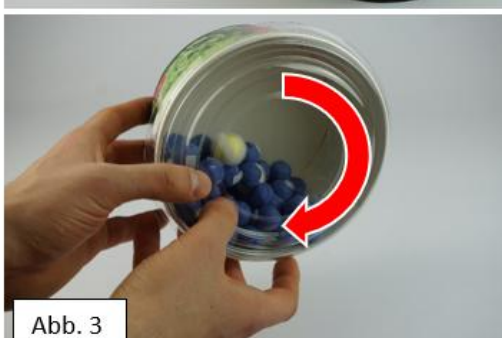
Materialien:	
• 1 großer Pappbecher • 4 Papierkugeln mit Plüschband (Salzteilchen) • 42 Papierkugeln mit Klettband (Wasserteilchen)	
Durchführung:	
Schritt 1: Präparieren des Modells • Gib die Papierkugeln mit dem Plüschband in die Mitte des großen Bechers (Abb. 1). • Schütte nun alle Papierkugeln mit dem Klettband rundum in den großen Becher (Abb. 2). • Jetzt kannst du den Deckel auf den großen Becher setzen. Das Modell ist nun einsatzbereit. Schritt 2: Simulieren der Teilchenbewegung und des Lösens in Wasser • Halte nun den großen Pappbecher in beiden Händen und dreh ihn wie eine Wäschetrommel (Abb. 3). • Nach 15 Umdrehungen kannst du das Modell öffnen und dir anschauen, was sich verändert hat! • Welche Teilchen befinden sich nun wo? • Nun kannst du gerne einzelne Teilchen aus dem Modell herausnehmen und dir genauer anschauen. • Wie sind die Teilchen zueinander geordnet?	 Abb. 1  Abb. 2  Abb. 3

Abbildung 13: Arbeitsmaterial und -auftrag zum haptisch-interaktiven Modell zur Bildung der Hydrathülle

6. Studie I – Entwicklung des Messinstruments „BeQuAn“ zur Bewertung der Qualität von Anschauungsmodellen

Die einzelnen Modelltypen weisen jeweils unterschiedliche Charakteristika auf. Dadurch ist es nicht möglich die Inhalte mit den unterschiedlichen Medien identisch abzubilden. Um die Wirkung der verschiedenen Modelltypen miteinander vergleichen zu können, müssen die unterschiedlichen Darstellungsformen in ihrer Qualität miteinander übereinstimmen. Daher bedarf es einer Beurteilung der einzelnen Modelle, um eine gleichwertige Qualität der einzelnen Interventionen im Rahmen der Vergleichsstudie zu gewährleisten. Da die einzelnen Kriterien auf Beobachtungen mit geringem Komplexitätsgrad zurückzuführen sind, fällt die Wahl der Verfahrensweise auf die Methode des Urteils von Fachexpert*innen (Döring & Bortz, 2016). Im Folgenden werden die Gütekriterien für naturwissenschaftsdidaktische Anschauungsmodelle (Eschenhagen, Kattmann, & Rodi, 2008) und Modelle in der Informatik (Fischer et al., 1995; Rauh & Stickel, 1997; Mahr, 2008a) auf ihre Eignung zur quantitativen Gütebewertung diskutiert. Studie I verfolgt das Ziel, ein valides Messinstrument zur Bestimmung der Qualität von naturwissenschaftlichen Anschauungsmodellen „BeQuAn“ zu entwickeln. Dazu wird mittels Expert*inneninterviews und anschließender qualitativer Inhaltsanalyse folgende Hypothese überprüft:

*(H1) Die Antworten von Expert*innen zu den entwickelten Items entsprechen inhaltlich den Beschreibungen der Gütekriterien aus der Primärliteratur.*

Die Untersuchung dieser Hypothese durch Expert*inneninterviews und anschließende qualitative Inhaltsanalyse ermöglicht es, Einsichten in die Übereinstimmung zwischen den Expert*innenantworten und den Beschreibungen der Gütekriterien in der Primärliteratur zu gewinnen. Diese Erkenntnisse sind entscheidend für die Validierung und Weiterentwicklung der Modellbewertungsmethoden sowie für die Gewährleistung ihrer Anwendbarkeit und Verlässlichkeit in der wissenschaftlichen Praxis.

6.1 Formulierung der Items

Zur Bestimmung der Modellgüte wurden die zuvor unter 2.1 beschriebenen Gütekriterien herangezogen. Die Gütekriterien wurden auf fünf Items ausgelegt, die den gesamten Charakter der Modellqualität abbilden. Einige Kriterien von Eschenhagen, Kattmann & Rodi (2008) entsprechen in ihrer Charakterisierung den Kriterien von Rauh und Stickel (1997). Daher wurden diese Kriterien jeweils einem Item zusammengefasst. Da die Verständlichkeit die Basis für didaktische Anschauungsmodelle darstellt (Eschenhagen, Kattmann & Rodi, 2008), wurden die Unterpunkte des Kriteriums für Verständlichkeit in einzelne Items überführt. Somit soll gewährleistet werden, dass das Modell in jedem Fall seinem Zweck dienlich wird. In Bezug auf

die geplante Lernsituation besteht der Zweck der Modelle darin, den Lernenden Informationen zum Erreichen der Lernziele zu vermitteln. Das Erreichen der Lernziele wird in den Items daher explizit thematisiert, um die Abgrenzung der Items auf die Einsatzsituation zu schärfen. Weiterhin ergibt sich daraus die Notwendigkeit jedes Modell immer im Kontext des zugehörigen Lernziels zu bewerten. Der Einsatz des Messinstruments sieht es vor die Einschätzung von Expert*innen zu einzelnen Merkmalen des Modells zu erfragen. Um Akquieszenzeffekte zu vermeiden, wurden die Items in Form von Fragen formuliert (Jonkisz, Moosbrugger, & Brandt, 2012). Aufgrund eines besseren Verständnisses des Items zur Überprüfung der Redundanz, wurde ein positiv formulierter Itemwortlaut bei negativer Itempolung gewählt (Jonkisz, Moosbrugger, & Brandt, 2012).

Im Folgenden sind die Items mit den abgebildeten Gütekriterien dargestellt.

1) Ist das Phänomen auf visueller Ebene fachlich korrekt dargestellt?

Dieses Item bildet die Eigenschaft der **fachlichen Korrektheit** nach Rauh und Stickel (1997) ab. Da ein Modell nur dann fachlich korrekt sein kann, wenn es in den wesentlichen Merkmalen mit dem Original übereinstimmt, wird auch das Kriterium der Ähnlichkeit und Entsprechung (Eschenhagen, Kattmann, & Rodi, 2008) von diesem Item erfasst.

2) Ist das Phänomen für die Zielgruppe verständlich dargestellt?

Mit diesem Item wird das Kriterium der **Verständlichkeit** nach Rauh und Stickel (1997) abgebildet. Wenn das Modell von der Zielgruppe verstanden wird, erlaubt dies Prognosen über die originale Modellvorstellung zu abzuleiten. Da die für die Zielgruppe verständliche Darstellung Voraussetzung für die Prognosen ist, erfasst dieses Item auch das Kriterium der Exaktheit und Fruchtbarkeit (Eschenhagen, Kattmann, & Rodi, 2008).

3) Ist das Phänomen zum Erreichen der Lernziele ausreichend ausführlich dargestellt?

Dieses Item dient zur Überprüfung des Kriteriums der **Vollständigkeit** nach Rauh und Stickel (1997). Da eine ausreichend ausführliche Darstellung alle wesentlichen Merkmale des Modells abdeckt, bildet dieses Item auch das Prinzip der Adäquatheit (Eschenhagen, Kattmann, & Rodi, 2008) ab.

4) Ist das Modell übersichtlich gestaltet?

Die **Übersichtlichkeit** stellt eine Subkategorie des Kriteriums der Verständlichkeit dar. Ist ein Inhalt überpräsent, hat dies eine verringerte Übersichtlichkeit zur Folge und stellt darüber hinaus einen Verstoß gegen das Prinzip der Genügsamkeit (Eschenhagen, Kattmann, & Rodi, 2008) bzw. dem Gesetz der Einfachheit (Winnenburg, 1993) dar. Daher erfasst dieses Item die beiden Kategorien Übersichtlichkeit und Genügsamkeit.

5) *Enthält das Modell Informationen, die zum Erreichen der Lernziele überflüssig sind?*

Sind im Modell Informationen enthalten, die nicht dem Zweck des Modells dienen, sind diese als überflüssig anzusehen. Wird dieses Item abgelehnt, ist damit die **Redundanzfreiheit** des Modells gegeben. Ebenso wird das Modell auf das Prinzip der Einfachheit überprüft, da überflüssige Informationen einer Einfachheit entgegenwirken.

Um eine Grundlage für den Vergleich der Modelle zu schaffen, ist es erforderlich, dass die Qualitätskriterien in einer Weise konzipiert sind, die eine Quantifizierung ermöglicht. (Döring & Bortz, 2016). Das bevorzugte Format einer Beurteilungsskala für eine solche Bewertung stellen diskrete Ratingskala-Aufgaben dar (Jonkisz, Moosbrugger, & Brandt, 2012). Hierbei wird eine Likert-Skala verwendet, um die Zustimmung zur Erfüllung der einzelnen Kriterien zu erfassen. Um die Komplexität der Befragung auch für eine große Anzahl an Modellen möglichst klein zu halten, wurde eine vierstufige Skala gewählt (Bühner, 2006). Das Messinstrument soll eingesetzt werden, um die Stärken und Schwächen der einzelnen Modelltypen herauszustellen. Daher wurde sich für eine gerade gestufte Skala entschieden, um eindeutig positive beziehungsweise eindeutig negative Beurteilungen der Kriterien zu forcieren. Weiterhin wurde sich für das Format der bipolaren Skala entscheiden um eine klare Stellungnahme zu den einzelnen Kriterien zu erhalten (Bühner, 2006). Durch die gerade Stufung und Bipolarität der Skala kann zusätzlich eine Tendenz zur Mitte im Antwortverhalten der Befragten reduziert werden (ebd.). Darüber hinaus gibt dies den Rater*innen die Möglichkeit zu sowohl positiver als auch negativer Kritik. Die negative Polung des Items zur Redundanzfreiheit ermöglicht es zudem Ankreuzmuster zu identifizieren (Busker, 2014) und Akquieszenzeffekte zu verringern (Moosbrugger & Kelava, 2012). Die Beobachtungen, die der Fragebogen von den Expert*innen einfordert, beziehen sich auf einzelne zuvor festgelegte Aspekte. Daher zählen diese als Beobachtungen mit geringem Komplexitätsgrad (Döring & Bortz, 2016). Zusätzlich zur Likert Skala, gibt es zu jedem Kriterium ein Kommentarfeld für offene Antworten auf die Fragen.

6.2 Validierung des Messinstruments

Methoden der Datenerhebung

Studie I zielt darauf ab, die Tauglichkeit der von Eschenhagen et al. (2008) sowie von Rauh und Stickel (1997) entwickelten Kriterien zur Evaluierung der Modellqualität zu untersuchen und diese in eine valide Kurzskala zu überführen (vgl. Przywarra et al., 2021). Dazu wurde folgende Hypothese überprüft:

*(H1) Die Antworten von Expert*innen zu den entwickelten Items entsprechen inhaltlich den Beschreibungen der Gütekriterien aus der Primärliteratur.*

Für die Bestimmung der Inhaltsvalidität der Items wurde ein mehrstufiges, zyklisches Verfahren angewendet. Ein erprobtes Vorgehen zur Validierung und Optimierung von Items stellt die Methode des lauten Denkens (Sandmann, 2014) im Rahmen einer Expert*innenbefragung (Bogner & Menz, 2002) dar. In einem ersten Schritt wurden mittels lauten Denkens die Formulierungen der Items auf Verständlichkeit sowie die deduzierten Kategorien der Modellqualität auf Vollständigkeit überprüft. Gegenstand der Untersuchungen sind die Interpretationen und Kognitionen der Proband*innen während der Bearbeitung der Bewertungsbögen. Aus diesem Grund wurde sich für das laute Denken in Form einer Introspektion entschieden (Funke & Spering, 2006; Konrad, 2010). Dadurch können die dynamischen und prozeduralen Aspekte der kognitiven Prozesse handlungsnah erfasst werden (Funke & Spering, 2006). Basierend auf dem Modell menschlicher Informationsverarbeitung nach Ericsson und Simon (1993) lässt sich die Tätigkeit während der Erhebung der zweiten Verbalisierungsebene zuordnen. Proband*innen verbalisieren Inhalte, die noch nicht in verbal kodierter Form im Kurzzeitgedächtnis vorliegen. Ericsson und Simon (1993) bezeichnen Verbalisierungen dieser Art als „think aloud“. Aus den Verbalisierungen während der Bearbeitung eines Items kann darauf geschlossen werden, ob tatsächlich Wissen zum passenden Gütekriterium aktiviert wurde, oder ob andere Denkprozesse im Vordergrund stehen (Schmiemann & Lücken, 2014). Die Erhebung der Daten erfolgte dabei unter Einhaltung der sechs Punkte einer typischen Lernsituation zum lauten Denken nach Sandmann (2014). Die Proband*innen erhielten eine kurze Einführung in den Ablauf und das Thema der Think-Aloud-Sitzung. Zwecks einer besseren Vorbereitung auf die bevorstehende Aufgabe, erhielten die Versuchspersonen als nächstes Informationen über den geplanten Einsatz des Fragebogens und zum Ziel der Think-Aloud-Sitzung. Um die Testpersonen auf die Tätigkeit des lauten Denkens einzustimmen, wurden mit jedem/jeder Proband*in zwei Probedurchgänge absolviert. Im ersten Übungsdurchlauf hatten die Proband*innen den Auftrag einen beliebigen Papierflieger zu falten und dabei ihre Absichten während ihrer Handlungen zu beschreiben (Sandmann, 2014). Die zweite Übungssequenz forderte die Proband*innen dazu auf einen für sie perfekten Tag inklusive Begründungen für ihre Entscheidungen zu beschreiben. Im Anschluss daran wurde den Versuchspersonen das AR-Modell zur Diffusion mit den dazugehörigen Lernzielen und der Klassenstufe vorgelegt, um den kontextuellen Rahmen für die Bewertung zu schaffen. Dann wurden die Proband*innen aufgefordert den Bewertungsbogen zu lesen und das zuvor gesehene Anschauungsmodell zu bewerten. Dabei sollten sie jeden Gedankengang, der ihnen während der Bearbeitung in den Sinn kommt, laut verbalisieren. Insbesondere die Aspekte und Begründungen, an welche die Expert*innen während der Bewertung denken, sollten dabei verbalisiert werden. Bei längeren Redepausen wurden die Proband*innen vom Versuchsleiter dazu aufgefordert, ihre Gedanken laut auszusprechen (Sandmann, 2014). Um die kognitiven Vorgänge beim Verbalisieren

(Ericsson & Simon, 1993) oder deren zeitliche Abfolge (Sasaki, 2003) nicht zu ändern, wurden darüber hinaus keine weiteren Prompts gesetzt. Obwohl die Items des Bewertungsbogens die Proband*innen dazu auffordern, konkrete Faktoren zu erläutern oder zu interpretieren, wird dadurch der „think aloud“-Charakter der Erhebung beibehalten, da diese Handlungen die eigentliche Primäraufgabe bilden (Ericsson & Simon, 1993). Der erste Durchgang diente neben einem ersten Validierungszyklus zusätzlich der Überprüfung des Kategoriensystems (Przywarra et al., 2021). Die Stichprobe für den Testdurchgang besteht aus einer Biologie- und Chemielehrerin (weiblich, 2 Jahre Lehrererfahrung, Alter = 27) sowie einer Chemielehrkraft, die zusätzlich in Chemiedidaktik promoviert (männlich, 2 Jahre Lehrererfahrung, Alter = 28). Die technische Datensicherung erfolgte in Form von Audioaufzeichnungen.

Methoden der Datenauswertung

Zur Datenaufbereitung wurde das einfache Transkriptionssystem von Kuckartz et al. (2008) angewandt. Aufgrund einer Erleichterung beim späteren Kodieren wurde die Handhabung von Interpunktionen nach Dresing & Pehl (2015) verändert. Darüber hinaus wurde in Anlehnung an Sandmann (2014) im Transkript eine Unterscheidung zwischen Aufgabenmaterial und Äußerungen der Proband*innen vorgenommen. Die vollständigen Transkriptionsregeln sind in Anhang 15 zu finden. Im Zuge der Transkription wurden die Aussagen aufgrund einer besseren Lesbarkeit redigiert.

Zur Validierung der Items wurde anhand der aus dem lauten Denken entstandenen Transkripte (Anhang 19) eine qualitative Datenanalyse durchgeführt. Die einzelnen Schritte der Validierung erfolgten in Form einer inhaltlich-strukturierenden, qualitativen Inhaltsanalyse. Gläser und Laudel (2012) empfehlen zur Auswertung von Expert*inneninterviews das schrittweise Vorgehen nach Mayring (2010). Der Vorteil dieser Variante gegenüber der Vorgehensweise nach Kuckartz (2016) liegt in der Fokussierung auf die Überprüfung der Hypothesen durch Extraktion der Daten aus dem Material (Mayring, 2019; Mayring & Fenzl, 2019; Schreier, 2014). Aufgrund der relativ kurz ausgefallenen Interviews empfiehlt Mayring (2019) für den konkreten Prozess der Kodierung ein Vorgehen nach Kuckartz (2016). Das Vorgehen nach Kuckartz (2016) hat zunächst den Vorteil, dass das Gesamtmaterial und der Kontext auch im Zuge der Kodierung berücksichtigt werden. Dadurch ist möglich die Aussagen zu einzelnen Items im Gesamtkontext zu betrachten und erleichtert es sie zu den Äußerungen zu anderen Items abzugrenzen. Ebenfalls werden die Kodiereinheiten bei Kuckartz freier gewählt. Dies ermöglicht es, die Kodiereinheiten in ihrer Länge sowohl an die umfangreichen Beschreibungen der Kategorien/Items als auch in die Antworten der Proband*innen anzupassen. Auf diese Weise lässt sich das Problem einer Mehrfachkodierung einer oder mehrerer Kategorien im gleichen Abschnitt vermeiden (Schreier, 2014). Sowohl Kuckartz (2016) als auch Mayring (2010) sehen

es vor Kategorien im Nachhinein induktiv durch Subkategorien zu erweitern. Da Expert*innen zum Modelleinsatz interviewt werden, bietet dies die Möglichkeit auf Basis ihrer Beiträge induktiv weitere Kategorien zu generieren.

Die ermittelten Gütekriterien nach Rauh und Stickel (1997) sowie nach Eschenhagen, Kattmann, & Rodi (2008) bilden die Basis der thematischen Hauptkategorien. Diese Kategorien lauten fachliche Korrektheit, Verständlichkeit, Vollständigkeit, Übersichtlichkeit und Redundanzfreiheit und spiegeln damit jeweils ein Item des Fragebogens wider. Darüber hinaus dienen die Beschreibungen der Gütekriterien im weiteren Validierungsprozess als Basis für das Kategoriensystem (Anhang 16). Dieses Kategoriensystem wurde in einem ersten Testdurchlauf anhand von Material aus zwei Think-Aloud-Sitzungen auf seine konkrete Tauglichkeit für das empirisch gewonnene Material hin überprüft. Dies entspricht einem Drittel des über alle Phasen hinweg gesammelten Auswertungsmaterials. Damit ist der Umfang des Probematerials auch für umfangreiche und vielschichtige Datensätze ausreichend, um das Kategoriensystem und den Kodierleitfaden (Anhang 17) zu testen (Kuckartz, 2016). Zusätzlich wurde noch ein weiteres fiktives Referenzmaterial zur Überprüfung des Kodierleitfadens und der Validität des Kategoriensystems (Schreier, 2014) erstellt. Bei diesem *Material 0* handelt es sich um Textstellen der Primärliteratur zur Beschreibung der einzelnen Gütekriterien aus denen die Items abgeleitet wurden. *Material 0* wurde von den Testkodierenden als erstes bearbeitet. Dadurch soll die inhaltliche Validität der Kategorienbeschreibungen gewährleistet werden. Zusätzlich diene der Testzyklus am Probematerial zur Überprüfung der Verständlichkeit der Kategorienbeschreibungen. Das Probematerial wurde von drei Kodierenden kodiert. Als Kodierende wurden drei männliche Naturwissenschaftsdidaktiker mit jeweils mehr als 50 Stunden Kodiererfahrung eingesetzt.

Für die Bestimmung einer geeigneten Kodiereinheit ist zu beachten, dass diese den mentalen Repräsentationen bzw. den verbalisierten, kognitiven Prozessen der Versuchspersonen entspricht (Ericsson & Simon, 1993). Dabei ist darauf zu achten, dass das kodierte Segment dann noch verständlich ist, wenn es vom Kontext losgelöst steht (Kuckartz, 2016). Weiterhin soll die Kodiereinheit einen sich abgeschlossenen, nachvollziehbaren Gedankenschritt darstellen (Sandmann, 2014). Durch die Fragen des Bewertungsbogens hat die Befragung der Proband*innen den Charakter eines leitfadenorientierten Interviews (Hopf, 1995), genauer eines Expert*inneninterviews (Bogner & Menz, 2002; Flick, 2017). Kuckartz (2016) empfiehlt in diesem Fall die Größe einer Kodiereinheit auf die komplette Aussage einer Person zu einer Frage festzulegen. Die Kodierung des Datenmaterials erfolgte mittels des Programms MaxQDA 2020.

Auf Basis der Erkenntnisse des ersten Validierungszyklus wurden zwei Items angepasst und eine Kategorienbeschreibung erweitert. Das beschriebene methodische Vorgehen des ersten

Zyklus wurde in einem zweiten Validierungszyklus wiederholt. Die Stichprobe für diese Hauptuntersuchung besteht aus Naturwissenschaftsdidaktiker*innen und -lehrer*innen Fachdisziplinen ($n = 4$; ♀ = 50 %; Alter: $MW = 35.8$ Jahre, $SD = 6.3$; Lehrerfahrung: $MW = 8.3$ Jahre, $SD = 6.0$) und zeichnet sich dadurch aus, dass alle Expert*innen aktiv in Lehrtätigkeiten eingebunden sind und mithilfe von naturwissenschaftlichen Modellen lehren. Damit beläuft sich die Gesamtstichprobe auf sechs Expert*innen (♀ = 50 %; Alter: $MW = 33.3$ Jahre, $SD = 6.15$; Lehrerfahrung: $MW = 6.33$ Jahre, $SD = 5.5$). Als Abbruchkriterium zur Probandenakquise wurde das Prinzip der gesättigten Stichprobe angewandt (Glaser & Strauss, 1967; Marshall, 1996). Die Bestimmung der Inhaltsvalidität erfolgte über die relative Häufigkeit der Übereinstimmung zwischen den Aussagen der Expert*innen und den Beschreibungen der einzelnen Kategorien/Gütekriterien (Sandmann, 2014). Die Inter-Coder-Reliabilität wurde mithilfe des Gütemaßes Fleiss' κ und für eine genauere Berechnung bei kleinen Stichproben (Everitt & Skrondal, 2010) über die Intraklassen-Korrelation bestimmt (Shrout & Fleiss, 1979). Die Kodierenden wurden nicht zufällig ausgewählt und die Kodierung aller beteiligten Kodierenden wurden in die Auswertung mit aufgenommen. Daraus ergibt sich die ICC3.1 als geeignetes Reliabilitätsmaß (Shrout & Fleiss, 1979). Zur Bestimmung der Interraterreliabilitäten wurde die Software *R-Studio* (Version 2021.09.1) mit R (Version 3.14.0) unter Verwendung der Pakete *psych* und *irr* genutzt.

Ergebnisse des ersten Zyklus

Für die Kodierung des Testdurchlaufs ergab sich eine Interraterreliabilität von 100% (Fleiss' $\kappa = 1$; $p < 0.001$). Die Ergebnisse der Testkodierung sind in Tabelle 8 aufgelistet.

*Tabelle 8: Ergebnisse der Kodierung des Probematerials aus dem ersten Zyklus nach Proband*innen, Kodiereinheiten und Kodierenden sortiert*

Proband	Kodiereinheit	Coder 1	Coder 2	Coder 3
Material 0	1	1	1	1
Material 0	2	2	2	2
Material 0	3	3	3	3
Material 0	4	4	4	4
Material 0	5	5	5	5
Sa1	1	0	0	0
Sa1	2	2	2	2
Sa1	3	3	3	3
Sa1	4	4	4	4
Sa1	5	5	5	5
Mh1	1	1	1	1
Mh1	2	2	2	2
Mh1	3	2	2	2
Mh1	4	4	4	4
Mh1	5	5	5	5

Implementierung der Befunde des ersten Zyklus

Die Bearbeitung des Referenzmaterials (Material 0, Textabschnitte aus der Primärliteratur) ergab, dass alle Kategorien im Kategoriensystem ausreichend ausführlich beschrieben wurden und die Kodiereinheiten eindeutig den entsprechenden Kategorien zugeordnet werden konnten. Folglich können nicht zugeordnete Kodiereinheiten aus dem erhobenen Probematerial auf Probleme der Proband*innen beim Beantworten der Fragen zurückgeführt werden. Beim Probematerial ergab sich jeweils eine Abweichung für die Kodiereinheiten *Fachliche Korrektheit* (1) und *Vollständigkeit* (3).

Da sich das Item zur *Fachlichen Korrektheit* explizit auf die visuelle Ebene bezieht, hat sich ein Proband mit seiner Antwort ausschließlich auf optische Aspekte der Darstellung bezogen. Zumal im Modell jedoch nur optische und keine Informationen in Form von Text dargestellt werden, wurde dieser Bezug im Item *Fachliche Korrektheit* für die nächste Version des Bewertungsbogens entfernt. Zum Item *Vollständigkeit* wurde sich zur Verständlichkeit des Modells geäußert. Der Fokus lag jedoch darauf, wie ausführlich das Modell sein muss, um als gewinnbringend zu gelten. Da das Item auf die Vollständigkeit des Informationsgehalts des Modells abzielt, wurde das Item *Vollständigkeit* dahingehend abgeändert. Aus der Bearbeitung des Probematerials ergab sich zusätzlich eine Erweiterung des Gütekriteriums *Verständlichkeit*. Eine Testperson hat sich bei diesem Punkt stark auf das Vorwissen der Schüler*innen bezogen. Im schulischen Kontext stellt das anknüpfbare Vorwissen eine wichtige Voraussetzung für die Verständlichkeit dar (Stiller, 2003). Den Lernenden muss es möglich sein, den Inhalt des Modells mit ihrem bisherigen Wissen erfassen zu können. Nach Sumfleth (1992) bestimmt das Vorwissen, welche Informationen von Lernenden selektiert und verarbeitet werden, um infolgedessen Verbindungen zwischen den neuen Erkenntnissen und den vorhandenen Informationen herzustellen. Da das Vorwissen folglich eine unverzichtbare Variable für das Verständnis der Lernenden ist, wird die Beschreibung des Gütekriteriums *Verständlichkeit* um den Aspekt *Vorwissen* erweitert. Die abgeänderten Items mit der erweiterten Beschreibung der abgebildeten Gütekriterien sind im Folgenden dargestellt:

1) Ist das Phänomen fachlich korrekt dargestellt?

Dieses Item bildet die Eigenschaft der fachlichen Korrektheit nach Rauh und Stickel (1997) ab. Da ein Modell nur dann fachlich korrekt sein kann, wenn es in den wesentlichen Merkmalen mit dem Original übereinstimmt, wird auch das Kriterium der Ähnlichkeit und Entsprechung (Eschenhagen, Kattmann, & Rodi, 2008) von diesem Item erfasst.

2) Ist das Phänomen für die Zielgruppe verständlich dargestellt?

Mit diesem Item wird das Kriterium der Verständlichkeit nach Rauh und Stickel (1997) abgebildet. Wenn das Modell von der Zielgruppe verstanden wird, erlaubt dies Prognosen

über die originale Modellvorstellung zu abzuleiten. Der Inhalt des Modells muss mit dem bisherigen Vorwissen der Zielgruppe erfasst werden können. Da die für die Zielgruppe verständliche Darstellung Voraussetzung für die Prognosen ist, erfasst dieses Item auch das Kriterium der Exaktheit und Fruchtbarkeit (Eschenhagen, Kattmann, & Rodi, 2008).

3) *Sind alle Informationen zum Erreichen der Lernziele dargestellt?*

Dieses Item dient zur Überprüfung des Kriteriums der Vollständigkeit nach Rauh und Stickel (1997). Da eine ausreichend ausführliche Darstellung alle wesentlichen Merkmale des Modells abdeckt, bildet dieses Item auch das Prinzip der Adäquatheit (Eschenhagen, Kattmann, & Rodi, 2008) ab.

Aus der Ergänzung des Kriteriums *Vorwissen* ergibt sich für die Kategorie *Verständlichkeit* eine Unterteilung in zwei Subkategorien. Rauh und Stickel (1997) charakterisieren Verständlichkeit überwiegend darüber, wie ausdrucksfähig ein Modell ist. Im ausdifferenzierten Kategoriensystem (Anhang 18) bildet die Charakterisierung von Rauh und Stickel damit die erste Subkategorie *Ausdrucksfähigkeit* der *Verständlichkeit*. Die zweite Subkategorie stellt die induktiv aus dem Material ermittelte Kategorie *Vorwissen* dar.

Ergebnisse des zweiten Zyklus

Die Aussagen der Expert*innen konnten zu 98% den passenden Beschreibungen der Gütekriterien zugeordnet werden. Beim gesamten Material gab es nur bei einem Kodierenden eine Abweichung beim Material des ersten Probanden (s. Tabelle 9) in Form einer Zuordnung zu einer anderen Kategorie. Die Kodierung des Materials der restlichen Proband*innen zeigte keine weiteren Abweichungen. Für die vollständige Kodierung des gesamten Datenmaterials ergab sich eine exzellente (Koo & Li, 2016) Interraterreliabilität ($ICC_{3.1} = 0.998$, $p < 0.001$; Fleiss' $\kappa = 0.959$, $p < 0.001$). Die Werte für die Interraterreliabilitäten für jedes einzelne Item sind in Tabelle 10 aufgelistet. Mit Ausnahme des Kriteriums zur Übersichtlichkeit wurde überall eine Übereinstimmung von Fleiss' $\kappa = 1$ erreicht. Für das Item zum Übersichtlichkeitskriterium ergab sich ein Wert von Fleiss' $\kappa = 0.786$.

Tabelle 9: Ergebnisse der Kodierung des zweiten Zyklus (Anmerkung: Die Ziffer vor dem Punkt gibt die Kategorie an, die Ziffer hinter dem Punkt gibt die Subkategorie an; Abweichungen sind fett markiert)

Proband*in	Kodiereinheit	Coder 1	Coder 2	Coder 3
Am2	1	1	1	1
Am2	2	2.1	2.1	2.1
Am2	3	3	3	3
Am2	4	4	4	2
Am2	5	5	5	5
Af2	1	1	1	1
Af2	2	2.1	2.1	2.1
Af2	3	3	3	3
Af2	4	4	4	4
Af2	5	5	5	5
Bm2	1	1	1	1
Bm2	2	2.1	2.1	2.1
Bm2	3	3	3	3
Bm2	4	4	4	4
Bm2	5	5	5	5
Lf1	1	1	1	1
Lf1	2	2.1	2.1	2.1
Lf1	3	3	3	3
Lf1	4	4	4	4
Lf1	5	5	5	5

Tabelle 10: Interraterreliabilität für die Items der jeweiligen Gütekriterien

Item	Fleiss' κ	p
Fachliche Korrektheit	1	<0.001
Verständlichkeit	1	<0.001
Vollständigkeit	1	<0.001
Übersichtlichkeit	0.786	<0.001
Redundanzfreiheit	1	<0.001

Diskussion zum zweiten Zyklus

Die Auswertung des gesamten Materials zeigte, dass alle Kategorien im Kategoriensystem ausreichend ausführlich beschrieben wurden und die Kodiereinheiten eindeutig den entsprechenden Kategorien zugeordnet werden konnten. Ebenfalls ergaben sich im zweiten Zyklus weder Erweiterungen von Kategorienbeschreibungen noch Angaben zu bisher nicht gefundenen Kategorien. Daraus lässt sich schließen, dass die Theorie als gesättigt angesehen werden kann und keine weiteren Daten erhoben werden müssen (Glaser & Strauss, 1967; Marshall, 1996). Der Stichprobenumfang gilt damit als ausreichend groß, zur Validierung der Fragebogenitems (Döring & Bortz, 2016). Basierend auf der Interraterreliabilität zu den Expert*inneninterviews lässt sich die Inhaltsvalidität erfassen. Für die Interraterreliabilität wurde im zweiten Validierungszyklus ein exzellenter Wert erzielt ($ICC_{3.1} > 0.9$, Koo & Li, 2016). Daraus ergibt sich sowohl ein hohes Maß für die Reliabilität als auch für die Objektivität der Datenauswertung, wenn diese Art der Datenauswertung auch stets mit einem subjektiven Bias einhergeht (Döring & Bortz, 2016). Aus der hohen Übereinstimmung von 98 % der Antworten mit den Beschreibungen der Gütekriterien und der exzellenten Interraterreliabilität leitet sich eine sehr gute Inhaltsvalidität ab (Sandmann, 2014). Generell gilt für Daten, die mithilfe von lautem Denken erhoben wurden, eine immanent hohe Validität (Brückmann & Duit, 2014). Folglich kann Hypothese H1 verifiziert werden, da die Items die Gütekriterien aus der Primärliteratur valide erfassen und abbilden. Die Abweichung zum Item zur *Verständlichkeit* lässt sich theoriekonform daher begründen, dass *Übersichtlichkeit* nach Rauh und Stickel (1997) eine Bedingung für *Verständlichkeit* ist und daher als Teildimension angesehen werden kann. Zudem entspricht eine Interraterreliabilität von $\kappa > 0.75$ einer sehr guten Übereinstimmung (Wirtz & Caspar, 2002), wodurch sich die Validität des Items zur Verständlichkeit bestätigen lässt (Schmiedemann & Lücken, 2014). Aus den Interraterreliabilitäten der einzelnen Items lässt sich schließen, dass der Fragebogen zur Modellqualität eine valide Erfassung der jeweiligen Gütekriterien ermöglicht. Die Aussagekraft der Ergebnisse kann durch Kritik sowohl an den Methoden der Datenerhebung als auch Datenauswertung eingeschränkt werden. Zunächst bringt die Methode des lauten Denkens inhärente Kritikpunkte mit sich. Nicht alle kognitiven Vorgänge können tatsächlich erhoben werden, da ein bestimmter Anteil unbewusst abläuft und daher nicht verbalisiert wird (Funke & Spering, 2006). Zudem besteht die Gefahr, dass die Verbalisierung der kognitiven Prozesse ebendiese beeinflusst (ebd.). Weiterhin können selbst bei einer Gesprächsdauer von nur sechs bis zehn Minuten die Gedanken mit zunehmender Testdauer weniger ausführlich verbalisiert werden (Sandmann, 2014). Die teilnehmenden Versuchspersonen der Studie I sind mittels Gelegenheitsstichproben statt durch Zufallsstichproben ermittelt worden. Die Stichprobe der Expert*innenbefragung aus Studie I ergab sich aus bestehenden Kooperationen mit lokalen Schulen und Arbeitsgruppen innerhalb des Instituts

für naturwissenschaftliche Bildung der RPTU Kaiserslautern-Landau. Daher setzt sich die Stichprobe aus Expertinnen und Experten zusammen, die didaktischen Neuerungen eher offen gegenüberstehen und nicht als unbefangen von sozialer Erwünschtheit anzusehen sind.

6.3 Reliabilitätsbestimmung des Messinstruments

Methodisches Vorgehen

Um eine geeignete Methode zur Überprüfung der Zuverlässigkeit der Kurzskala des BeQuAn zu finden, müssen die unterschiedlichen Verfahren zur Reliabilitätsbestimmung gegeneinander abgewogen werden. Aufgrund der Unterschiede im lauten Denken der Proband*innen zu den unterschiedlichen Formulierungen der Items in den beiden Validierungszyklen, wurde sich gegen eine Bestimmung der Paralleltestreliabilität entschieden. Zudem spricht dies dafür, dass die Formulierung weiterer valider Items sich als schwierig erweisen und zudem eine erneute Überprüfung der Validität erfordern würde. Ebenso wird die Bestimmung der internen Konsistenz abgelehnt, da statt einer psychometrischen Skala eine Ratingskala für Objekte vorliegt (Bühner, 2006). Daher wurde sich für die Bestimmung der Test-Retest-Reliabilität entschieden. Dazu wurden 16 Studierende des Lehramts Chemie (5 Bachelor, 11 Master, ♀ = 56 %, Alter: $MW = 23$, $SD = 2$) zur Bewertung des haptisch-interaktiven Modells zur Bildung der Hydrathülle in einem Online-Fragebogen (LimeSurvey) im Zeitraum vom 11.01.21 bis 13.07.21 befragt. Zwischen den Testzeitpunkten von Test und Retest lagen mindestens zwei bis maximal drei Monate Abstand.

Die Auswertung der Daten erfolgte über die Bestimmung der Intraraterreliabilität zwischen Testzeitpunkt eins und zwei. Da die Rater*innen aufgefordert sind mehreren Bewertungsstufen zu wählen, liegt ein ordinale Skalenniveau vor. Daher sollten größere Abweichungen zwischen den Werten der Testzeitpunkte untereinander stärker gewichtet werden als kleinere Differenzen. In diesem Fall sollte das gewichtete Kappa als Spezialfall des Intraklassen-Korrelationskoeffizienten (Fleiss & Cohen, 1973) verwendet werden (Bortz & Schuster, 2010). Zudem wird bei einer Stichprobengröße von weniger als 15 Versuchspersonen die Bestimmung der Intraklassen-Korrelation ICC empfohlen (Everitt & Skrondal, 2010). Da die Kodierenden zufällig ausgewählt wurden und die Kodierungen aller beteiligten Kodierenden in die Auswertung mit eingeflossen sind, ergibt sich die ICC2.1 als geeignetes Reliabilitätsmaß (Shrout & Fleiss, 1979). Zur Bestimmung der Intraraterreliabilitäten wurde die Software *R-Studio* (Version 2021.09.1) mit R (Version 3.14.0) unter Verwendung der Pakete *psych* und *irr* genutzt.

Ergebnisse der Reliabilitätsbestimmung

Für die Gesamtintraraterreliabilität ergab sich Werte für ICC2.1 von 0.83 ($p < 0.001$) und Cohen's κ von 0.83 ($p < 0.001$). Die Werte für die einzelnen Gütekriterien sind in Tabelle 11 aufgelistet.

Tabelle 11: Kennwerte für die Intraraterreliabilitäten zu den Gütekriterien

	Fachliche Korrektheit	Verständlichkeit	Vollständigkeit	Übersichtlichkeit	Redundanzfreiheit
ICC2.1	0.92	0.85	0.66	0.78	0.88
Cohen's κ	0.92	0.84	0.64	0.76	0.88

Diskussion zur Reliabilitätsbestimmung

Die Stichprobe setzt sich aus Teilnehmer*innen fachdidaktischer Seminare, in denen die Modellmethode gelehrt und kritisiert wird. Damit ist die Stichprobe als Gelegenheitsstichprobe anzusehen. Der Pre-Test erfolgte stets direkt im Anschluss an die Lehrveranstaltungen zum Thema Modelle. Dadurch ist jedoch nicht gewährleistet, dass die Studierenden bereits ausreichend Kompetenz zur Modellkritik erworben haben oder sich ihre Fähigkeiten zwischen den Messzeitpunkten (MZP) nicht weiterentwickelt haben. Damit unterscheidet sich die Stichprobe von der eigentlichen Zielgruppe in Form von voll ausgebildeten Lehrkräften als erprobte Expert*innen. Weiterhin ist anzumerken, dass die Antworten der Teilnehmenden aufgrund von sozialer Erwünschtheit (Paulhus, 2002) beeinflusst sein können. Jedoch ist dieser Einfluss zu beiden Testzeitpunkten als gleich groß einzuschätzen. Daher ist eine Beeinflussung von Intraklassen-Korrelationskoeffizienten nicht zu erwarten (Fleiss & Cohen, 1973). Die Werte der Intraraterreliabilitäten für alle Kriterien und die Gesamtskala befinden sich in einem guten bis sehr guten Bereich (ICC: gut = 0.6 bis 0.74, sehr gut > 0.75 , Cicchetti, 1994; Cohen's κ : gut = 0.6 bis 0.74, sehr gut > 0.75 , Wirtz & Caspar, 2002). Beim Test-Retest Verfahren besteht zusätzlich das Problem, dass die Befragten dazu verleitet sind, bei der Retestung dieselben Antworten wie zum ersten Testzeitpunkt zu geben, um sicher in ihrer Einschätzung zu wirken (Moosbrugger & Kelava, 2012). Dies hat zur Folge, dass eigentliche Messfehler unterschätzt und die Reliabilität überschätzt werden. Die Bipolarität und gerade Stufung der Likert-Skala wirken einer Überschätzung der Reliabilität jedoch entgegen (Bühner, 2006). Folglich weisen die guten bis sehr guten Werte für die Intraraterreliabilität darauf hin, dass das Bewertungsinstrument eine reliable Messung der Qualität von Anschauungsmodellen ermöglicht. Die statistische Aussagekraft der Untersuchungen kann aufgrund möglicher Verzerrungen durch die verwendete Stichprobe eingeschränkt sein. Die Schlussfolgerungen, die aus den Ergebnissen gezogen werden können, sind daher möglicherweise eingeschränkt belastbar (Berk & Freedman, 2003).

6.4 Implikationen aus Studie I für die weiteren Untersuchungen

Im Rahmen von Studie I wurde die Bedeutung eines fundierten Maßstabs zur Beurteilung der Modellqualität hervorgehoben. Die Ergebnisse der Untersuchungen zeigen, dass es möglich ist, Qualitätskriterien zur Bewertung von Anschauungsmodellen zu formulieren und diese in ein valides und reliables Messinstrument zu übertragen. Die in Studie I gewonnenen Erkenntnisse zur Modellbewertung bilden somit die Grundlage für die Konzeption und Bewertung der Modelle anhand valider Kriterien. Dies ermöglicht zunächst die Entwicklung qualitativ hochwertiger Anschauungsmodelle. Mithilfe des entwickelten Messinstruments lassen sich die entwickelten Anschauungsmodelle zuverlässig und zielsicher bewerten. Durch die kriteriengeleitete Entwicklung der Modelle für die Studien II und III wird zudem die interne Validität dieser Untersuchungen erhöht, da die Konzeption und Bewertung der Modelle anhand der gleichen, für das Lernen mit der Modellmethode relevanten, Kriterien erfolgt.

7. Studie II: Gütebestimmung der Qualität der drei Modelltypen haptisch-interaktiv, AR und IL mittels Expert*innenratings

Im Rahmen von Interventionsstudien, welche einen Vergleich verschiedener Medien zum Gegenstand haben, ist ein fairer Vergleich eine notwendige Voraussetzung, um valide Aussagen über den Einfluss des jeweiligen Mediums treffen zu können. Um einen fairen Vergleich während der Untersuchung zur Wirksamkeit der drei Modelltypen digital-erweitert, haptisch-interaktiv & illustrativ zu garantieren, muss eine gleichwertige Modellgüte gewährleistet sein (Tversky et al., 2002).

7.1 Methoden der Datenerhebung

Die Ergebnisse der Studie II zielen darauf ab, eine Antwort auf die Frage zu finden, ob die entwickelten Modelle der drei Modelltypen eine gleichwertige Qualität aufweisen (Przywarra et al., 2021). Dazu wurde folgende Hypothese überprüft:

(H2) Die Gesamtgüte der Modelle unterscheidet sich nicht signifikant in Abhängigkeit des Modelltyps.

Da die einzelnen Kriterien der Modellqualität auf Beobachtungen mit geringem Komplexitätsgrad beruhen, fällt die Wahl der Bewertungsmethode auf die des Urteils von Fachexpert*innen (Döring & Bortz, 2016). Als Expert*innen wurden 15 aktive Lehrpersonen der Fächer Naturwissenschaften oder Chemie aus Rheinland-Pfalz als Expert*innen aus dem Kreis der Beraterinnen und Berater für naturwissenschaftlichen Unterricht des Pädagogischen Landesinstituts Rheinland-Pfalz rekrutiert (♀ = 60 %; Alter: MW = 42.5 Jahre, SD = 10.9; Lehrerfahrung: MW = 13.5 Jahre, SD = 9.4). Die Datenerhebung erfolgte per Paper-Pencil-Befragung mittels des validierten BeQuAn-Fragebogens. Zusätzlich bestand die Möglichkeit, Kommentare zu den einzelnen Kriterien jedes bewerteten Modells abzugeben. Aufgrund der Lockdown-Situation während des Erhebungszeitraums im April 2020 wurden die Modelle und die Bewertungsbögen postalisch an die Proband*innen verschickt. Um einen standardisierten Gebrauch der Modelle und des Bewertungsbogens zu gewährleisten, wurden die Expert*innen mittels eines zehnminütigen Tutorial-Videos durch die Bewertung der Modelle geführt (Link: <https://youtu.be/FYmyij92vCg>). Zur Vermeidung von Reihenfolgeeffekten, wurde dabei die chronologische Abfolge, in der die Modelle bewertet wurden, variiert. Die Proband*innen bewerteten zu jedem Modelltyp jeweils sechs unterschiedliche Modelle, die jeweils ein anderes Phänomen auf Teilchenebene zeigen.

7.2 Methoden der Datenauswertung

Weichen Werte um mehr als den dreifachen Interquartilsabstand nach unten vom ersten Quartil bzw. nach oben vom dritten Quartil ab, werden diese als Extremwerte definiert (Eid et al., 2015). Sofern Extremwerte im Datensatz vorlagen, wurden diese nach Ausschluss von Eingabefehlern einzelfallweise gelöscht. Bei fehlenden Werten wurde die betreffende Person für die Auswertung der einzelnen Kriterien fallweise ausgeschlossen. Zur Bestimmung der Gesamtgüte wurden einzelne fehlende Werte durch die Mittelwerte der betroffenen Personen zum jeweiligen Modell imputiert. Zur Analyse der Antwortverteilungen wurden die deskriptiven Kriterien Mittelwert, Standardabweichung, Minimum, Maximum, Schiefe, Kurtosis und Itemschwierigkeit nach Dahl betrachtet. Für den Betrag der Schiefe gilt ein Überschreiten des Grenzwerts von 2 und für den Betrag der Kurtosis von 7 als gravierende Abweichung der Normalverteilung (Eid et al., 2015). Die Grenzwerte für die Itemschwierigkeit liegen bei $20 < P_{\text{Dahl}} < 80$ (Döring & Bortz, 2016; Kelava & Moosbrugger, 2012). Aufgrund der ordinalskalierten Daten erfolgte die weitere Datenauswertung anhand von Unterschiedsprüfungen der unabhängigen Stichproben mittels Kruskal-Wallis-Test (Field et al., 2012). Als post-hoc Verfahren wurde der paarweise Mann-Whitney U Test angewendet (ebd.). Die Effektgrößen geben bei partiellem $\eta^2 > 0.14$ einen großen, bei $\eta^2 > 0.06$ einen mittleren oder bei $\eta^2 > 0.01$ einen kleinen Effekt an (Cohen et al., 2013). Zunächst wurde untersucht, ob sich die Modellqualität innerhalb eines Modelltyps signifikant unterscheidet oder alle Modelle im Mittel über alle Gütekriterien hinweg eine vergleichbare Qualität aufweisen. Im zweiten Schritt galt es zu prüfen, ob ein Unterschied zwischen den Modelltypen im Mittel über alle Modelle hinweg und spezifisch für jedes Modell einzeln unabhängig von den einzelnen Gütekriterien vorliegt (vgl. Hypothese 2). Weiterhin wurden Unterschiede zwischen den Modelltypen im Mittel über alle Modelle hinweg spezifisch pro Gütekriterium analysiert. Die deskriptiven und inferenzstatistischen Auswertungen erfolgten mittels der Software *R-Studio* (Version 2021.09.1) mit R (Version 3.14.0) unter Verwendung der Pakete *psych*, *dplyr*, und *stats*. Zur abschließenden Prüfung der Aussagekraft der Ergebnisse zur Unterschiedsprüfung der Gesamtgüte der drei Modelltypen gilt es zu prüfen, mit welcher Wahrscheinlichkeit ein Fehler zweiter Art vorliegt. Dazu wurde mittels G*Power (Version 3.1.9.4) eine Post-Hoc-Poweranalyse durchgeführt (Cohen, 1988). Zur Bestimmung der Effektstärke f wurden die Mittelwerte der Gesamtgüte der Gruppen (IL = 2.44, HI = 2.66, AR = 2.68) und die gemittelte Standardabweichung über alle drei Gruppen ($\sigma = 0.44$) verwendet (Rasch et al., 2014a).²

² Dieser Absatz wurde bereits in Przywarra et al. (2021) veröffentlicht, modifiziert.

Zur Klärung der Frage nach den Ursachen für die Kritik der Expert*innen wurden die Kommentare nach dem Prinzip der Grounded Theory nach Charmaz (2006) kodiert. Als Kodieren wird hier der Prozess der intellektuellen Bearbeitung des Materials mittels Generierens und Zuteilen von Codes verstanden. Codes sind dabei Schilderungen, die den inhaltlichen Kern der Kodiersegmente abbilden. Die Kommentare der Expert*innen wurden inhaltlich zusammengefasst und auf die einzelnen Gütekriterien bezogen, dargestellt. Den Proband*innen fehlt die Information, welche Kritikpunkte zu jedem Modell in der abschließenden Modellkritik am Ende der Interventionen besprochen werden. Daher wurden Kritikpunkte, die aufgrund dieses fehlenden Hintergrundwissens subjektiv als nicht zielführend eingestuft wurden und in geringer Anzahl vertreten sind (<5%), von der inhaltlichen Darstellung ausgeschlossen. Mithilfe dieser Abbruchkriterien wurde das gesamte Datenmaterial bearbeitet. Aufgrund des explorativen Charakters der betrachteten Frage, sieht die Grounded Theory keine kategorienbasierte Analyse des Materials vor (Kuckartz, 2016).

7.3 Ergebnisse

Die Prüfung auf Extremwerte zeigt, dass Ausreißer vorliegen, welche jedoch nach Eid et al. (2015) nicht als Extremwerte definiert werden konnten. Die Unterschiedsprüfung zwischen den Modellen jeweils innerhalb des Modelltyps ergab keine signifikanten Unterschiede (vgl. Tabelle 13, Zeile 1-3). Die gemittelte Qualität über sämtliche Gütekriterien und alle sechs Modelle hinweg unterscheidet sich signifikant in Abhängigkeit des Modelltyps (vgl. Tabelle 13, Zeile 4). Während die AR-Modelle sowie die haptisch-interaktiven Modelle qualitativ gleichwertig eingeschätzt werden, ist die Gesamtgüte der illustrativen Modelle mit einem kleinen Effekt ($\eta^2 = 0.04$) signifikant geringer bewertet worden. In der post-hoc Analyse zeigte sich, dass dieser Effekt sowohl auf dem Unterschied zwischen den AR-Modellen und den illustrativen Modellen ($p < 0.01$) als auch auf dem Unterschied zwischen den haptisch-interaktiven Modellen und den illustrativen Modellen ($p < 0.05$) beruht. Dieser Effekt wurde durch die nicht signifikanten Unterschiede zwischen den einzelnen Modellen eines Modelltyps aufsummiert (vgl. Tabelle 13, Zeile 5-10). Der Vergleich auf der Ebene der einzelnen Gütekriterien ergab für die Kriterien *Fachliche Korrektheit*, *Vollständigkeit*, *Übersichtlichkeit* und *Redundanzfreiheit* keine signifikanten Gruppenunterschiede zwischen den Modelltypen (vgl. Tabelle 13, Zeile 11-14). Im Kriterium der *Verständlichkeit* unterschied sich allerdings die Qualität der Modelle in Abhängigkeit vom Modelltyp signifikant mit einem großen Effekt (vgl. Tabelle 13, Zeile 15). Laut post-hoc Analyse ist dieser Effekt nur auf den Unterschied zwischen den AR-Modellen und den illustrativen Modellen zurückzuführen ($p = 0.02$). Die Ergebnisse sind in Tabelle 13 und Tabelle 14 sowie in Abbildung 14 zusammengefasst und nach den drei Modelltypen, den fünf Gütekriterien sowie der Gesamtgüte aufgeschlüsselt.³

Tabelle 12: Deskriptive Kennwerte der Items und Gesamtskala des BeQuAn

Bezugsskala	MW (SD)	Median	Min	Max	Range	Schiefe (<2)	Kurtosis (<7)	P _{Dahl} (20-80)
Fachliche Korrektheit	2.48 (0.50)	2.67	1.33	3	1.67	-0.79	-0.52	61.90
Verständlichkeit	2.47 (0.48)	2.6	1	3	2	-0.99	0.51	61.71
Vollständigkeit	2.55 (0.50)	2.67	0.67	3	2.33	-1.5	2.64	63.85
Übersichtlichkeit	2.65 (0.45)	2.83	1.33	3	1.67	-1.43	1.29	66.26
Redundanzfreiheit	2.8 (0.31)	3	1.83	3	1.17	-1.54	1.48	70.07
Gesamtgüte	2.59 (0.47)	2.67	0.67	3	2.33	-1.3	1.36	(64.76)

³ Dieser Absatz wurde bereits in Przywarra et al. (2021) veröffentlicht, modifiziert.

Tabelle 13: Ergebnisse der inferenzstatistischen Auswertungen mit dem Kruskal Wallis Test (Przywarra et al., 2021, S. 32)

	Zur Unterschiedsprüfung betrachtete Parameter	Anzahl der Freiheitsgrade	χ^2	p	η^2
1	IL: Modelle 1-6	4	4.26	> 0.05	< 0.01
2	HI: Modelle 1-6	4	5.05	> 0.05	0.02
3	AR: Modelle 1-6	4	6.14	> 0.05	0.03
4	Gemittelte Gütekriterien: IL, HI, AR	2	10.33	0.006*	0.04*
5	Diffusion: IL, HI, AR	2	3.42	> 0.05	0.03
6	Temp.abh. Diffusion: IL, HI, AR	2	2.13	> 0.05	< 0.01
7	Salzlösung: IL, HI, AR	2	4.21	> 0.05	0.05
8	Abdampfen: IL, HI, AR	2	2.45	> 0.05	0.01
9	Teilchensieb: IL, HI, AR	2	5.91	> 0.05	0.09
10	Osmose: IL, HI, AR	2	3.77	> 0.05	0.04
11	Fachliche Korrektheit: IL, HI, AR	2	0.38	> 0.05	0.04
12	Vollständigkeit: IL, HI, AR	2	3.11	> 0.05	0.03
13	Übersichtlichkeit: IL, HI, AR	2	1.81	> 0.05	< 0.01
14	Redundanzfreiheit: IL, HI, AR	2	3.42	> 0.05	0.03
15	Verständlichkeit: IL, HI, AR	2	8.08	0.02*	0.15*

Anmerkungen: IL = Illustrative Modelle; HI = haptisch-interaktive Modelle; AR = Augmented Reality Modelle; Temp.abh. = Temperaturabhängigkeit; Signifikante p-Werte und die zugehörigen Effektgrößen sind mit „*“ markiert

Tabelle 14: Bewertung der einzelnen Gütekriterien sowie der Gesamtgüte (Przywarra et al., 2021, S. 33)

	Illustrative Modelle	Haptisch-interaktive Modelle	Augmented Reality Modelle
Fachliche Korrektheit	2.40 (0.64)	2.51 (0.33)	2.56 (0.46)
Verständlichkeit	2.21 (0.52)	2.53 (0.37)	2.69 (0.38)
Vollständigkeit	2.34 (0.63)	2.66 (0.37)	2.67 (0.41)
Übersichtlichkeit	2.53 (0.51)	2.81 (0.23)	2.62 (0.54)
Redundanzfreiheit	2.70 (0.38)	2.81 (0.26)	2.89 (0.27)
Gesamtgüte	2.44 (0.56)	2.66 (0.33)	2.68 (0.42)

Anmerkung: Dargestellt sind die Mittelwerte mit den dazugehörigen Standardabweichungen in Klammern

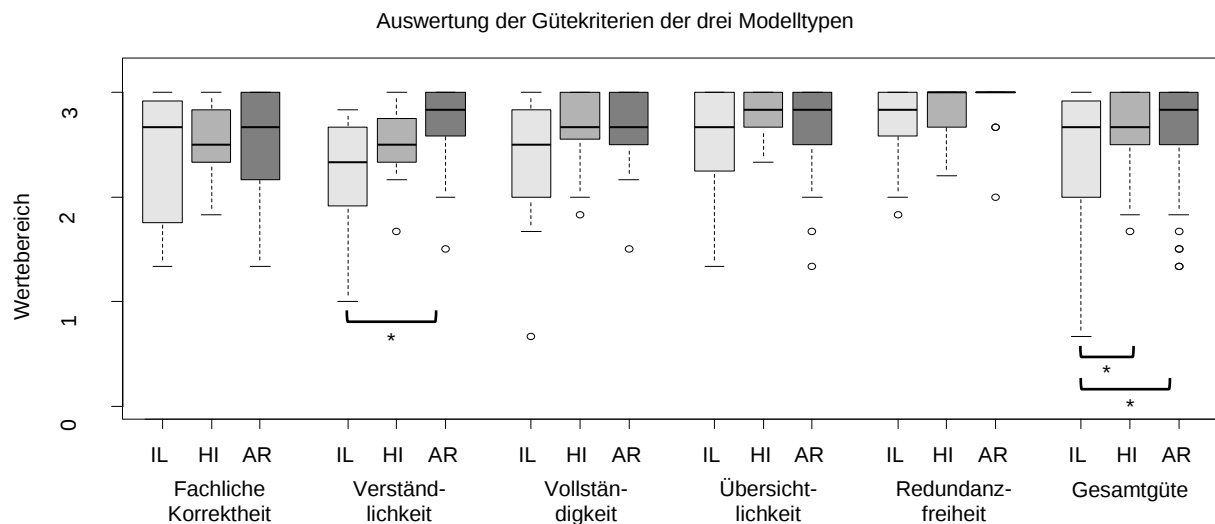


Abbildung 14: Gemittelte Gesamtgüte sowie die fünf Gütekriterien der drei Modelltypen IL = Illustratives Modell, HI = Haptisch-interaktives Modell und AR = Augmented Reality Modell; Wertebereich: 0 = „stimme nicht zu“, 1 = „stimme eher nicht zu“, 2 = „stimme eher zu“, 3 = „stimme zu“; * = $p < 0.05$ (Przywarra et al., 2021, S. 33)

Unter Berücksichtigung der kalkulierten Effektstärke von $f = 0.247$ sowie der Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.05 (Cohen, 1988) lässt sich für die Gesamtstichprobe von 225 Observationen ($\triangleq 15$ Rater*innen * 3 Modelltypen/Gruppen * 5 Gütekriterien) eine Power von 0.919 ableiten. Folglich liegt die Wahrscheinlichkeit eines Fehlers zweiter Ordnung bei 8,1% und die Power des Tests gilt als ausreichend (Cohen, 1988; Eid et al., 2015). Bei einer Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.05 und einer Power von 0.8 ergibt sich für die ermittelte Effektstärke eine Gesamtstichprobengröße von 162 Observationen. Folglich wäre eine Stichprobengröße von elf Expert*innen ausreichend, um die gefundenen Effekte in der Unterschiedlichkeitsprüfung aufzudecken.

Im Zuge der Bewertung durch die Expert*innen gaben diese insgesamt 180 Kommentare zu den Modellen ab. Von den insgesamt 15 Expert*innen haben 11 ihre Bewertungen kommentiert. Die Kommentare verteilten sich dabei wie folgt auf die Modelltypen: 26,7% zu AR, 25% HI, und 48,3% IL (vgl. Tabelle 15).

Tabelle 15: Anzahl der Kommentare zu den Modelltypen aufgelistet nach den Gütekriterien (die Prozentzahl zu den einzelnen Kriterien bezieht sich immer auf den Anteil an der Gesamtanzahl [N = 180] der Kommentare)

	Illustrativ	Haptisch-Interaktiv	Augmented Reality
Fachliche Korrektheit	24 (13,3%)	19 (10,5%)	19 (10,5%)
Verständlichkeit	24 (13,3%)	15 (8,3%)	10 (5,5%)
Vollständigkeit	18 (10%)	6 (3,3%)	1 (0,5%)
Übersichtlichkeit	12 (6,7%)	1 (0,5%)	13 (7,2%)
Redundanzfreiheit	9 (5%)	4 (2,2%)	5 (2,8%)
Summe	87 (48,3%)	45 (25%)	48 (26,7%)

Inhaltlich lassen sich die Aussagen wie folgt zusammenfassen (vgl. Tabelle 16). Die Kommentare zur fachlichen Korrektheit stellen in erster Linie Kritik an den Größenverhältnissen der Teilchen dar. Zu den Modelltypen AR und IL äußern die Expert*innen zusätzlich Kritik an den Abständen zwischen den Teilchen. Für den Typ IL wird Kritik geäußert, die in der Einschränkung auf einen festen Blickwinkel begründet liegt. Die Kritik an der Verständlichkeit der Modelle bezieht sich bei HI auf die Geschwindigkeit, in der das Phänomen veranschaulicht wird. Zu den IL-Modellen wird die Verständlichkeit aufgrund der Beschränkung auf einen Blickwinkel kritisiert. Die Verständlichkeit der AR-Modelle wird laut Kritik durch eine zu große Freiheit der Nutzer bei der Wahl der Geschwindigkeit und der Wahl des Bildausschnitts begrenzt. Kritikpunkte zur Vollständigkeit werden beim Modelltyp IL ebenfalls zum Blickwinkel geäußert. Zum Kriterium Übersichtlichkeit wird der inhomogene Hintergrund der AR-Modelle und das Fehlen weiterer Blickwinkel in den IL-Modellen angemerkt. Zur Redundanzfreiheit wird bei den IL-Modellen angemerkt, dass einige Standbilder in der Bildfolge dieselben Inhalte zeigen wie andere.

Tabelle 16: Kategorien zu den Kritikpunkten der drei Modelltypen

Modelltyp	Kategorie	Code	Ankerbeispiel
Illustrativ	Fachliche Korrektheit	Blickwinkel vermittelt Größenfehler	„Aus diesem Winkel könnte man den Eindruck haben, dass nicht alle Teilchen gleich aussehen.“
		Falsche Teilchenabstände	„Die Teilchenabstände sind so groß wie im flüssigen Zustand.“
	Verständlichkeit	Eingeschränkte Dreidimensionalität	„Die Dreidimensionalität der Hydrathülle ist nicht erkennbar.“
	Vollständigkeit	Eingeschränkte Dreidimensionalität	„Es fehlt die Perspektive, um den Vorgang der Diffusion in alle Raumrichtungen nachvollziehen zu können!“
	Übersichtlichkeit	Eingeschränkter Blickwinkel	„Aus diesem Winkel kann man manches nur schwer erkennen.“
Haptisch-Interaktiv	Redundanzfreiheit	Gleicher Inhalt	„Bild 1 und 7 zeigen das gleiche.“
	Fachliche Korrektheit	Größenfehler der Teilchen	„Die Chlorid-Ionen müssten größer als die Natrium-Ionen sein“
Augmented Reality	Verständlichkeit	Eingeschränkte Phänomendauer	„Der Zeitpunkt, an dem das Wichtige passiert, ist schon sehr kurz.“
	Fachliche Korrektheit	Falsche Teilchenabstände	„Die Abstände zwischen den Teilchen sind zu groß für den flüssigen Zustand.“
	Verständlichkeit	Fehlende Fokussierung	„Man weiß nicht, worauf man sich fokussieren soll“
	Übersichtlichkeit	Inhomogener Hintergrund	„Das Fehlen eines Hintergrunds, macht es schwer alle Teilchen zu verfolgen.“

7.4 Diskussion zu Studie II

Die deskriptiven Kennwerte der Items und der Gesamtskala des BeQuAn (vgl. Tabelle 12) zeigen keine extreme Kurtosis (<7) oder Schiefe (<2) der Verteilungen. Zusätzlich sprechen die Kennwerte für die Itemschwierigkeit nach Dahl gegen das Vorliegen von Decken- oder Bodeneffekten (Tepner & Dollny, 2014). Aufgrund der Ergebnisse der Post-Hoc-Poweranalyse wird die Stichprobengröße als ausreichend groß und angemessen ökonomisch angesehen (Kraft, 2020). Die Ergebnisse des Expert*innenratings zu den drei Modelltypen zeigen, dass die Inhalte der Phänomene mit den unterschiedlichen Modelltypen nur mit gewissen Einschränkungen gleichwertig abgebildet werden können. Dennoch werden die Modelle innerhalb eines Modelltyps als qualitativ vergleichbar eingeschätzt. Dies liefert eine notwendige Voraussetzung für die Interpretation der hypothesenbezogenen Ergebnisse, da andernfalls nicht auf eine übergeordnete Vergleichbarkeit der Modelltypen rückgeschlossen werden könnte, sondern möglicherweise lediglich gegenläufige Unterschiede von spezifischen Modellen vorliegen. Allerdings wurde die Qualität der illustrativen Modelle mit einem kleinen Effekt signifikant geringer bewertet, weshalb die Hypothese zur Forschungsfrage falsifiziert werden muss. Dieser Effekt findet sich deskriptiv und inferenzstatistisch in den einzelnen Modellen des illustrativen Modelltyps und allen Gütekriterien wieder. Folglich besteht bisher noch keine Vergleichbarkeit zwischen den drei Modelltypen. Limitationen der Studie sind möglicherweise nicht erfasste Persönlichkeitsmerkmale der Expert*innen, die einen Einfluss auf die Bewertung der Modelle haben könnten. Beispielsweise könnte bei den Proband*innen das Selbstkonzept zum Umgang mit digitalen Medien schwach ausgeprägt sein (Dickhäuser, 2001), was sich eventuell auf die Bewertung der AR Modelle auswirken könnte. Ebenso könnten die Einstellungen der Lehrkräfte zu den Modelltypen eine vorurteilsbehaftete Bewertung mit sich bringen. Weiterhin ist nicht auszuschließen, dass die Bewertung durch einen Neuheitseffekt beeinflusst wurde (Theyßen, 2014), da die Expert*innen möglicherweise wenig Vorerfahrungen beispielsweise mit haptisch-interaktiven oder AR-Modellen aufweisen.⁴ Weitere Kritikpunkte ergeben sich für die Standardisierung der Erhebung. Aufgrund der Pandemiesituation wurden die Modelle zu den Befragten nachhause geschickt und auch dort von ihnen bewertet. Die Erhebung in unterschiedlichen Settings führt zu variierenden externen Einflüssen. Diese Einflüsse resultierten in unvermeidbaren Variationen der Untersuchungsbedingungen, wie sie typischerweise in Feldstudien auftreten (Döring & Bortz, 2016). Dadurch konnten Variablen wie der Ort, die Testzeit, Pausen und die Reihenfolge der bewerteten Modelle nicht kontrolliert werden. Zur Vermeidung von Reihenfolgeeffekten wurde die Reihenfolge der Modelltypen in den Bewertungsbögen variiert. Um eine Standardisierung der Bewertung zu gewährleisten, wurden die

⁴ Dieser Absatz wurde bereits ins Przywarra et al. (2021) veröffentlicht, modifiziert.

Befragten mithilfe eines Video-Tutorials durch die Verwendung und Bewertung der Modelle geführt. Für das Expert*innenrating in Studie II kam die Stichprobe auf Basis freiwilliger Teilnahme von Lehrpersonen aus dem Kreis der Beraterinnen und Berater für naturwissenschaftlichen Unterricht des Pädagogischen Landesinstituts Rheinland-Pfalz zustande. Somit handelt es sich um eine Gelegenheitsstichprobe, die nicht als repräsentativ für die Gesamtpopulation angesehen werden kann. Folglich wird die Generalisierbarkeit der Ergebnisse eingeschränkt (Döring & Bortz, 2016). Weiterhin setzt sich die Stichprobe aus Expertinnen und Experten zusammen, die didaktischen Neuerungen eher offen gegenüberstehen und nicht als unbefangen von sozialer Erwünschtheit anzusehen sind. Somit kann durch die Auswahl der Teilnehmenden aufgrund ihrer Erreichbarkeit können systematische Verzerrungen in der Stichprobe auftreten (Etikan et al., 2016). Da jedoch stets alle Modelle von allen Expert*innen bewertet wurden, wird der Einfluss auf die Gruppenvergleiche als geringfügig eingeschätzt.

Die Quantifizierung der Kategorien der Expert*innenkommentare ermöglicht eine methodeninterne Triangulation der Ratingergebnisse. Die Verteilung der Kommentare steht dabei inkongruent im Kontrast zu den Mittelwerten der Expert*innenratings innerhalb der einzelnen Kategorien. Eine ausführliche, qualitative Inhaltsanalyse der Expert*innenkommentare unter Verwendung der induktiv gebildeten Codes könnte mehr Aufschluss über das Verbesserungspotenzial geben. Da jedoch lediglich die illustrativen Modelle verbessert werden müssen, um eine Gleichwertigkeit der Modelle zu erreichen und die haptisch-interaktiven Modelle sich nicht besser konstruieren lassen, wird dieses Vorgehen hier als ausreichend eingeschätzt. Die identifizierten Unterschiede liegen zudem in der Natur der jeweiligen Modelltypen (Wu & Puntambekar, 2012). Darüber hinaus wurde der Großteil der Kritik zu allen drei Modelltypen zur *Fachlichen Korrektheit* geäußert. Da diese Kritikpunkte jedoch in der abschließenden Modellkritik der Intervention zusammen mit den Lernenden aufbereitet werden, ist deren Berücksichtigung im Zuge der Auswertung nicht zielführend. Die geäußerte Kritik an den illustrativen Modellen zur *Verständlichkeit*, *Vollständigkeit* und *Übersichtlichkeit* basiert vorwiegend auf Problemen bezüglich der jeweils abgebildeten Perspektiven. In Anbetracht dieser Kritik werden die illustrativen Modelle um Abbildungen aus einem anderen Blickwinkel ergänzt (Przywarra & Risch, 2021). Da somit die einzige, unwiderlegbare Kritik an den illustrativen Modellen umgesetzt wird,⁵ können die illustrativen Modelle folglich als gleichwertig zu den anderen Modellen angesehen werden. Daher wird aus testökonomischen Gründen auf eine weitere Bewertung der überarbeiteten Modelle verzichtet und eine gleichwertige Modellgüte der Modelltypen angenommen.

⁵ Die Kritik an der Redundanzfreiheit wird nicht berücksichtigt, da die angesprochene Dopplung verdeutlichen soll, das Anfangs- und Endzustand identisch sind.

8. Studie III – Unterschiede in der Wirksamkeit der drei Modelltypen digital-erweitert, haptisch-interaktiv & illustrativ

8.1 Pilotierungsstudie

Das hier vorgestellte Forschungsprojekt ist sowohl konzeptionell mit starkem Bezug zur experimentell ausgelegten Unterrichtspraxis als auch empirisch grundlagenorientiert ausgerichtet. Das Ziel dieser praxisnahen Vorgehensweise besteht darin, die oft kritisierte Kluft zwischen Theorie und Praxis zu überbrücken (Prediger et al., 2012). Daher besteht das vorrangige Ziel der Pilotierung darin, die Praxistauglichkeit der konzipierten Lernumgebungen zu überprüfen.

Die gleichwertige Einbettung unterschiedlicher Modelltypen in speziell für Vergleichszwecke konzipierte Lernumgebungen gewährleistet zunächst, dass die Lernenden annähernd gleich intensiv mit den Modellen interagieren und ein ähnlich hohes Maß an "Time on Task" aufwenden. Für den Vergleich der Wirksamkeit verschiedener Modelltypen sollte die Qualität der eingesetzten Aufgaben und Materialien ebenfalls gewährleistet sein. Zudem ermöglichen die Erprobung und Evaluation in der Schulpraxis einen weitestgehend reibungslosen Ablauf der Intervention (Flint, 2014).⁶

Der Pilotierung der vorliegenden Studie liegt ein im Zuge des Datenmanagements reduzierter Datensatz mit insgesamt 54 Schüler*innen ($\varnothing = 41,5\%$; Alter: $MW = 13,5y (0,5)$) der Klassenstufe 8 einer Realschule-Plus (Oktober 2020) und eines Gymnasiums (Juni 2021) in Rheinland-Pfalz vor. Die Pilotierung adressierte drei übergeordnete Ziele, welche der Qualitätsprüfung sowie der Identifizierung erforderlicher Adaptionen dienten:

- (1) Die Testkonstruktion einer Skala zum Konstrukt Fachwissen
- (2) Die Evaluation der eingesetzten Lernmaterialien und -umgebungen
- (3) Die Analyse der eingesetzten Messinstrumente

Die Erhebung erfolgte zu fünf Messzeitpunkten, an denen unterschiedliche Methoden zur Überprüfung der verschiedenen Variablen eingesetzt wurden. Die Proband*innen nutzten zu den drei Interventionszeitpunkten alternierend die unterschiedlichen Modelltypen, sodass jede*r Proband*in mit jedem Modelltyp gearbeitet hat. Der Prozess der Entwicklung der Materialien und Messinstrumente ist in Abbildung 15 dargestellt. Der Ablauf der Pilotierung mit den eingesetzten Methoden zum Pre- und Postzeitpunkt sowie zu den drei

⁶Dieser Absatz wurde bereits in Przywarra & Risch (2022) veröffentlicht.

Interventionszeitpunkten ist in Tabelle 17 aufgelistet und umfasst alle Konstrukte, die in der Pilotierungs- und Interventionsstudie erhoben werden.

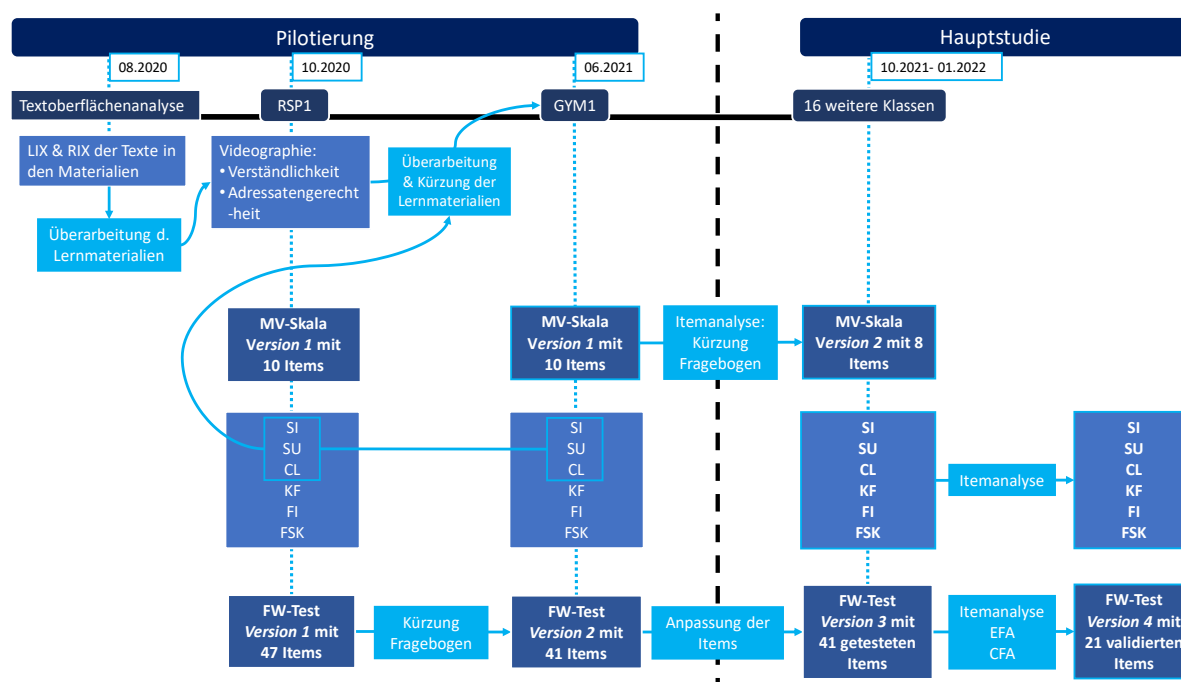


Abbildung 15: Ablauf der Entwicklung der Materialien und Messinstrumente sowie deren Analyse (RSP1 = Klasse 1 an der RealschulePlus, GYM1 = Klasse 1 am Gymnasium, LIX = Lesbarkeitsindex, RIX = Regenburger Index, MV = Modellverständnis, SI = situationales Interesse, SU = System Usability, CL = Cognitive Load, KF = Kognitive Fähigkeiten, FI = Fachinteresse, FSK = Fähigkeitsselbstkonzept, FW = Fachwissen)

Tabelle 17: Ablauf der Datenerhebung während der Pilotierung

MZP 0 Pre	MZP 1 Intervention I	MZP 2 Intervention II	MZP 3 Intervention III	MZP 4 Post
Single-Choice-Test: - Fachwissen - Kognitive Fähigkeiten	Videographie: - Adressatengerechtigkeit - Verständlichkeit	Videographie: - Adressatengerechtigkeit - Verständlichkeit	Videographie: - Adressatengerechtigkeit - Verständlichkeit	Single-Choice-Test: - Fachwissen - Kognitive Fähigkeiten
Likert-Skalen: - Modellverständnis - Fachinteresse - Fachbezogenes Selbstkonzept	Likert-Skalen: - Situationales Interesse - Cognitive Load - System Usability	Likert-Skalen: - Situationales Interesse - Cognitive Load - System Usability	Likert-Skalen: - Situationales Interesse - Cognitive Load - System Usability	Likert-Skalen: - Modellverständnis - Fachinteresse - Fachbezogenes Selbstkonzept
Fachnote und Gesamtschnitt				
Demografische Daten				

8.1.1 Testkonstruktion

Die eingesetzten Skalen sind zu jedem Testzeitpunkt jeweils in Testhefte eingearbeitet. Die Struktur der Testhefte unterteilt sich in ein Deckblatt mit Kontaktdaten, einen Begrüßungstext mit der Abfrage einer anonymen ID, dem Arbeitsauftrag und einem Beispiel zum Umentscheiden bei versehentlichem Ankreuzen (Jonkisz et al., 2012). Durch das Betonen der Anonymität der Umfrage wird neben dem Sicherstellen des Datenschutzes auch reduziert, dass sich die Befragten in einer Prüfungssituation fühlen oder ihre Antworten aufgrund sozialer Erwünschtheit verfälscht werden (Busker, 2014). Im Layout aller Skalen, mit Ausnahme des Fachwissenstests, wurde zur verbesserten Übersichtlichkeit jede zweite Zeile schraffiert und die Indikatoren zur einfacheren Bearbeitung rechtsbündig angeordnet (Jonkisz et al., 2012). Zur Erfassung des Fachwissens wurde eigens ein Single-Choice-Test entwickelt. Die Formulierung erfolgte ohne die Verwendung doppelter Verneinungen, Verallgemeinerungen oder mehrdeutiger Begriffe (Bühner, 2006; Jonkisz et al., 2012). Fachbegriffe wurden nur dann verwendet, wenn ihre Erklärung Teil der Fragestellung ist. Zur Operationalisierung des *Fachwissens* dienten die Basiskonzepte „Teilchen-Materie-Stoff“ und „Struktur-Eigenschaft-Funktion“ im Fach Chemie für die Sekundarstufe I (Ministerium für Bildung, Wissenschaft, Weiterbildung und Kultur, 2014). Inhaltlich ist der Fachwissenstest nach dem geltenden Curriculum und den zuvor definierten Lernzielen zur Lernumgebung ausgerichtet (s. Abschnitt 5.1). Die 47 Single Choice Items (Lienert & Raatz, 1998) wurden orientiert an Fachliteratur (Freyer, 2013) und den Themen der Intervention entwickelt. Neben der korrekten Antwortmöglichkeit wurden drei ansprechende Distraktoren entwickelt (Riese & Reinhold, 2014). Die Konzepte und Formulierungen der Distraktoren entstammen qualitativen Untersuchungen zu Schüler*innenvorstellungen (vgl. Abschnitt 2.3) und stellen bereits empirisch gesicherte Präkonzepte von Schüler*innen dar (vgl. Abbildung 16). Es wurde darauf geachtet, dass die Anzahl der Items sich gleichmäßig auf die beiden Basiskonzepte verteilt (Schmiemann & Lücken, 2014). Um Reihenfolgeeffekte zu vermeiden, befinden sich leichter eingeschätzte Items eher am Anfang bzw. am Ende des Tests (Bühner, 2006). Dies hat zudem den Vorteil, das Aufkommen von Frustration bei den Testpersonen während der Fragebogenbearbeitung zu reduzieren (Theyßen, 2014).

05 iAM1	Welche Aussage zum Teilchenverbund trifft zu?
☐ A	Zwischen den Teilchen eines Körpers ist Vakuum (leerer Raum).
☐ B	Zwischen den Teilchen jeder Flüssigkeit befindet sich die eigentliche Flüssigkeit.
☐ C	Zwischen den Teilchen eines beliebigen Gases ist Luft.
☐ D	Zwischen den Teilchen eines Körpers ist Luft, unabhängig davon, ob er im festen, flüssigen oder gasförmigen Zustand vorliegt.

Abbildung 16: Beispiel-Item des Fachwissenstests, dessen Antwortmöglichkeiten auf existierenden Präkonzepten basieren

8.1.2 Evaluation der Lernmaterialien und -umgebungen

Methoden der Datenerhebung

Die evidenzbasierte Optimierung der Lernmaterialien der drei Unterrichtsstunden zu den drei Modelltypen erfolgte durch Pilotierung anhand unterschiedlicher Methoden. Zunächst wurden die schriftlichen Unterrichtsmaterialien (Aufgabenstellungen, Erklär- und Lückentexte, Versuchs- und Modellanleitungen) mittels Textoberflächenanalyse auf ihre Verständlichkeit hin überprüft (Hansen-Schirra & Gutermuth, 2019). Diese Überprüfung wurde nach der kategoriengeleiteten Überarbeitung der Texte im Anschluss an die qualitative Videostudie erneut vorgenommen.

Zur weiteren Evaluation der Materialien, wurden die Kriterien *Adressatengerechtheit* der Modelle und Experimente und *Verständlichkeit* der Skripte für die Lernenden zusammen mit dem empfundenen, *Situationalen Interesse*, *Cognitive Load* und *Usability* der Lernmaterialien bestimmt. Zunächst erfolgte die Erhebung der Adressatengerechtheit und der Verständlichkeit mit den Schüler*innen der Realschule. Dazu wurden zu jedem der drei Interventionszeitpunkte jeweils zwei Schüler*innen pro Modelltyp videographiert. Zusätzlich wurden an allen drei Messzeitpunkten die Konstrukte *situationales Interesse* (Rheinberg et al., 2001), *System Usability* (Brooke, 1996) und *Cognitive Load* (Leppink et al., 2013) mittels Likert-Skalen zur Selbsteinschätzung in beiden Klassen erhoben. Anhand dieser Konstrukte lässt sich bereits während der Pilotierung prüfen, ob die Aufgaben ein angemessenes Niveau aufweisen und die Anleitungen der Modelle für die Adressaten verständlich formuliert wurden.

Zur Erhebung des *situationalen Interesses* wurde die Interessensskala des Fragebogens zur aktuellen Motivation (FAM) von Rheinberg, Vollmeyer und Burns (2001) verwendet. Die Interessensskala des FAM umfasst fünf Items, die mit einer siebenstufigen Likert-Skala beantwortet werden können. Die Skala reicht von 1=„stimme gar nicht zu“, bis 7=„stimme voll zu“ und weist Reliabilitäten zwischen $\alpha = 0.71 - 0.90$. Das situationale Interesse wurde immer direkt im Anschluss an das Arbeiten mit den Modellen erhoben. Da zu diesem Zeitpunkt, das „Lesen der Instruktion“ bereits länger vergangen ist, wird Item Nummer 7 („Nach dem Lesen der Instruktion erscheint mir die Aufgabe sehr interessant.“) zur besseren Passung durch das entsprechende Item (FAM-V6 „Ich fand diese Aufgabe sehr interessant.“) aus dem Fragebogen zur aktuellen Motivation Video (FAM-Video) von Cauet (2016) ersetzt. Der FAM-Video ist eine adaptierte Version des FAM zur Erfassung der aktuellen Motivation in Unterrichtsstunden. Die Reliabilität des FAM-Videos beläuft sich auf $\alpha = 0.74 - 0.80$ (Cauet, 2016).

Die Skala zur Erfassung der drei Dimensionen des *Cognitive Loads* von Leppink et al. (2013) besteht aus insgesamt zehn Items. Jeweils drei dieser Items bilden die Subskalen ICL und ECL ab, während vier Items die GCL messen. Die Skala wurde zunächst unter Einhaltung der

Leitlinien der ITC (International Test Commission, 2017) ins Deutsche übersetzt und an die Zieldomäne angepasst. Dazu wurden insgesamt Übersetzungen von drei unabhängigen Übersetzer*innen erstellt. Die Übersetzenden weisen Sprachexpertisen aus professionellen sprachlichen bzw. bilingualen Hintergrund auf. Zudem verfügen die Übersetzer*innen über akademische Bildungsabschlüsse und sind mit dem Verfahren psychometrischer Testungen vertraut. Unter Berücksichtigung der jungen Zielgruppe der Befragung wurde die einfache Verständlichkeit der Items im Zuge der Übersetzungen fokussiert. Aufgrund der methodischen Limitationen von Rückübersetzungen (z.B. Behr, 2017) wurde auf diesen Schritt ausdrücklich im Zuge der Förderung eines Konsenses zwischen den Expert*innen verzichtet. Die Expert*innen zum Prüfen und Diskutieren der Übersetzungen im Abgleich mit der Original-Skala stellte eine Gruppe aus fünf fachdidaktisch Forschenden dar. Unter Anwendung der Kriterien nach Hambleton und Zenisky (2010) erarbeitete die Expert*innenrunde die finalen Itemformulierungen aus den drei vorliegenden Übersetzungen. Als Anpassungen an die Zieldomäne wurde die Übersetzung des Worts „activity“ zum deutschen Wort „Lernumgebung“ vorgenommen. Zur Selbsteinschätzung der Proband*innen dient eine elfstufige Likert-Skala, deren Stufen als 0 = „stimmt überhaupt nicht“ bis 10 = „stimmt genau“ übersetzt wurden. Die ICL-Subskala besteht aus drei Items und weist Konsistenzkoeffizienten zwischen $\alpha = 0.86 - 0.91$ auf (Lepink et al., 2013). Die drei Items zum ECL zeigen eine interne Konsistenz zwischen $\alpha = 0.63 - 0.87$ (ebd.). Die interne Konsistenz der GCL-Skala liegt zwischen $\alpha = 0.94 - 0.97$ (ebd.).

Die *System-Usability* Skala nach Brooke (1996) besteht aus insgesamt zehn Items. Für die fünfstufige Likert-Skala wurden die maximalen Ausprägungen mit 1 = „stimme überhaupt nicht zu“ und 5 = „stimme vollständig zu“ übersetzt. Die Übersetzung und Anpassung der Skala erfolgte identisch zum Verfahren, das bei der Situationalen Interessensskala angewendet wurde. Da die vorliegende Studie den Einfluss des Modelltyps als unabhängige Variable untersucht, wurde der Begriff „system“ durch Verwenden des Begriffs „Modell“ konkretisiert. Zur Überprüfung der Usability eines Softwareprodukt konnten Bangor et al. (2008) durch Ersetzen des Begriffs „system“ mit „product“ ähnliche Item- und Skalenkennwerte wie die Originalskala von Brooke (1996) erzielen. So ergab sich für die Skala eine interne Konsistenz von $\alpha = 0.91$ (Bangor et al., 2008). Zudem wurde die einfaktorielle Struktur des Testinstruments von Bangor et al. (2008) bestätigt. Daher werden durch den Austausch des Objekts im Satzbau keine Einschränkungen der Validität oder Reliabilität erwartet. Da die Skala auch bei kleinen Stichproben (12 Versuchspersonen) eine hohe Reliabilität aufweist (Tullis & Stetson, 2004), erlaubt dies bereits zu einem frühen Zeitpunkt der Pilotierung ein formatives Assessment der Usability (Neff, 2023). Das Sicherstellen einer hohen Nutzerfreundlichkeit ermöglicht eine Verzerrung der restlichen Lernendenvariablen zu verhindern (ebd.), da die Usability einen Einfluss auf

ebendiese Variablen ausübt (vgl. Abschnitt 2.4). Usabilitywerte kleiner 50 werden als inakzeptabel eingestuft, während Werte ab 73 als gut gelten (Bangor et al., 2008). Bessere Anwendungen weisen eine Usability im Bereich der hohen 70er bis 85 auf (ebd.). Ab einem Wert von 85 gilt die Usability als exzellent, wobei ein Wert größer 90 für ein absolut überlegenes Produkt spricht (ebd.).

Methoden der Datenauswertung

Im nachfolgenden Kapitel werden sowohl das Datenmanagement der Pilotierungsstudie als auch die Auswertung der quantitativen Fragebogendaten und der Daten aus der Textoberflächenanalyse näher ausgeführt. Darüber hinaus wird die qualitative Auswertung der Video-Daten zur Beurteilung der Adressatengerechtheit und Verständlichkeit der Lernmaterialien der dargestellt.

Textoberflächenanalyse

Im Rahmen der Textoberflächenanalyse wurden der Lesbarkeitsindex (LIX) sowie der Regensburger Index (RIX) und der Simple Measure of Gobbledygook – german (gSMOG) mit dem Regensburger Analysetool für Texte (RATTE Version 1.6) bestimmt. Die Kriterien, die zur Bestimmung der Indizes dienen, sind die Wort- und Satzlänge, die Anzahl der Wörter und Sätze, die relative Häufigkeit von Nebensätzen und Füllwörtern sowie die lexikalische Dichte. Als Metaindex erfasst der RIX zusätzliche Schwierigkeitsparameter wie Satzkomplexität, Passiv und Substantivierungen sowie die Gesamtlänge des Textes (J. Wild & Pissarek, 2016). Die Bestimmung der Indizes RIX und gSMOG ermöglicht es, Auflistungen anhand der Sonderzeichen „:“ und „.“ miteinzubeziehen. Der LIX errechnet sich aus der gemittelten Satzlänge ohne die Berücksichtigung von Aufzählungen und dem Anteil an Wörtern aus mehr als sechs Zeichen (Lenhard & Lenhard, 2011). Anhand der Indizes und Zuordnungen zu Klassenstufen lässt sich der Schwierigkeitsgrad der Texte bestimmen (Bamberger, 2006). Da der RIX die Gesamtlänge des Textes berücksichtigt, die Lesetätigkeit der Schüler*innen jedoch durch das praktische Arbeiten mit den Versuchen und Modellen oder Anfertigen der Zeichnungen unterbrochen wird, wurden die analysierten Texte in Sinnabschnitte unterteilt. Als Sinnabschnitte wurden die Versuchs- bzw. Modellanleitung, Aufgaben bzw. Lückentexte gewählt. Die Analyse des kompletten Materials zu allen Interventionszeitpunkten erfolgte über die Indizes LIX und gSMOG. Zum Vergleich der Lesbarkeit der Materialien zwischen den drei Gruppen (AR, HI & IL) wurden die Modellanleitungen aus den Materialien extrahiert und nach Modelltyp getrennt analysiert. Daraus ergibt sich ein Stichprobenumfang von 18 Beobachtungen (jeweils sechs Modelle pro Gruppe) anhand derer Unterschiedsprüfungen durchgeführt wurden. Da sich die restlichen Materialien nicht unterscheiden, wurde die Lesbarkeit der Aufgaben, Versuchsanleitungen und Lückentexte gesammelt überprüft.

Qualitative Inhaltsanalyse zur Bestimmung der Adressatengerechtigkeit und Verständlichkeit

Zur Überprüfung der Adressatengerechtigkeit und Verständlichkeit der Lernmaterialien wurden die videographierten Daten anschließend in einer inhaltlich strukturierenden qualitativen Inhaltsanalyse nach Mayring (2010) ausgewertet. Die Kodierung der Videos der Lerneinheiten erfolgte eventbasiert. Der Vorteil gegenüber der zeitbasierten Auswertung liegt darin, dass eventbasiert alle aufgetretenen Probleme erfasst werden können. Die betrachteten Kategorien stellen keine Oberflächenmerkmale dar. Da bereits während des Kodierens eine Interpretation erforderlich wird, muss die inhaltliche und zeitliche Begrenzung des Events klar zu erfassen sein (Seidel, 2003). Zu diesem Ziel wurde die qualitative Analyse der Videos optimiert. Zunächst wurde die Kodierung selbst direkt in den Textdateien der Lehrmaterialien vorgenommen. Weiterhin wurde als Kodiereinheit der vollständige Satz gewählt, zu dem in den Materialien das Problem aufgetreten ist. Um die Inferenz der Auswertung weiterhin niedrig zu halten, wurden die Kategorien so gewählt, dass sie auf direkt zugänglichen, visuellen und auditiven Strukturen beruhen (Brückmann & Duit, 2014). Zur Bildung des Kategoriensystems wurden die Kategorien deduktiv aus Untersuchungen mit ähnlich konzipierten Lernumgebungen generiert (Mayring, 2010). Zudem erfolgte eine induktive Rückkopplung, um vorhandene Kategorien anzupassen bzw. fehlende Kategorien herauszuarbeiten (ebd.). In der vorliegenden Studie wurde das bereits für Schüler*innenexperimentierphasen erprobte Kategoriensystem nach Engl (2020) genutzt (s. Tabelle 18). Dabei wurde der Subcode „ungelöst“ immer dann verwendet, wenn die Testpersonen um Hilfestellung gebeten haben. Zusätzlich wurde die Lerneinheit auf eine Passung der Lerngruppe zur verfügbaren Arbeitszeit hin überprüft.

Tabelle 18: Kategoriensystem zur qualitativen Auswertung von Adressatengerechtigkeit und Verständlichkeit der Materialien auf Basis der Videodateien

Kategorie	Code	Subcode
Aufgabe	Aufgabe übersehen	Gelöst
	Aufgabe Schwierigkeitsgrad	Ungelöst
	Verständlichkeit der Aufgabenstellung	
Modell Handhabung	Beschriftung der Materialien	Gelöst
	Versuchsaufbau	Ungelöst
Modell Anleitung	Unpräzise Formulierung	Gelöst
	Falsche Reihenfolge	Ungelöst
	Unbekannte Wörter	
Versuch Handhabung	Beschriftung der Materialien	Gelöst
	Versuchsaufbau	Ungelöst
Versuch Anleitung	Unpräzise Formulierung	Gelöst
	Falsche Reihenfolge	Ungelöst
	Unbekannte Wörter	

Datenmanagement der Fragebogendaten

Zunächst wurden die Daten zu den eingesetzten Messinstrumenten gemäß den vorliegenden Manualen konvertiert. Die Skala zum *situationalen Interesse* des FAMs bedarf gemäß Manual keiner spezifischen Anpassung der Items (Rheinberg et al., 2001). Die Auswertung der Skala zum *Cognitive Load* erfolgt nach dem Manual von Leppink et al. (2013) auf Basis der empirisch gesicherten Dreifaktorenstruktur der Skala getrennt nach *ICL*, *ECL* und *GCL*. Dazu wurden die Items in entsprechende Subsets aufgeteilt. Die entsprechenden Items der *System Usability Scale* wurden gemäß Leitfaden nach Brooke (1996) umgepolt. Zur Itemanalyse wurde die Skalierung der Items von 0 bis 4 verwendet. Für die normierte Gesamtbeurteilung der Usability wurde die Summe der Ausprägungen aller Items der Skala mit 2,5 multipliziert, um einen Wertebereich von 0-100 pro Personenrating zu erhalten (Brooke, 1996).

Umgang mit fehlenden Werten

Die Pilotierung diente nicht der Identifizierung inter- oder intraindividuelle Unterschiede, sondern der Evaluation der Lernmaterialien. Zudem wurden die gemessenen Konstrukte unabhängig voneinander deskriptiv betrachtet. Die Analysen erfolgten dabei auf Ebene der Gesamtstichprobe. Das Aufkommen fehlender Werte wurde über alle Messzeitpunkte hin breit und zufällig gestreut erwartet. Da das Auftreten eines fehlenden Wertes unabhängig von der Ausprägung der Variable selbst und den restlichen Variablen im Datensatz ist, gilt die Missing Completely At Random (MCAR) Annahme (Wirtz, 2004). Daher wurden Teilstichproben für jedes (Sub-)Konstrukt zu allen Messzeitpunkten gesammelt gebildet. Die Stichproben von $n < 100$ schließt dies die Anwendung des Expectation-Maximization Algorithmus oder multipler Imputationen aus (Wirtz, 2004). Daher wurde sich zum Umgang mit fehlenden Werten für eine einfache Imputation als Ersetzung durch die Skalenmittelwerte der betroffenen Personen zum jeweiligen Konstrukt entschieden.

Personenbezogene Analyse

Weiterhin wurde eine personenbezogene Analyse durchgeführt. Zunächst wurde der Datensatz auf das Vorhandensein von Extremwerten geprüft. Weichen Werte um mehr als den dreifachen Interquartilsabstand nach unten vom ersten Quartil bzw. nach oben vom dritten Quartil ab, werden diese als Extremwerte definiert (Eid et al., 2015). Sofern Extremwerte im Datensatz vorlagen, wurden diese nach Ausschluss von Eingabefehlern fallweise gelöscht. Zur Überprüfung der Plausibilität der Antworten wurde diese auf Kreuzungsmuster untersucht (Döring & Bortz, 2016). Dazu wurde die personenbezogene Standardabweichung innerhalb der jeweiligen Skala bestimmt (ebd.). Als Cut-Off-Wert wurde eine Standardabweichung kleiner als 0.2 definiert (ebd.).

Deskriptive und inferenzstatistische Datenauswertung

Zur Auswertung der deskriptiven Statistiken, Itemkennwerte und der Unterschiedsprüfungen wurde die Software *R-Studio* (Version 2021.09.1) mit R (Version 3.14.0) unter Verwendung der Pakete *utils*, *likert*, *psych*, *dplyr*, *stats*, *WRS2*, *car*, *graphics*, *MBESS*, *agricolae* und *effectsize*. Aufgrund der Stichprobengröße von $n = 27$ erfolgte die Überprüfung auf Normalverteilung zur Textoberflächenanalyse mittels Shapiro-Wilk-Tests, da dieser in kleinen Stichproben wenig anfällig gegenüber Ausreißern ist. Weiterhin wurde die Normalverteilung der Residuen der Faktoren mittels QQ-Plots (Abbildung 17) begründet (Eckstein, 2012). Die Überprüfung auf Varianzhomogenität wurde mittels Levene-Tests durchgeführt. Zur inferenzstatistischen Auswertung der Indizes LIX und RIX wurde eine einfaktorielle ANOVA (Döring & Bortz, 2016) und des gSMOG eine Oneway-Varianzanalyse mit Welch-Korrektur (Welch, 1951) durchgeführt. Zur Post-Hoc-Analyse wurde der TukeyHSD-Test verwendet.

Ergebnisse

Im Anschluss an die Überarbeitung aller Materialien wurde die Verständlichkeit dieser für Lernende der achten Jahrgangsstufe mittels Textoberflächenanalyse geprüft. Die Ergebnisse für das Gesamtmaterial belaufen sich für AR auf $LIX = 41.01$ und $gSMOG = 5.9$, für HI auf $LIX = 40.28$ und $gSMOG = 5.9$ sowie für IL auf $LIX = 41.58$ und $gSMOG = 6.0$. Die deskriptiven Kennwerte aller Aufgaben, Versuchsbeschreibungen und Lückentexte sind in Tabelle 19 aufgelistet. Die Ergebnisse der Vergleichsgruppen sind in Tabelle 20 dargestellt. Die ermittelten Maximalwerte für LIX und gSMOG überschreiten die empfohlenen Grenzwerte. Als Ursache hierfür lässt sich die Verwendung der Leitfrage der Stunde als Nebensatz ausmachen: *Nachdem du das Modell benutzt hast, wollen wir nun die Frage „Was passiert mit den Wasserteilchen und den Salzteilchen, wenn die Salzlösung eingedampft wird?“ beantworten [Tag 2, S. 21].* Aus dem Mittelwert aller Beobachtungen ergibt sich für den gSMOG jedoch eine Eignung für die Klassenstufe sechs. Mit einem Mittelwert des Lesbarkeitsindex von 41.37 liegt dieser knapp über dem Wert für Kinder- und Jugendliteratur (<40 ; Lenhard & Lenhard, 2011). Alle Textabschnitte lassen sich vom Komplexitätsniveau unter Belletristik (40-50) einordnen (ebd.). Die Werte für den Verständlichkeitsindikator RIX liegen im Bereich der dritten bis siebten Jahrgangsstufe. Aus dem Mittelwert des RIX ergibt sich eine Verständlichkeit ab Klassenstufe fünf.

Tabelle 19: Deskriptive Statistiken für die Aufgaben, Versuchsanleitungen und Lückentexte. Für die Indizes Lesbarkeitsindex (LIX), Regensburger Index (RIX) und Simple Measure of Gobbledygook – german (gSMOG) sind die deskriptiven Kennwerte MW = Mittelwert, Median, SD = Standardabweichung, Schiefe, Kurtosis und Anzahl der Beobachtungen n angegeben. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert.

Index	MW	Median	SD	Min	Max	Schiefe (<2) Kurtosis (<7)	n
gSMOG (≤ 8.00)	5.95	5.88	1.37	3.86	9.40	0.72 0.04	27
RIX (≤ 8.00)	4.55	4.71	0.72	2.41	5.69	-0.79 0.60	27
LIX (≤ 40.00)	41.37	39.69	4.56	34.25	50.12	0.57 -0.90	27

Die Textoberflächenanalyse der Modellbeschreibungen ergibt für den gSMOG eine Eignung ab den Jahrgangsstufen fünf beziehungsweise sechs. Aufgrund der vorliegenden Varianzheterogenität ($F = 3.76$, $p = 0.04$) und Normalverteilung ($W = 0.92$, $p = 0.10$) wurde eine Varianzanalyse mit Welch-Korrektur zu Prüfung der Unterschiede zwischen den Gruppen durchgeführt (Welch, 1951). Es konnte ein großer Effekt ($\eta^2 = 0.25$, $p = 0.26$) festgestellt werden (Cohen et al., 2013), der jedoch keine Signifikanz aufweist. Deskriptiv ergibt der RIX eine Eignung ab den Klassenstufen vier bei IL und fünf bei AR und HI. Aufgrund von

Varianzhomogenität ($F = 0.23$, $p = 0.80$) und Normalverteilung ($W = 0.96$, $p = 0.64$) wurde eine einfaktorielle ANOVA zur Unterschiedsprüfung durchgeführt. Auch hier konnte kein signifikanter Unterschied ($\eta^2 = 0.21$, $p = 0.17$) festgestellt werden. Die deskriptiven Ergebnisse des LIX lassen eine Zuordnung der Texte von HI zu Kinder- und Jugendliteratur und der Anleitungen von AR und IL zu Belletristik zu. Aufgrund der Struktur der Daten zum LIX (Normalverteilung: $W = 0.96$, $p = 0.64$; Varianzhomogenität: $F = 0.23$, $p = 0.80$) fällt die Wahl der Methode zur Unterschiedsprüfung auf eine einfaktorielle ANOVA. Die ANOVA zum Gruppenvergleich zeigt einen signifikanten, großen Effekt ($\eta^2 = 0.36$, $p = 0.03$). Der TukeyHSD-Test zeigt, dass dieser Effekt nur für den Unterschied zwischen IL und HI signifikant wurde (HI-AR: $p = 0.07$; IL-AR: $p = 0.95$; IL-HI: $p = 0.04$).

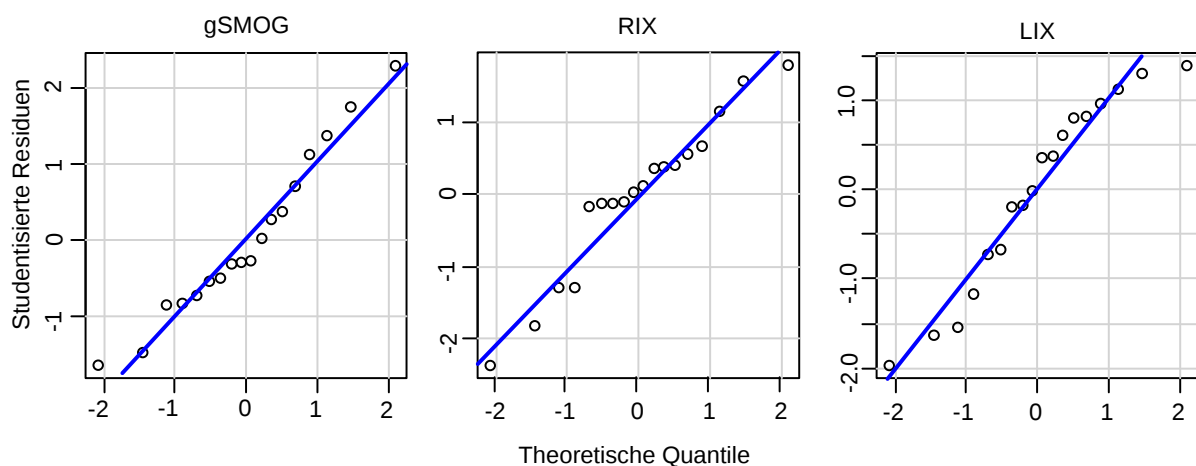


Abbildung 17: QQ-Plots der Verteilung der studentisierten Residuen auf die theoretischen Quantile der Indizes gSMOG, RIX und LIX: Die Darstellungen weisen keine gravierenden Abweichungen von der Normalverteilung auf.

Tabelle 20: Für die Indizes Lesbarkeitsindex (LIX) sowie der Regensburger Index (RIX) und der Simple Measure of Gobbledygook – german (gSMOG) sind die deskriptiven Kennwerte MW = Mittelwert, SD = Standardabweichung, die Testvoraussetzungen Normalverteilung und Varianzhomogenität sowie die Ergebnisse der Unterschiedsprüfungen mit Effektstärken. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert.

Index	MW AR (SD)	MW HI (SD)	MW IL (SD)	Testvoraussetzungen	Unterschiedsprüfungen und Effektstärken
gSMOG (≤ 8.00)	5.62 (0.70)	5.24 (0.34)	5.62 (0.43)	$W = 0.91$, $p = 0.10$ $F = 3.76$, $p = 0.048$	$F = 1.58$, $p = 0.26$ $\eta^2 = 0.25$
RIX (≤ 8.00)	4.97 (0.77)	4.83 (0.93)	4.03 (0.95)	$W = 0.96$, $p = 0.64$ $F = 0.23$, $p = 0.80$	$F = 4.30$, $p = 0.17$ $\eta^2 = 0.21$
LIX (≤ 40.00)	41.65 (2.02)	38.31 (2.97)	42.05 (2.15)	$W = 0.92$, $p = 0.15$ $F = 1.15$, $p = 0.34$	$F = 4.30$, $p = 0.03$ $\eta^2 = 0.36$

Während der qualitativen Datenauswertung der Videos wurden insgesamt 37 Probleme bei allen 18 videographierten Versuchspersonen identifiziert. Die codierten Kategorien sind in Tabelle 21 dargestellt. Dabei wurde keine Aufgabe als zu schwer kodiert. Zudem wurden keine Probleme in der Handhabung des Versuchsaufbaus der Modelle identifiziert.

Tabelle 21: Liste der Codes zu den identifizierten Schwierigkeiten während der Durchführung der Lerneinheiten; die Anzahl der kodierten Segmente steht in den Klammern.

Aufgaben	Versuche	Modelle
- Aufgabe übersehen	- Versuch Handhabung Beschriftung Materialien ungelöst	- Modell Handhabung Beschriftung Materialien ungelöst
- Aufgabe Verständlichkeit der Aufgabenstellung	- Versuch Handhabung Beschriftung Materialien gelöst	- Modell Handhabung Beschriftung Materialien gelöst
	- Versuch Anleitung unvollständig/unpräzise ungelöst	- Modell Anleitung unpräzise gelöst
	- Versuch Anleitung unvollständig/unpräzise gelöst	
	- Versuch Handhabung Versuchsaufbau ungelöst	

Die Verteilung der arithmetischen Mittelwerte der Skalen *Situationales Interesse*, *Intrinsic Cognitive Load*, *Extraneous Cognitive Load*, *Germane Cognitive Load* und *System Usability* ist in Abbildung 18 dargestellt. Die Schüler*innen zeigen beim Bearbeiten der Lernumgebung ein mittleres bis hohes situationales Interesse (Rheinberg et al., 2001). Die kognitive Belastung fällt im niedrigen bis mittleren Bereich ICL ($MW = 3.64$, $SD = 3.07$) aus (Leppink et al., 2013). Der ECL ($MW = 2.33$, $SD = 2.48$) wurde von den Lernenden als niedrig bewertet (ebd.). Schließlich bewerteten die Schüler*innen den GCL ($MW = 5.52$, $SD = 2.7$) als mittel bis hoch (ebd.). Die System Usability der Modelle ($MW = 83.1$, $SD = 14.61$) wurde von den Lernenden im besseren Bereich eingestuft (Bangor et al., 2008). Die erzielten Itemkennwerte interne Konsistenz (Cronbachs α), korrigierte Itemtrennschärfe (r_{it}), die Itemschwierigkeit nach Dahl (P_{Dahl}), Schiefe und Kurtosis sind in Tabelle 22 aufgelistet. Abschließend ist festzuhalten, dass Schüler*innen im festgelegten Zeitraum den letzten Versuch „Osmose“ nicht vollständig bearbeiten konnten.

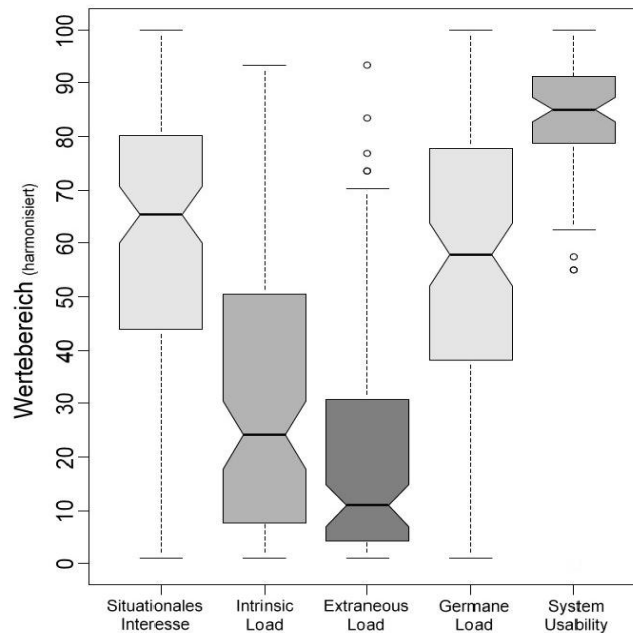


Abbildung 18: Personenbezogene Mittelwerte der Skalen Situationales Interesse, Intrinsic Cognitive Load, Extraneous Cognitive Load, Germane Cognitive Load und System Usability

Tabelle 22: Für die Konstrukte Situationales Interesse, Cognitive Load und System Usability sind die Itemkennwerte MW = Mittelwert, Median SD = Standardabweichung, Schiefe, Kurtosis, d = Schwierigkeitsindex nach Dahl, r_{it} = Korrigierte Trennschärfe und die Reliabilität nach Cronbach α angegeben; n entspricht der Anzahl an messwiederholten Datenpunkten, die Zahl in der Klammer zu n entspricht der tatsächlichen Anzahl an Personen. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert.

Konstrukte	r_{it} ($>.3$)	P_{Dahl} (20-80)	α ($>.5$) ω ($>.5$)	MW	Me- dian	SD	Schiefe (<2) Kurtosis (<7)	n
Situationales Interesse (5 Items, Skalierung 1-7)	.71 – .86	48.23 – 73.37	.88 .88	4.66	4.9	1.48	-0.86 – 0.37 -1.20 – -0.07	117 (39)
Cognitive Load (10 Items, Skalierung 0-10)	.39 – .82	20.54 – 58.04	.86 .84	3.96	3.6	2.01	-0.30 – 1.24 -1.30 – -0.28	147 (49)
Intrinsic Cognitive Load (3 Items, Skalierung 0-10)	.92 – .95	31.27 – 36.55	.96 .96	3.64	3.33	3.03	0.30 – 0.51 -1.30 – -0.89	147 (49)
Extraneous Cognitive Load (3 Items, Skalierung 0-10)	.73 – .82	20.54 – 26.51	.83 .84	2.34	1.5	2.48	-0.30 – 1.24 -1.30 – -0.28	147 (49)
Germane Cognitive Load (4 Items, Skalierung 0-10)	.79 – .90	51.89 – 58.04	.86 .93	5.52	5.5	2.7	-0.30 – -0.12 -1.18 – -0.88	147 (49)
System Usability (10 Items, Skalierung 0-4)	.32 – .79	50.13 – 74.67	.72 .71	3.36	3.4	0.41	-3.10 – -0.19 -0.80 – 11.72	117 (39)

Diskussion zur Eignung der Lernmaterialien

Die hier aufgeführten Diskussionspunkte thematisieren die konkrete inhaltliche Diskussion und die Methodenkritik explizit zur Evaluation der Lernmaterialien. Eine übergeordnete Diskussion, die gemeinsame Diskussionspunkte zur gesamten Studie III aufgreift, ist in Abschnitt 8.3.5 dargestellt. Die Ergebnisse aus der Auswertung der Textverständnisindizes zeigen überwiegend zufriedenstellende Ergebnisse. Die Höchstwerte erzielten Textstellen, die die Fragestellung der Stunde beinhalten. Da es sich bei diesen Stellen stets um Wiederholungen des Stundeneinstiegs handelt, ist der Halbsatz mit der Fragestellung den Lernenden bereits bekannt. Daher wird die Einschränkung der Verständlichkeit für die Zielgruppe als vernachlässigbar eingeschätzt. Da die Modellanleitungen der Gruppen AR und IL jedes Mal die Nutzung der Apps bzw. Legebilder thematisieren, wiederholt sich der Großteil dieser Anleitungen in identischer Form. Folglich relativieren sich die Unterschiede im LIX zwischen den Gruppen in Bezug auf ihre Auswirkung auf die kognitive Belastung. Zudem ist die Bewertung der Texte anhand des LIX kritisch zu betrachten. Die Durchschnittslänge der Wörter eines Satzes beeinflusst den LIX stark (J. Wild & Pissarek, 2016). Die Materialien bestehen zum Großteil aus Auflistungen, die keine kurzen Wörter wie Propositionen oder Konjunktionen enthalten. Daher führt die Betrachtung des LIX für solche Materialien zu Verzerrungen. Da der LIX auch die Trennung von Sinnabschnitten durch „;“ oder „:“ nicht berücksichtigt, ist dieser Index als nachteilig zur Bewertung der Lesbarkeit Versuchs- und Modellanleitungen einzuschätzen. Die Varianzanalysen zu gSMOG und RIX zeigen keine signifikanten Unterschiede zwischen den drei Gruppen. Die Passung der Indizes gSMOG und RIX zur Zielgruppe lässt darauf schließen, dass die Texte aller drei Vergleichsgruppen für die Lernenden als verständlich anzusehen sind.

Das Ausbleiben von Kodierungen zu Problemen im Umgang mit den Modellen spricht ebenso für eine sehr gute Passung der Modelle zur Zielgruppe. Es wurden lediglich Probleme auf Ebene der Materialien und Experimenten identifiziert. Alle identifizierten Probleme konnten zum Beispiel durch zusätzliche Beschriftungen der Materialien oder das Verwenden alternativer Materialien (Federstahlpinzetten statt anatomischer Pinzetten) behoben werden. Weiterhin lässt die sehr hohe Einschätzung der Usability der Modelle durch die Versuchspersonen auf eine angemessene Adressatengerechtigkeit schließen. Die Werte für Schiefe und Kurtosis können auf Deckeneffekte hinweisen (Eid et al., 2015). Aufgrund der Schwierigkeitsverteilungen der Items ist jedoch anzunehmen, dass die Skala im oberen Bereich noch ausreichend differenziert auflöst. Die Überschreitung der Lernzeit von 90 Minuten führte zum Ausschluss des Versuchs Osmose aus der Intervention. Als Ausgleich wurde der Versuch zum Teilchensieben in die Versuche „Herstellung einer Goji-Lösung“ und das eigentliche „Teilchensieben“ aufgeteilt. Entsprechend wurden die vier Items zum Phänomen Osmose aus dem Fachwissenstest entfernt.

Aus den Werten für den ICL lässt sich für die Lernaufgaben der Intervention ein niedriges bis mittleres Anspruchsniveau ableiten. Die niedrig ausgefallene Bewertung des ECL lässt die Schwierigkeit, die die Lernenden dem Material und den Modellen zuordnen, ebenfalls als niedrig einstufen. Die mittlere bis hohe Ausprägung des GCL lässt darauf schließen, dass den Schüler*innen ausreichend kognitive Kapazität zum Lernen zur Verfügung stand. Die Usability im besseren Bereich gilt als ein sehr guter Wert für Lernumgebungen (Karapanos et al., 2018). Zusammenfassend lässt sich konstatieren, dass die entwickelten Lernumgebungen und Modelle unter Berücksichtigung der gemessenen Konstrukte als geeignet für die Zielgruppe bezeichnet werden können.

8.2 Analyse der eingesetzten Messinstrumente

Aufgrund der besonderen Situation während der COVID-19-Pandemie mussten für die Analyse der eingesetzten Messinstrumente sowohl die Stichprobe der Pilotstudie als auch die Stichprobe der Interventionsstudie zugrunde gelegt werden (vgl. Abbildung 15).

Methoden der Datenerhebung

Neben den, zur Evaluation der Lernmaterialien eingesetzten, Skalen zu *situationalem Interesse*, *Cognitive Load* und *System Usability* (vgl. 8.1.2) kamen die im Folgenden beschriebenen Messinstrumente in der Pilotierung und Interventionsstudie zum Einsatz.

Zur Messung des Konstrukts *Modellverständnis* wird die Skala von Leisner (2005a) für Schüler*innen der Klassenstufe sieben verwendet. Die Skala umfasst 10 Items zur „Natur“ des Teilchenmodells basierend auf dem Test von Mikelskis-Seifert (2002). Die Befragten ordnen ihre Vorstellungen zur Existenz von Teilchen und dem Teilchenmodell als Abbild der Realität oder als Modellvorstellung auf einer quasi-fünfstufigen Likert-Skala ein. Die Stufen der Skala reichen von 0 = „lehne ab“ bis 4 = „stimme zu“, wobei Stufe 2 = „habe keine Meinung“ als zusätzliches Feld abgesetzt erscheint. In der Untersuchung von Leisner (2005a) weist die Skala eine interne Konsistenz von $\alpha = 0.77$ auf.

Die Messung des *chemiebezogenen Fähigkeitsselbstkonzepts* erfolgte mittels der validierten Skala nach Engl (2020). Diese setzt sich aus sieben Items mit einer bipolaren und vierstufigen Likert-Skala zusammen. Die Stufen der Skala reichen von 1 = „stimmt gar nicht“ bis 4 = „stimmt völlig“. Fünf Items dieser Skala entstammen ursprünglich der Skala von Köller et al. (2000) und zwei Items dem Test von Pawek (2009). Die Items dienen zur Selbsteinschätzung der Befragten über ihre theoretischen und praktischen Fähigkeiten im Chemieunterricht, auch im Vergleich zu Mitschüler*innen. Da die Skala von Engl (2020) auf den Ergebnissen einer Itemanalyse in einem vergleichbaren Untersuchungssetting zur vorliegenden Studie basiert, wurde sich zugunsten der Testökonomie für diese Vorauswahl entschieden. In der Untersuchung von Engl (2020) weist die Skala eine interne Konsistenz von $\alpha = 0.89$ auf.

Aus den gleichen Gründen entstammt die Skala zur Messung des *chemiebezogenen Fachinteresses* entstammt ebenfalls der Untersuchung von Engl (2020). Die Skala setzt sich aus fünf Items mit einer bipolaren und vierstufigen Likert-Skala zusammen. Die Stufen der Skala gehen von 1 = „stimmt gar nicht“ bis 4 = „stimmt völlig“. Die Items der Skala entstammen ursprünglich der Skalen von Hoffmann et al. (1998), Köller et al. (2000), Pawek (2009) und Weißnigk (2012). Die Items dienen der Selbsteinschätzung zur Einstellung zum und dem Interesse am Chemieunterricht. Die nach Engl (2020) gekürzte Skala erreicht eine Reliabilität von $\alpha = 0.89$.

Zur Messung der *kognitiven Fähigkeiten* wurde der mini-q-Test nach Baudson und Preckel (2016) zur Einschätzung von speeded reasoning eingesetzt. Innerhalb einer begrenzten Zeitspanne von drei Minuten sind die Proband*innen dazu aufgefordert möglichst viele der insgesamt 64 Single-Choice-Items zu beantworten. Die Items bestehen aus einer Symbolfolge und einer zugeordneten Aussage, deren Korrektheit in Bezug auf die Symbolfolge beurteilt werden soll (Baudson & Preckel, 2016). Dabei kann zwischen den beiden Antwortoptionen „richtig“ und „falsch“ gewählt werden. Das Messinstrument erreichte dabei in mehreren Testungen von Baudson und Preckel (2016) eine Reliabilität von $\alpha = 0.98$.

Methoden der Datenauswertung

Die Auswertung der deskriptiven Statistiken und der Itemanalyse erfolgte mittels der Software *R-Studio* (Version 2021.09.1) mit *R* (Version 3.14.0) unter Verwendung der Pakete *utils*, *likert*, *psych*, *dplyr*, *stats*, *WRS2*, *car*, *graphics*, *MBESS* und *effectsize*.

Datenmanagement

Die Konvertierung der Daten zu den Konstrukten *situationales Interesse*, *Cognitive Load*, *System* und *System Usability* erfolgte nach dem gleichen Schema wie zuvor unter 8.1.2 beschrieben. Die Daten aller eingesetzten Messinstrumenten wurden gemäß den vorliegenden Manu-
alen konvertiert. Die Transformationen der Werte der Skalen *chemiebezogenes Fachinteresse* und *Fähigkeitsselbstkonzept* erfolgten dabei nach Engl (2020). Die Anpassung der Daten zur *Modellverständnis*-Skala wurde dabei nach Leisner (2005a) durchgeführt. Zur Aufbereitung der *Fachwissensitems* wurden korrekten Antworten ein Wert von 1 sowie falschen oder fehlenden Antworten ein Wert von 0 zugeschrieben.

Umgang mit fehlenden Werten

Der Umgang mit fehlenden Werten und dem Ausschluss von Personen aus der Stichprobe während der Pilotierung ist bereits unter 8.1.2 beschrieben. Im Folgenden wird das Schema zur Handhabung mit den Versuchspersonen der Interventionsstudie skizziert.

Auch für die Analyse der Messinstrumente der Interventionsstudie gilt die *Missing Completely At Random* (MCAR) Annahme (Wirtz, 2004). Da keine multivariate Normalverteilung vorliegt und die Gruppen (AR, HI, IL) kleiner als $n = 100$ sind, schließt dies die Anwendung des Expectation-Maximization Algorithmus oder multipler Imputationen aus (Wirtz, 2004). Daher wurde sich zu Umgang mit fehlenden Werten für einen listenweisen Fallausschluss über alle Messzeitpunkte und erfassten Konstrukte hinweg entschieden. Dies betraf Proband*innen mit einem Fehltermin zu einem der fünf Messzeitpunkte. Sofern nur Werte einzelner Items zu einem Konstrukt und Messzeitpunkt fehlten, wurden diese Werte durch einfache Imputation in Form von einfacher Ersetzung durch die Skalenmittelwerte der betroffenen Personen zum

jeweiligen Konstrukt ergänzt. Als Cut-Off-Kriterium für Versuchspersonen wurde das Vorliegen einer eindeutigen Mehrheit an vorhandenen Werten definiert. Eine Ausnahme stellt das Konstrukt Fachwissen dar. Hier wurden fehlende Werte mit null Punkten codiert. Dieses Vorgehen entspricht nach Little und Rubin (2019) einer available-case-analysis. Demnach basieren die Kennwerte der Items und Unterschiedsprüfungen auf einer vollständig reduzierten Stichprobe. Folglich werden sowohl zur Itemanalyse als auch deskriptiven und inferenzstatistischen Betrachtung der zugehörigen Konstrukte identische Stichproben betrachtet.

Personenbezogene Analyse

Die personenbezogene Analyse wurde identisch zum Vorgehen in der Pilotierung durchgeführt. Zur Überprüfung der Plausibilität der Antworten wurde diese auf Kreuzungsmuster untersucht (Döring & Bortz, 2016). Dazu wurde die personenbezogene Standardabweichung innerhalb der jeweiligen Skala bestimmt (ebd.). Als Cut-Off-Wert wurde eine Standardabweichung kleiner als 0.2 definiert (ebd.). Sofern Extremwerte im Datensatz vorlagen, wurden diese nach Ausschluss von Eingabefehlern listenweise gelöscht.

Analyse der Itemkennwerte und Verteilungen

Zur explorativen Analyse der Verteilungen wurden die deskriptiven Kriterien Mittelwert, Standardabweichung, Minimum, Maximum sowie Schiefe und Kurtosis betrachtet. Für den Betrag der Schiefe gilt ein Überschreiten des Grenzwerts von 2 und für den Betrag der Kurtosis von 7 als gravierende Abweichung der Normalverteilung (Eid et al., 2015). Für die Überarbeitung der Fachwissens-Items wurde die Verteilung der Antworten auf die vier Optionen ausgewertet. Um die Testgütekriterien sicherzustellen, wurden für alle Skalen die Kennwerte Reliabilität, Trennschärfe und Schwierigkeit ermittelt. Zur Bestimmung der Reliabilität und Skalenhomogenität kann die Berechnung der internen Konsistenz der Skala mit dem Cronbach's- α -Koeffizienten dienen (Cronbach, 1951). Als Voraussetzung dafür muss ein essentiell τ -äquivalentes und unidimensionales Messmodell vorliegen (Eid et al., 2015). Wird die erste Voraussetzung verletzt, gilt Cronbach's α im Vergleich zur allgemeineren Form der internen Konsistenz, McDonald's ω , als weniger exakt (Hayes & Coutts, 2020). Laut Schermelleh-Engel und Werner (2012) kann α jedoch noch als untere Grenze der Zuverlässigkeit angenommen werden. Da für die vorliegenden Skalen weder die Ladungen der Faktoren (manifeste Variablen in Form der Items) gleich sind, noch die Items die gleiche Itemschwierigkeit aufweisen, wird ein τ -kongenerisches Messmodell angenommen und McDonald's ω bestimmt (Döring & Bortz, 2016; Trizano-Hermosilla & Alvarado, 2016). Weiterhin betonen Schermelleh-Engel und Werner (2012), dass Persönlichkeitstests häufig nur Reliabilitäten um $\alpha = 0.7$ erreichen. Lienert und Raatz (1998) definieren daher eine interne Konsistenz von 0.5 bis 0.7 für Gruppenvergleiche als ausreichend. Im Fall der Multidimensionalität der *Cognitive Load* Skala mit ihren drei expliziten Subskalen wurde die Reliabilität für jede Subskala einzeln bestimmt (Döring & Bortz,

2016). Aufgrund der bereits vorliegenden ausführlichen Validierung der Skalen zu Modellverständnis, Fähigkeitsselbstkonzept, Fachinteresse, kognitive Fähigkeiten, System Usability, situationalem Interesse und Cognitive Load wurde auf eine Dimensionalitätsprüfung dieser Skalen verzichtet (ebd.). Zudem liefern die guten Reliabilitäten der Skalen und Subskalen weitere Indizien für die Korrektheit der angegebenen Faktorenstrukturen (ebd.).

Zur Schätzung der Itemschwierigkeit der Likert-skalierten Messinstrumente wurde der Schwierigkeitsindex nach Dahl bestimmt. Dieser Wert kann als arithmetisches Mittel der Teilnehmendenantworten interpretiert und somit für intervallskalierte Daten herangezogen werden. Je größer ein Wert im Wertebereich von 0-100 ausgeprägt ist, desto leichter stimmten die Befragten dem Item zu. Sofern die Skalen nicht mit dem Wert „0“ starten, wurden unter Einbezug der maximalen und minimalen Skalenwerte die modifizierte Formel nach Döring und Bortz (2016, S. 477) als Korrektur angewendet. Die Bestimmung der Itemschwierigkeit der Items aus dem Fachwissenstest erfolgt über die relative Häufigkeit an richtigen Antworten aller Proband*innen im Verhältnis zur maximalen Punktzahl (Bühner, 2006; Kelava & Moosbrugger, 2012). Für den Fachwissenstest wurde keine Zufallskorrektur vorgenommen. Da die Versuchspersonen angewiesen wurden stets die Antwortoption zu wählen, die ihnen am wahrscheinlichsten erscheint und die Distraktoren empirisch fundierte Präkonzepte abbilden, würde eine Zufallskorrektur die Ratewahrscheinlichkeit überschätzen (Bühner, 2006). Da jeder Versuchsperson ausreichend Zeit zur Bearbeitung aller Aufgaben zur Verfügung stand, wurde eine Inangriffnahmekorrektur ebenfalls abgelehnt (Bühner, 2006). Die Untersuchung zielt nicht darauf ab, Unterschiede zwischen den extremen Ausprägungen der Personenmerkmale herauszustellen. Daher wurden die Grenzwerte für die Itemschwierigkeit nach Kelava und Moosbrugger (2012) und Döring und Bortz (2016) als $20 < P_{\text{Dahl}} < 80$ festgelegt. Zur Bestimmung der Trennschärfe der Items wurde die korrigierte Trennschärfe r_{it} per Part-Whole-Korrektur bestimmt, um einer Überschätzung der Trennschärfe vorzubeugen (Luhmann, 2020; Schermelleh-Engel & Werner, 2012). Als Grenzwert für die korrigierte Trennschärfe der Items wird $r_{it} > 0.3$ angenommen (Field et al., 2012).

Explorative Faktorenanalyse

Im Zuge der datengeleiteten Entwicklung des Fachwissenstests wurde das Verfahren der explorativen Faktorenanalyse (EFA) angewendet. Aufgrund der kleinen Stichprobe in der Pilotierung und der großen Anzahl an Fachwissens-Items diente der bereinigte Datensatz der Interventionsstudie zum Post-Zeitpunkt als Grundlage für die EFA. Das Hauptinteresse dieser darin besteht, die Struktur der latenten Faktoren zu erschließen. Aufgrund der Annahme, dass die manifesten Variablen Messfehlervarianz aufweisen wird auf die Form der Hauptachsenanalyse (= Principal Axis Factor Analysis) als Extraktionsmethode zurückgegriffen (Moosbrugger

& Schermelleh-Engel, 2012). Da die Oblimin-Rotation eine Korrelation zwischen den einzelnen Faktoren erlaubt, wurde diese zur Transformation der Faktorladungen in eine Einfachstruktur angewendet (Field et al., 2012). Items mit einer Kommunalität außerhalb der geschätzten Reliabilität (Eid et al., 2015) und einer Faktorladung < 0.30 wurden von der Skala ausgeschlossen. Die Überprüfung der Eignung des Datensatzes erfolgte mittels des Bartlett-Test auf Sphärizität und des Kaiser-Meyer-Olkin (KMO) Kriteriums (ebd.). Während ein KMO-Wert > 0.5 in der Praxis weitläufig als akzeptabel gilt (Field et al., 2012), bemängeln Expert*innen die ausreichende Expertise der meisten Praxisanwender zur Einschätzung des Datensatzes und tendieren daher zu einem Mindestwert von KMO > 0.7 (Ledesma et al., 2021). Folglich wurde sich hier für den höheren Grenzwert entschieden. Zur Reduktion der, in den Datensatz einfließenden, manifesten Indikatoren wurde eine KMO Measure of sampling adequacy (MSA) > 0.5 als Cut-Off-Wert gesetzt (Shrestha, 2021). Die Bestimmung der Dimensionalität (= Faktorenanzahl der Skala) erfolgte unter Zuhilfenahme der Abbruchkriterien von Eigenwerten > 1 (= Kaiserkriterium) sowie des Scree-Tests (Moosbrugger & Schermelleh-Engel, 2012).

Konfirmatorische Faktorenanalyse

Auf Grundlage des bereinigten Datensatzes der Interventionsstudie zum Pre-Zeitpunkt wurde die Passung der in der EFA extrapolierten Faktorenstruktur des Fachwissenstests zur Stichprobe der Interventionsstudie getestet. Zusätzlich sollen über die Analyse einer Kovarianzmatrix die Struktur den latenten Faktoren anhand der manifesten Variablen theoriegeleitet geprüft werden (Moosbrugger & Schermelleh-Engel, 2012). Dazu wurde der Datensatz aufgrund der ordinalen Struktur des Fachwissenstests und der Annahme eines τ -kongenerischen Messmodells einer kategorialen, konfirmatorischen Faktorenanalyse (CCFA) unterzogen (Brown, 2015). Mit 248 Versuchspersonen erfüllt der bereinigte Datensatz die Voraussetzung einer fünf Mal größeren Stichprobe als die Anzahl von 22 Items (Bühner, 2006). Zusätzlich wurde sichergestellt, dass keine gravierenden Abweichungen der Normalverteilung vorliegen (Eid et al., 2015). Aufgrund der ordinalen Datenstruktur sind robuste Schätzer den Maximum Likelihood Verfahren vorzuziehen (Li, 2016). Da die Diagonally Weighted Least Squares (DWLS) Methode bei Stichprobengrößen unter $n = 1000$ (Forero et al., 2009) präzisere Ergebnisse als der Unweighted Least Squares (ULS) Schätzer liefert und zudem weniger anfällig gegenüber der Verletzung einer Normalverteilung ist (Finney & DiStefano, 2006), wird der DWLS-Schätzer verwendet. Zur Bestimmung der Modellgüte wurden verschiedene Indizes zur Modellpassung herangezogen. Zunächst wurde als obere Grenze für den χ^2 -Wert bei einem nicht signifikanten Ergebnis der dreifache Wert der Freiheitsgrade gewählt (Moosbrugger & Schermelleh-Engel, 2012). Die Grenzen wurden für Root Mean Square Error of Approximation (RMSEA) bei < 0.08 und für das Standardized Root Mean Residual (SRMR) als < 0.10 gewählt (ebd.). Zudem

wurden für die deskriptiven Gütemaße Tucker-Lewis Index (TLI) und Comparative Fit Index (CFI) Grenzwerte ≥ 0.95 definiert (Schermelleh-Engel et al., 2003). Darauffolgend wurde die Modellgüte unter Zuhilfenahme des Modifikationsindex verbessert. Die CCFA verfolgt das Ziel die Fachwissensskala zu validieren. Das Verfahren des Item-Parcelings wird daher abgelehnt, da dieses etwaige Fehlspezifizierungen zwischen Modell und Datensatz nicht aufzudecken vermag (Marsh et al., 2013).

Abschließend wurden aus allen Skalen die Items mit unzureichenden Itemkennwerten und Werten aus der Faktorenanalyse nach kritischer und inhaltlicher Betrachtung aus dem Datensatz entfernt (Bühner, 2006).

Ergebnisse

Ergebnisse und Diskussion der Itemanalyse

Die Analyse der Antwortverteilungen der Teilnehmer*innen der Pilotierung ($n = 54$; $n_{\text{♀}} = 41,5\%$; Alter: $MW = 13,5y$ ($0,5$)) hat ergeben, dass die dritte Antwortoption zu Item 01_BmB1 von keiner Versuchsperson ausgewählt wurde.

Die Itemanalyse im Zuge der Skalenentwicklung wurde mit der unbereinigten Stichprobe zum Post-Zeitpunkt der Interventionsstudie durchgeführt. Diese setzt sich aus 11 Klassen an Gymnasien und 5 Klassen an Realschulen-Plus in Rheinland-Pfalz zusammen ($n = 292$ ($n_{\text{♀}} = 124$ ($42,5\%$), $\text{divers} = 1$; Alter: $MW = 13.32y$ (1.06)). Zur Überprüfung der Eignung der Skalen für die inferenzstatistischen Untersuchungen wurde dieser Datensatz reduziert (s. 6.1.3 „Umgang mit fehlenden Werten“ und „Personenbezogene Analyse“) woraus sich letztlich eine Stichprobe $n = 248$ ($n_{\text{♀}} = 106$ ($42,7\%$), $\text{divers} = 1$; Alter: $MW = 13.33y$ (1.26)) ergab. 17 Items der zweiten Version des Fachwissenstests (02_BmB2, 03_BmB3, 04_BmB4, 05_iAM1, 06_Kr1, 09_Dif3, 12_Lsg2, 14_Lsg4, 18_Abd2, 19_Abd3, 27_Allg1, 22_Sieb2, 29_DifS1, 30_DifS1, 34_LsgS2, 36_TrenS2, 37_TrenS3) wurden aufgrund unzureichender Itemkennwerte ausgeschlossen.

Zwei Items der Skala zum ECL weisen eine Schwierigkeit von < 20.00 auf und damit auf Bodeneffekte hin. Da die Items während der Evaluation der Lernmaterialien jedoch noch ausreichende Itemschwierigkeiten aufwiesen, ist davon auszugehen, dass die Überarbeitung der Materialien eine entsprechende Verbesserung in der Bewertung durch die Lernenden mit sich gebracht hat. Auf Basis dieser Annahme werden die Items weiterhin für die Auswertung berücksichtigt. Dies bringt den Vorteil einer besseren Differenzierung im unteren Schwierigkeitsbereich mit sich und lässt somit eine genauere Prüfung auf Unterschiede zu. Zwei Items der Modellverständnisskala (MV7 und MV8) wurden aufgrund unzureichender Trennschärfe ($r_{it} < 0.3$) von der Skala ausgeschlossen.

Tabelle 23: Für die Skalen SI = Situationales Interesse, ICL = Intrinsic Cognitive Load, ECL = Extraneous Cognitive Load, GCL = Germane Cognitive Load, SU = System Usability, FIC = Chemiebezogenes Fachinteresse, FSK = Chemiebezogenes Fähigkeitsselbstkonzept, MV1 = Modellverständnis Version 1, MV2 = Modellverständnis Version 2, FW3 = Fachwissen Version 3 und FW4 = Fachwissen Version 4 sind die Itemkennwerte r_{it} = Korrigierte Trennschärfe, P_{Dahl} = Schwierigkeitsindex nach Dahl, α = Reliabilität nach Cronbach, ω = Reliabilität nach McDonald, Schiefe, Kurtosis, Varianzaufklärung, $n_{effektiv}$ = effektive Stichprobengröße und der Dropout der Items auf Grundlage der Itemanalyse angegeben. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert.

Skala	r_{it} (>.3)	P_{Dahl} (20-80)	α (>.5) ω (>.5)	Schiefe (<2) Kurtosis (<7)	Varianz -aufklärung	$n_{effektiv}$	Dropout
SI	.54 – .77	34.01 – 63.02	.85 .85	-0.50 – 0.56 -0.94 – -0.55	54%	248	0/5
ICL	.89 – .90	21.70 – 26.40	.93 .94	0.96 – 1.29 0.19 – 0.98	83%	248	0/3
ECL	.53 – .77	14.01 – 22.86	.71 .71	1.18 – 1.83 0.49 – 3.23	51%	248	0/3
GCL	.68 – .86	60.90 – 63.89	.86 .90	-0.34 – -0.57 -0.65 – -0.40	71%	248	0/5
SU	.34 – .72	41.61 – 70.33	.82 .83	-1.43 – -0.11 -0.65 – -0.40	32%	248	0/10
FIC	.48 – .83	24.80 – 57.56	.84 .85	-0.84 – 0.37 -0.87 – 0.05	54%	248	0/5
FSK	.48 – .79	52.12 – 61.79	.86 .86	-1.30 – -0.65 -0.65 – 1.59	48%	248	0/7
MV1 (10 Items)	.19 – .71	23.48 – 62.04	.78 .78	-0.20 – 1.31 -1.72 – 1.72	29%	248	2/10
MV2 (8 Items)	.36 – .72	24.39 – 62.04	.80 .81	-0.48 – 1.19 -1.52 – 0.15	35%	248	0/8
FW3 (41 Items)	.14 – .55	13 – 90	.83 .84	-2.67 – 2.14 -2.01 – 5.12	14%	292	20/41
FW4 (21 Items)	.31 – .57	42 – 78	.81 .82	-1.56 – 0.40 -1.77 – 0.91	18%	248	0/21

Ergebnisse und Diskussion der EFA

Die explorative Faktorenanalyse der neu entwickelten Skala zum chemiebezogenen Fachwissen FW3 hat ergeben, dass der Datensatz (Post, ungekürzt; $n = 292$) mit einem Overall-KMO-MSA = 0.75 für die explorative Datenanalyse geeignet ist. Da alle Items einen KMO-MSA > 0.5 aufweisen, wurde kein Item aufgrund dieses Kriteriums ausgeschlossen. Die Nullhypothese des Bartlett-Tests ($\chi^2(820) = 2064.032$, $p < 0.0001$) wurde abgelehnt, wonach die Itemwerte

als nicht unabhängig voneinander angesehen und die Sphärizität als Voraussetzung angenommen werden kann.

Das Screeplot zu den Daten der 24 nach der Itemanalyse verbleibenden Items trägt die Eigenwerte der Faktoren gegen ihre Anzahl auf (Abbildung 19). Die Linie der simulierten Daten schneidet die Linie der empirischen Daten zwischen dem fünften und sechsten Datenpunkt. Die Parallelanalyse weist folglich auf eine mehrfaktorielle Struktur hin (Moosbrugger & Schermelleh-Engel, 2012). Die Linie der empirischen Daten zeigt den typischen Ellenbogenverlauf. Da sich der Knick hier bereits nach dem ersten Faktor zeigt, spricht dies für eine einfaktorielle Lösung (ebd.). Die Differenz zwischen den Eigenwerten von Faktor 1 und Faktor 2 fällt größer aus als die Spanne der Eigenwerte zwischen den restlichen Datenpunkte. Dies spricht für eine Einfaktor-Struktur der Skala (Moosbrugger & Schermelleh-Engel, 2012). Auch wenn der Scree-Test dem Kaiserkriterium vorzuziehen ist, bietet die Betrachtung der Eigenwerte bei

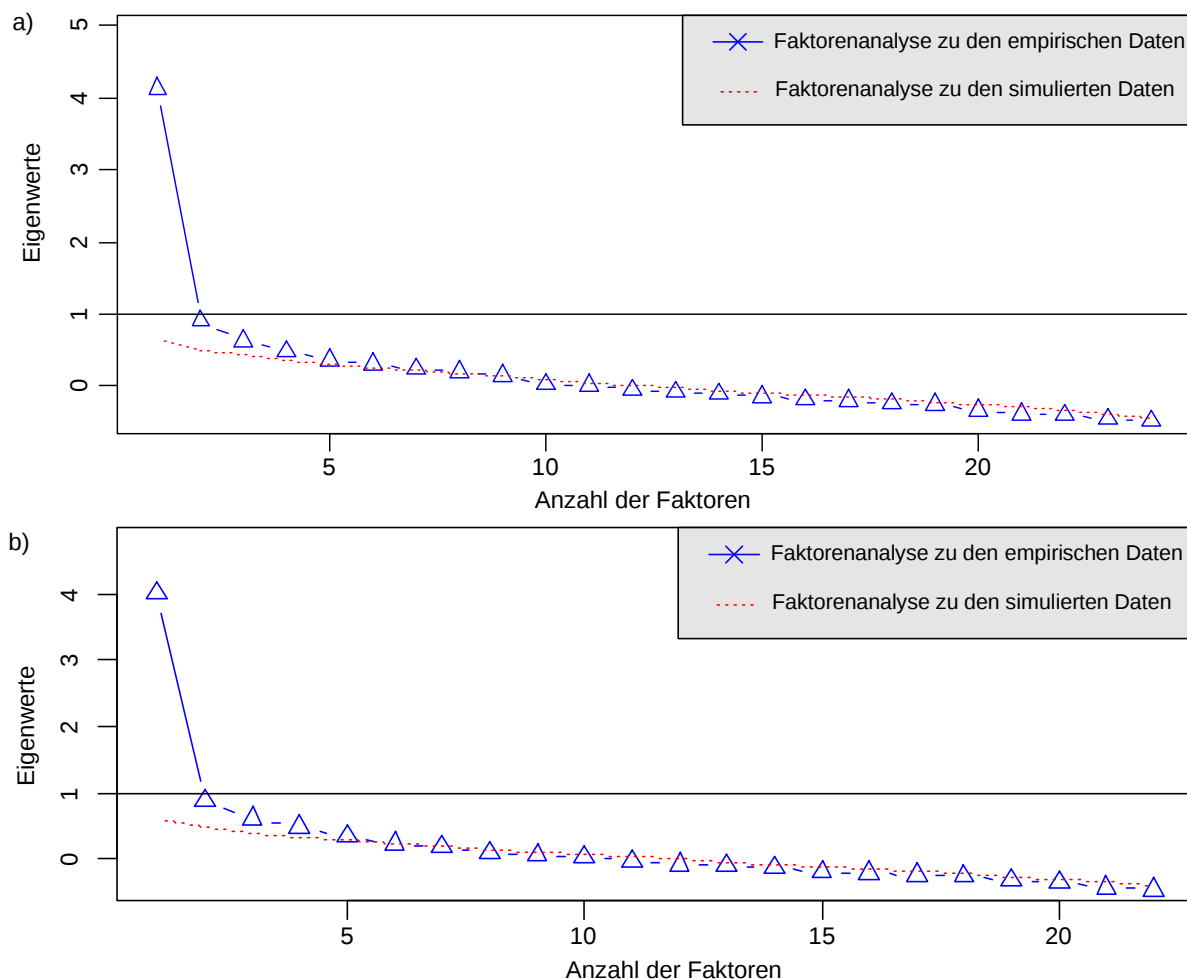


Abbildung 19: Screeplot zur Parallelanalyse zum Fachwissenstest (FW3) nach Ausschluss der Items mit unzureichenden Itemkennwerten. Der Datenverlauf der empirischen Faktoren in blau wird mit dem Verlauf der simulierten Daten in rot verglichen. a) stellt die Parallelanalyse zu den 24 Items verbleibenden Items nach der Itemanalyse dar. b) zeigt die Parallelanalyse zu den 22 Items verbleibenden Items nach der explorativen Faktorenanalyse

uneindeutigen Screeplots eine weitere Betrachtungsmöglichkeit zur Bestimmung der Faktorenstruktur (Eid et al., 2015). Lediglich der erste Faktor weist einen Eigenwert größer als 1 auf. Folglich ist aufgrund der Betrachtung der Eigenwerte ebenfalls von einem einfaktoriellen Modell auszugehen. Der Eigenwert des zweiten Faktors liegt zwar nur knapp unter dem Grenzwert. Die Items, die den zweiten Faktor auf Basis der zweifaktoriellen Hauptachsenanalyse mit obliquer Rotation bilden, weisen jedoch keine inhaltlich begründbaren Zusammenhänge auf. Aufgrund der Mehrzahl an Kriterien, die für eine einfaktorielle Struktur der Skala sprechen und die Annahme weiterer Faktoren sich nicht theoretisch begründen lässt, wird ein einfaktorielles Modell angenommen. Die Ergebnisse der Hauptachsenanalyse zur einfaktoriellen Lösung weisen für die Items *23_Sieb3* und *28_Allg2* mit Werten $< .30$ zu geringe Faktorladungen auf (Luhmann, 2020), worauf die Items aus dem Datensatz entfernt wurden.

Ergebnisse der CCFA

Anschließend wurde die Skala zum Fachwissen (FW3) mit den 22 verbleibenden Items einer kategorialen, konfirmatorischen Faktorenanalyse (CCFA) unterzogen. Da alle Items mit unzureichenden Werten für Schiefe und Kurtosis von der Skala ausgeschlossen wurden, können gravierende Abweichungen von der Normalverteilung ausgeschlossen werden (Eid et al., 2015). Die Ergebnisse zum robusten Schätzverfahren zeigen, dass das entsprechende Modell nicht ausreichend zur Struktur der vorliegenden Daten passt (vgl. Tabelle 24). Daher wurde die Modellgüte für die CCFA optimiert, indem das Item mit dem höchsten Modifikationsindex (Item *07_Dif1*) ausgeschlossen wurde (Döring & Bortz, 2016). Die erneute CCFA zum auf 21 Items reduzierten Modell zeigt für alle Fitindices zufriedenstellende Ergebnisse (vgl. Tabelle 24).

Tabelle 24: Vergleich der zwei Messmodelle zur Fachwissensskala anhand des χ^2 -Tests mit zugehörigen p -Werten im Verhältnis des χ^2 -Werts mit der dreifachen Anzahl der Freiheitsgrade df , CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Residual. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert. Alle Werte wurden mit dem robusten DWLS-Schätzer und NLMINB-Optimierungsmethode berechnet.

Messmodell	Frei geschätzte Modellparameter	χ^2 -Test $p > 0.05$ $\chi^2 < 3*df$	CFI TLI > 0.95	RMSEA < 0.08 $p > 0.05$	SRMR < 0.10
mit Item 07	44 $< 0.2*n (= 222)$	0.014 256.297 < 3*209	0.915 0.906	0.032 0.993	0.098
ohne Item 07	42 $< 0.2*n (= 222)$	0.128 211.278 < 3*189	0.955 0.950	0.023 0.999	0.093

Weitere Anpassungen der Skalen

Die Erkenntnisse aus der Itemanalyse der zweiten Version des Fachwissenstests auf Basis der Antwortverteilungen der Teilnehmer*innen der Pilotierung zeigen Anpassungsbedarf auf. Da keine Person die dritte Antwortalternative zu Item 01_BmB1 gewählt hat, wurde diese überarbeitet. Die Aussage wurde durch das bereits bekannte Präkonzept der Übertragung des Phänomens der Volumenvergrößerung auf Teilchen ersetzt (s. Tabelle 25).

Tabelle 25: Anpassung der Itemformulierung. Die Formulierung vor der Überarbeitung ist durchgestrichen.

01 BmB1	Je höher die Temperatur eines Stoffs ist, ...
☐ A	... desto schwerer werden seine Teilchen.
☐ B	... desto leichter werden seine Teilchen.
☐ C	... desto langsamer werden seine Teilchen. ... desto größer werden seine Teilchen.
☐ D	... desto schneller bewegen sich seine Teilchen.

Ergebnisse der personenbezogenen Analyse

Nach der Überarbeitung der Skalen erfolgte die personenbezogene Analyse zur Stichprobe der Interventionsstudie. Die Maximalwerte an fehlenden Werten pro Versuchsperson belaufen sich auf 8 Items zu Messzeitpunkt 0, 12 Items zu Messzeitpunkt 4 und zu den Messzeitpunkten 1 bis 3 auf jeweils 25 Items. Die hohe Anzahl an fehlenden Werten zu den Messzeitpunkten 1 bis 3 beruht auf fehlenden Fragebögen von einzelnen Versuchspersonen. Aufgrund von fehlenden Werten zu mindesten einer gesamten Skala wurden die Daten von 43 Proband*innen listenweise gelöscht. Dies beinhaltet sowohl Personen, die ihren Fragebogen nicht eingereicht haben, als auch Personen mit mindestens einem Fehltermin. Aufgrund eines Extremwerts im MiniQ-Test erfolgte der listenweise Ausschluss einer Probandin. Somit beläuft sich die Stichprobe des bereinigten Datensatzes auf 248 Proband*innen.

Bei 26 Versuchspersonen wurden fehlende Werte im Fachwissenstest mit null Punkten kodiert. Für die Skala Modellverständnis traten bei zehn Teilnehmenden fehlende Werte auf, die durch den Skalenmittelwert der Versuchsperson ergänzt wurden. Gleiches gilt für zwölf Personen in Bezug auf die Skala *System Usability*, zwei Personen *ICL*, drei Personen *ECL*, vier Personen *GCL*, zwei Personen *Situationales Interesse*, eine Person *chemiebezogenes Fachinteresse* und eine Person *chemiebezogenes Fähigkeitsselbstkonzept*. Weiterhin wurden bei keinem Konstrukt Kreuzungsmuster festgestellt und folglich keine weiteren Teilnehmenden ausgeschlossen.

8.3 Interventionsstudie

Die vorliegende Interventionsstudie geht der zentrale Forschungsfrage nach, wie sich die unterschiedlichen Modelltypen auf das situationale Interesse, das Modellverständnis und das Fachwissen von Schüler*innen der Klassenstufe acht auswirken. Tabelle 26 bietet eine Übersicht über die Studie, einschließlich der eingesetzten Methoden zur Datenerhebung und Datenauswertung sowie der verfügbaren Stichprobe.

Tabelle 26: Methoden und Stichprobe der Untersuchungen auf Gruppenunterschiede

Untersuchungen auf Gruppenunterschiede	
Methode der Datenerhebung	Levene-Test
Methoden der Datenauswertung	Einfaktorielle, messwiederholte ANOVA
	Oneway-Varianzanalyse mit Welch-Korrektur
	Mauchly-Test
	TukeyHSD-Test
Stichprobe	n = 248 Schülerinnen und Schüler

8.3.1 Methoden der Datenerhebung

Die Versuchspersonen wurden klassenweise randomisiert, sodass die vorliegende Studie ein quasi-experimentelles Design aufweist (Miller et al., 2020). Die Daten wurden im Zeitraum von Oktober 2020 bis Januar 2021 in insgesamt 16 Klassen erhoben. Die Einteilung der Klassen in die jeweilige Experimentalgruppe ist in Tabelle 27 dargestellt.

Tabelle 27: Einteilung der Klassen in die Experimentallgruppen. Den teilnehmenden Schulen wurde ein Farbcode zugeordnet. Der Lehrkräfte-Code ist kursiv dargestellt.

AR	HI	IL
GYM1 8a <i>WI</i>	GYM1 8c <i>WI</i>	GYM1 8d <i>BY</i>
GYM1 8b <i>BY</i>	GYM2 8b <i>JU</i>	GYM2 8e <i>JU</i>
GYM2 8c <i>KA</i>	GYM3 8a <i>AL</i>	GYM2 8a <i>FL</i>
GYM3 8c <i>JA</i>	IGS1 8c <i>DI</i>	GYM3 8d <i>JA</i>
IGS1 8a <i>SC</i>	IGS1 8d <i>SC</i>	IGS1 8b <i>DI</i>
		IGS2 8b <i>DD</i>

8.3.2 Fragebogenerhebung

Die Datenerhebungen erfolgten zum Pre- und Postzeitpunkt – in der Unterrichtsstunde vor Intervention I beziehungsweise nach Intervention III – sowie begleitend jeweils direkt im Anschluss an die Verwendung der Modelle während der Interventionszeitpunkte. Das Forschungsdesign mit den erhobenen Konstrukten zu den konkreten Erhebungszeitpunkten ist in Tabelle 28 veranschaulicht. Die eingesetzten Fragebögen befinden sich in Anhang 10 bis 11. Als Methode zur Erhebung der Daten kamen Fragebögen mit geschlossenem Antwortformat zu Einsatz. Die Wahl der genutzten Skalen ist durch ihre inhaltliche Passung und der Passung zur Zielgruppe sowie deren ausführliche Erprobung und ausreichende Testgüte in vorangegangenen Studien begründet. Die eingesetzten Messinstrumente inklusive der Analysen der Skalen sind in Abschnitt 8.2 beschrieben.

Tabelle 28: Ablauf der Datenerhebung mittels Fragebögen während der Interventionsstudie. Grau unterlegt sind die Konstrukte, die Pre und Post erhoben wurden; weiß unterlegt: Konstrukte, die begleitend zu den Interventionszeitpunkten erhoben wurden.

MZP 0 Pre	MZP 1 Intervention I	MZP 2 Intervention II	MZP 3 Intervention III	MZP 4 Post
Modellverständnis	Situationales Interesse	Situationales Interesse	Situationales Interesse	Modellverständnis
Fachwissen	Cognitive Load	Cognitive Load	Cognitive Load	Fachwissen
Fachinteresse	System Usability	System Usability	System Usability	Fachinteresse
Fachbezogenes Selbstkonzept				Fachbezogenes Selbstkonzept
Kognitive Fähigkeiten				
Fachnote und Gesamtschnitt				
Demografische Daten				

Basierend auf Abschnitt 4 *Ableitung der Forschungsfragen und Hypothesen* wurden folgende Hypothesen überprüft:

H3 Die Entwicklung des situationalen Interesses der einzelnen Gruppen (AR, HI & IL) unterscheidet sich über den Interventionszeitraum hinweg.

H3.1 Die Lernenden zeigen während der Intervention ein höheres situationales Interesse im Umgang mit den AR-Modellen als im Umgang mit den haptisch-interaktiven Modellen.

H3.2 Die Lernenden zeigen während der Intervention ein höheres situationales Interesse im Umgang mit den AR-Modellen als im Umgang mit den illustrativen Modellen.

H4 Das Modellverständnis der verschiedenen Gruppen (AR, HI, IL) entwickelt sich über die einzelnen Messzeitpunkte unterschiedlich stark.

H4.1 Das Modellverständnis der Schüler*innen in der Gruppe AR nimmt mehr zu als in der Gruppe HI.

H4.2 Das Modellverständnis der Schüler*innen in der Gruppe AR nimmt mehr zu als das der IL.

H5 Das Fachwissen der verschiedenen Gruppen (AR, HI, IL) entwickelt sich über die einzelnen Messzeitpunkte unterschiedlich stark.

H5.1 Das Fachwissen der Schüler*innen in der Gruppe AR nimmt mehr zu als in der Gruppe HI.

H5.2 Das Fachwissen der Schüler*innen in der Gruppe AR nimmt mehr zu als das der IL.

Die Konstrukte *Cognitive Load* (Leppink et al., 2013) und *System Usability* (Brooke, 1996), *chemiebezogenes Fähigkeitsselbstkonzept* (Engl, 2020), *chemiebezogenes Fachinteresse* (Engl, 2020) und *kognitive Fähigkeiten* (Baudson & Preckel, 2016) sowie *demografische Daten* und *Zeugnisnoten* dienten als Kontrollvariablen (vgl. Tabelle 28).

Interventionsmaßnahmen

Die Intervention besteht aus drei Doppelstunden unter Aufsicht des Versuchsleiters und von wissenschaftlichen Hilfskräften. Die Lerneinheiten wurden alle in den gewohnten Chemiesälen der jeweiligen Schulen vor Ort durchgeführt. Die restlichen Einflussfaktoren wurden bestmöglich standardisiert, um einen fairen Gruppenvergleich zu ermöglichen (Theyßen, 2014). Alle Lerngruppen entstammen der Klassenstufe acht. Nahmen mehrere Klassen pro Schule und oder Lehrkraft teil, erfolgte eine gleichmäßige Verteilung auf die Experimentalgruppen eins bis drei. Um den Einfluss des Versuchsleiters zu minimieren, wurden die Interventionen in allen Klassen von derselben Person durchgeführt (ebd.). Zudem wurden die Lernumgebungen so konzipiert, dass nahezu alle Informationen durch das Skript vermittelt werden. Auch die praktische Arbeitsphase erfolgte eigenverantwortlich und selbstständig in Einzelarbeit (ebd.). Es wurde darauf geachtet, dass die Interventionen zu allen drei Modelltypen annähernd gleich konzipiert sind. Bis auf den eingesetzten Modelltyp sind die Rahmenbedingungen der Interventionen in den Vergleichsgruppen komplett identisch gestaltet (s. Abschnitt 5.65). Dies gilt für die Experimente, Aufgabenstellungen, Handlungsaufforderungen und Lernziele ebenso wie für das Layout der Materialien. Lediglich der Wortlaut zu den Modellanleitungen weist notwendige Unterschiede auf. Folglich galt der Neuheitseffekt aufgrund des Eventcharakters gleichermaßen für alle drei Vergleichsgruppen (ebd.). Die Inhalte der Lerneinheiten unterscheiden sich nur in Bezug auf die genutzten Modelltypen. Im Zuge der Pilotierung wurde sichergestellt, dass alle Lerneinheiten in allen Vergleichsgruppen eine Zeitspanne von 90

Minuten und damit eine gleiche „Time on Task“ aufweisen. Der Ablauf jeder Lerneinheit ist dreiteilig gegliedert. Nach einer fünfminütigen Einführung in den Ablauf des jeweiligen Tages erstreckte sich die Erarbeitungsphase über 70 Minuten. In den 15 Minuten der Abschlussbesprechung wurden die Beobachtungen und Ergebnisse zu den Experimenten und Modellen besprochen und interpretiert sowie kritisch reflektiert.

8.3.3 Methoden der Datenauswertung

Die Überprüfung der Hypothesen erfolgt über mehrere Schritte:

(1) Datenmanagement, (2) Itemanalyse, (3) Personenbezogene Analyse und (4) Unterschiedsprüfungen. Die Schritte eins bis drei wurden bereits in den Abschnitten 8.1.2 und 8.2 beschrieben, sodass im Folgenden lediglich die Unterschiedsprüfungen aufgeführt werden.

Zur Auswertung der deskriptiven Statistiken und der Unterschiedsprüfungen wurde die Software *R-Studio* (Version 2021.09.1) mit *R* (Version 3.14.0) unter Verwendung der Pakete *utils*, *likert*, *psych*, *dplyr*, *stats*, *WRS2*, *car*, *graphics*, *welchADF*, *MBESS*, *effectsize*, *rstatix*, *ez*, *ggplot2* und *agricolae* genutzt. Zunächst wurden Unterschiedsprüfungen zu den Kontrollvariablen durchgeführt, um die Vergleichbarkeit der Stichproben in den einzelnen Experimentalgruppen zu überprüfen. Die Geschlechterverteilung zwischen den unterschiedlichen Experimentalgruppen wurde mithilfe des nonparametrischen χ^2 -Tests überprüft (Luhmann, 2020). Zur Bestimmung der Effektgröße wird das Odds Ratio (OR) berechnet. Diese Effektgröße wird in folgende Kategorien eingeteilt: ein kleiner Effekt liegt bei $OR > 2$ vor, ein mittlerer Effekt bei $OR > 3$ und ein großer Effekt bei $OR > 7$ (Döring & Bortz, 2016). Infolge der Robustheit der ANOVA gegenüber Verletzungen der Normalverteilungsannahme und einer Stichprobengröße von $n > 30$ ($n = 248$) wurde der zentrale Grenzwertsatz angewendet (Blanca et al., 2017; Schmider et al., 2010). Weiterhin wurde die Normalverteilung der Residuen der Faktoren mithilfe QQ-Plots und Histogrammen begründet (Eckstein, 2012). Die Voraussetzung der Varianzhomogenität wurde mittels des Levene-Tests überprüft. Zur inferenzstatistischen Auswertung bei vorhandener Varianzhomogenität (*Alter*, *Note NaWi*, *Zeugnisschnitt*, *Fachinteresse*, *Fähigkeitsselbstkonzept*, *Kognitive Fähigkeiten*, *ICL*, *ECL*, *GCL*, *Situationales Interesse*, *Fachwissen zu MZP1* und *Modellverständnis zu MZP1*) wurde eine einfaktorielle (messwiederholte) ANOVA (Döring & Bortz, 2016) durchgeführt. Zur Post-Hoc-Analyse fand der TukeyHSD-Test Anwendung. Die Effektgrößen geben bei $\eta_g^2 > 0.26$ einen großen, bei $\eta_g^2 > 0.13$ einen mittleren oder bei $\eta_g^2 > 0.02$ einen kleinen Effekt an (Olejnik & Algina, 2003). Bei Verletzung der Varianzhomogenität wurden die entsprechenden Konstrukte (*Fachwissen-messwiederholt* und *Modellverständnis-messwiederholt*) mittels messwiederholtem Welch-James-ADF-Test (Welch, 1951) unter Anwendung des Winsorizing- und Trimmverfahrens untersucht (Villacorta, 2017). Für die Nutzung des hier empfohlenen Winsorizings um den Faktor 0.642 wird ein Trimmen der Daten zum Entfernen von Ausreißern vorausgesetzt (ebd.). Anstelle des konservativen Ansatzes eines 20-prozentigen Trimmens (ebd.) wurde sich zugunsten einer größeren Varianz im Datensatz (Prendergast, 2008) zu einem Trimm-Maß auf Basis der Cook'schen Distanz als Cut-Off Kriterium entschieden (R. D. Cook & Weisberg, 1982). Über einen Vergleich der paarweisen Kontraste wurde die Post-Hoc-Analyse durchgeführt. Die Effektgrößen für den Winsorized Welch-James-ADF-Test entsprechen den Grenzwerten für

Cohen's d (Cohen, 1988; Sawilowsky, 2009) und geben bei $\hat{\delta}_j^{(R)} > 1.2$ einen sehr großen, bei $\hat{\delta}_j^{(R)} > 0.8$ einen großen, bei $\hat{\delta}_j^{(R)} > 0.5$ einen mittleren und bei $\hat{\delta}_j^{(R)} > 0.2$ einen kleinen Effekt an (Villacorta, 2017). Die Effektstärken für den Welch-James-ADF-Test lassen sich nur über paarweise Kontraste berechnen (ebd.). Daher wurde die Effektstärke für den Interaktionseffekt zum Modellverständnis über das ANOVA-Modell zum vorliegenden Datensatz berechnet.

Bei Faktoren mit Testungen zu drei Messzeitpunkten (*situationales Interesse*, *System Usability*, *ICL*, *ECL* und *GCL*), wurden die Daten unter der Annahme, dass eine durchgehende Compound Symmetry keine Voraussetzung für die F-Statistik einer einfaktoriellen, messwiederholten ANOVA ist, lediglich auf die Bedingung Sphärizität mittels Mauchly-Test geprüft (Eid et al., 2015; Huynh & Feldt, 1976). Bei Verletzung der Sphärizität ($\epsilon \neq 1$) wurde eine Korrektur der Freiheitsgrade vorgenommen (Huynh & Feldt, 1976). Für Faktoren, die nur zwei Stufen aufweisen, kann die Sphärizität nicht verletzt sein (Rasch et al., 2014b). Insofern erfolgte auch keine Korrektur der Freiheitsgrade (ebd.).

Im Anschluss an die inferenzstatistischen Auswertungen zum Modellverständnis wurden die Post-Ergebnisse zur Modellverständnisskala mithilfe eines Einstichproben-t-Tests auf die generelle Lernwirksamkeit der konzipierten Lernumgebungen überprüft (Eid et al., 2015). Die verglichene Normierungsstudie wurde aufgrund der Konformität zur Stichprobe und der inhaltlichen Zielsetzung der vorliegenden Interventionsstudie ausgewählt. In der Normierungsstudie von Leisner (2005a) nahmen 110 Schüler*innen ($n_{\text{♀}} = 70 \triangleq 63,6\%$) am Ende der siebten Klasse an einem sechs bis achtwöchigen Unterricht über Modelle teil. Zunächst gilt für den Einstichproben-t-Test die Voraussetzung der Intervallskalierung der Daten (Eid et al., 2015). Aufgrund neuerer Erkenntnisse, die eine Robustheit des t-Tests bei Verletzung der Normalverteilungsannahme bei großen Stichproben ($n > 30$) zeigen, wird keine ausführliche Verteilungsprüfung durchgeführt (Pagano, 2013). Die Kennwerte auf Ebene der Stichprobe werden den Testungen der Voraussetzungen für die Anwendung der varianzanalytischen Verfahren entnommen.

Abschließend wurden die gefundenen Effekte der inferenzstatistischen Untersuchungen auf die Wahrscheinlichkeit, mit der ein Fehler zweiter Art vorliegt, geprüft (Cohen, 1988). Dazu wurde mittels G*Power (Version 3.1.9.4) eine Post-Hoc-Poweranalyse durchgeführt. Zur Bestimmung der Effektstärke f wurde diese aus den Effektstärken η_g^2 und $\hat{\delta}_j^{(R)}$ aus den varianzanalytischen Verfahren zu den jeweiligen Konstrukten unter Anwendung der Formeln nach Cohen (1988) berechnet. Die Korrelationen zwischen den Messwiederholungen wurden mithilfe der *cor*-Funktion aus dem Paket *stats* in R-Studio berechnet. Die Werte für Epsilon zur Korrektur der Nicht-Sphärizität werden den Ergebnissen der varianzanalytischen Verfahren entnommen. Aufgrund der kontinuierlichen, intervallskalierten Datenstruktur und der linearen

Beziehung der Daten wurde die Korrelation zwischen den Messzeitpunkten nach Pearson berechnet (Hauke & Kossowski, 2011).

8.3.4 Ergebnisse

Stichprobe:

- 248 Schüler*innen, $n = 106$ (42,7%), divers = 1; Alter: MW = 13.33y, aus 16 Klassen

Kovariablen:

- Es zeigten sich keine signifikanten Unterschiede zwischen den Gruppen zu MZP0/MZP1 oder Interaktionseffekte in den Variablen Geschlecht, Alter, Schulnote, Fachinteresse, Fähigkeitsselbstkonzept, Kognitive Fähigkeiten, Cognitive Load und System Usability.
- Es traten signifikanten Effekte zwischen den Messzeitpunkten im ICL und ECL auf.

Situationales Interesse:

- Es wurden keine signifikanten Unterschiede zwischen den Gruppen zu MZP1 oder Interaktionseffekte gefunden.
- Es trat ein signifikanter Abfall zwischen MZP1 und MZP2 sowie zwischen MZP1 und MZP3 wurde auf.

Fachwissen:

- Die Gruppe IL weist ein signifikant höheres Niveau im Vergleich zu den beiden Gruppen AR und HI zu MZP0 auf.
- Es wurden keine signifikanten Interaktionseffekte festgestellt.
- Es trat ein signifikanter, großer Lerneffekt zwischen den Messzeitpunkten auf.

Modellverständnis:

- Es liegen keine signifikanten Unterschiede zwischen den Gruppen zu MZP0 vor.
- Es trat ein signifikanter, sehr großer Lerneffekt zwischen den Messzeitpunkten auf.
- Signifikante Interaktionseffekte:
 - IL weist einen Vorteil mittlerer Effektstärke gegenüber HI auf.
 - AR weist einen Vorteil großer Effektstärke gegenüber HI auf.
- Die Gesamtstichprobe zeigt ein mit einem kleinen Effekt signifikant niedriger ausgeprägtes Modellverständnis zu MZP4 als die Stichprobe der Normierungsstudie von Leisner 2005a.

Im Zuge des Datenmanagements (s. Abschnitt 8.2) ergab sich für die Interventionsstudie eine Stichprobe von insgesamt 248 Schüler*innen [$n_{\text{♀}} = 106$ (42,7%), divers = 1; Alter: MW = 13.33y (SD = 1.26)] aus 16 achten Klassen an zwei Realschulen-Plus und vier Gymnasien in Rheinland-Pfalz. Es wurden 127 Personen (33,6%) aufgrund von Fehlterminen bei Messzeitpunkten sowie eine Person aufgrund eines fehlenden Wertes und eine Person aufgrund eines Extremwerts in der MiniQ-Skala ausgeschlossen (insgesamt 34,2%).

Ergebnisse der Analyse der Kovariablen

Begleitend zu den hypothesengeleiteten Unterschiedsprüfungen wurde die Stichprobe in einem ersten Schritt auf Gruppenunterschiede in Bezug auf die Kovariablen zum Messzeitpunkt 0 (Pre) untersucht. Die deskriptiven Statistiken der Kovariablen und die Ergebnisse der Unterschiedsprüfungen sind in Tabelle 29 dargestellt. Für keine der Kovariablen konnte in den inferenzstatistischen Analysen ein signifikanter Unterschied zum Pre-Zeitpunkt aufgezeigt werden. Beim Faktor Geschlecht liegt nur eine Person in der HI-Gruppe vor, die sich als divers identifiziert. Da diese Ausprägung in den anderen Gruppen gleich Null ist, ließ sich hier keine Odds-Ratio berechnen.

Weiterhin wurden für die Konstrukte *Fachinteresse Chemie (FIC)*, *Fähigkeitsselbstkonzept (FSK)*, *System Usability (SU)* und die drei Teildimensionen des *Cognitive Loads* messwiederholte Unterschiedsprüfungen durchgeführt. Aufgrund der ausbleibenden Signifikanzen der Levene-Tests zu den einzelnen Konstrukten erfolgten die Überprüfungen mittels messwiederholter, einfaktorieller ANOVA. Aufgrund der signifikanten Ergebnisse im Mauchly-Test wurde für die Konstrukte *SU*, *ICL*, *ECL*, *GCL* und *situationales Interesse* eine Korrektur der Freiheitsgrade vorgenommen. Da die Verletzungen der Sphärizität den Grenzwert von $\epsilon > 0.75$ nicht unterschreiten, erfolgte die Korrektur der Freiheitsgrade nach Huynh und Feldt (1976). Für die Konstrukte *Fachinteresse Chemie*, *Fähigkeitsselbstkonzept*, *System Usability* und *Germane Cognitive Load (GCL)* ergaben sich für die inferenzstatistischen Analysen keine signifikanten Unterschiede auf der Ebene der Gruppen, der Messzeitpunkte oder der Interaktionseffekte. Die deskriptiven Statistiken der Kovariablen zu allen Messzeitpunkten sind in Anhang 12 und die Ergebnisse der nicht signifikanten Unterschiedsprüfungen in Anhang 13 tabellarisch aufgelistet.

Tabelle 29: Unterschiede der Kovariablen zwischen den Experimentalgruppen zum Messzeitpunkt 0 (Pre) mit den Levene-Teststatistiken zur Analyse mittels einfaktorieller ANOVA und der zugehörigen Effektgröße. AR = Augmented Reality, HI = Haptisch-Interaktiv, MW = Mittelwert, SD = Standardabweichung, OR = Odds Ratio.

Kovariable	Deskriptive Statistiken	Levene-Teststatistik	Unterschiedsprüfungen und Effektstärken
Geschlecht	AR: n♀ = 47, n♂ = 29, ndiv. = 0	-	$\chi^2(2) = 4.00, p = 0.41$
	HI: n♀ = 45, n♂ = 31, ndiv. = 1		$OR_{AR,HI_{m/w}} = 1.12$
	IL: n♀ = 49, n♂ = 45, ndiv. = 0		$OR_{AR,IL_{m/w}} = 1.49$
			$OR_{HI,IL_{m/w}} = 1.33$
Alter	AR: MW = 13.19, SD = 0.49	$F(2, 245) = 2.09, p = 0.126$	$F(2, 245) = 1.00, p = 0.370$
	HI: MW = 13.48, SD = 2.08		$\eta_g^2 = 0.008$
	IL: MW = 13.33, SD = 0.65		
Note NaWi	AR: MW = 2.12, SD = 0.67	$F(2, 245) = 0.44, p = 0.646$	$F(2, 245) = 0.48, p = 0.621$
	HI: MW = 2.16, SD = 0.8		$\eta_g^2 = 0.004$
	IL: MW = 2.22, SD = 0.71		
Zeugnissschnitt	AR: MW = 2.32, SD = 0.69	$F(2, 245) = 0.87, p = 0.420$	$F(2, 245) = 0.32, p = 0.854$
	HI: MW = 2.26, SD = 0.71		$\eta_g^2 = 0.001$
	IL: MW = 2.29, SD = 0.61		
Fachinteresse	AR: MW = 3.03, SD = 0.57	$F(2, 245) = 0.05, p = 0.955$	$F(2, 245) = 1.00, p = 0.203$
	HI: MW = 2.83, SD = 0.58		$\eta_g^2 = 0.013$
	IL: MW = 3.02, SD = 0.59		
Fähigkeits-selbstkonzept	AR: MW = 3.27, SD = 0.53	$F(2, 245) = 1.39, p = 0.252$	$F(2, 245) = 0.14, p = 0.872$
	HI: MW = 3.21, SD = 0.6		$\eta_g^2 = 0.011$
	IL: MW = 3.21, SD = 0.63		
Kognitive Fähigkeiten	AR: MW = 29.21, SD = 9.38	$F(2, 245) = 0.43, p = 0.651$	$F(2, 245) = 0.55, p = 0.578$
	HI: MW = 27.74, SD = 8.67		$\eta_g^2 = 0.004$
	IL: MW = 28.62, SD = 8.25		

Die deskriptive Verteilung der durchschnittlichen Werte der Subskalen des *Cognitive Loads* (Wertebereich 0-10) über alle Gruppen und Messzeitpunkte hinweg gestaltet sich wie folgt. Der ECL fällt mit einem Mittelwert von MW = 1.86 am geringsten aus. Für den ICL ergibt sich ein Durchschnittswert von MW = 2.45. Mit einem Wert von MW = 6.26 fällt der GCL am größten aus.

Für die messwiederholte ANOVA zum *Intrinsic Cognitive Load (ICL)* zeigt sich nach Huynh-Feld-Korrektur ein signifikanter, marginaler Effekt zwischen den Messzeitpunkten ($F(2, 490) = 16.43, p[HF] < 0.001, \eta_g^2 < 0.02$). Das Ergebnis des TukeyHSD-Tests zeigt, dass dieser Unterschied nur zwischen MZP1 und MZP2 signifikant wird (MZP1 - MZP2: $p < 0.001$; MZP1 - MZP3: $p = 0.08$; MZP2 - MZP3: $p = 0.33$). Abbildung 20 zeigt die Entwicklung des ICLs in den Vergleichsgruppen über die drei Messzeitpunkte hinweg.

Die Post-Hoc-Poweranalyse zum gefundenen Effekt bringt folgende Ergebnisse. Mit der kalkulierten Effektstärke von $f = 0.140$ und eine der Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers mit einer Wahrscheinlichkeit von 0.01 (Cohen, 1988) ergibt sich für die Gesamtstichprobe von 248 Observationen eine Power von 1.0. Folglich gilt die Teststärke als ausreichend und ein Fehler zweiter Ordnung kann ausgeschlossen werden (Cohen, 1988; Eid et al., 2015). Bei einer Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.05 und einer Power von 0.95 ergibt sich für die ermittelte Effektstärke eine Gesamtstichprobengröße von 147 Observationen.

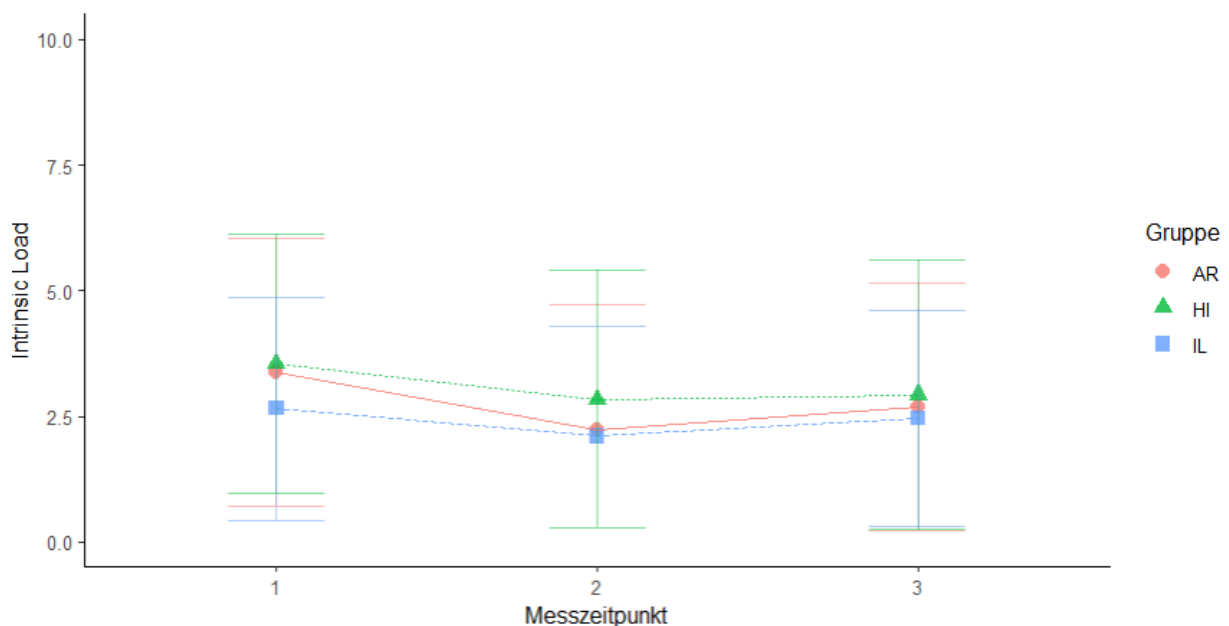


Abbildung 20: Entwicklung des *Intrinsic Cognitive Loads* zu den drei Interventionszeitpunkten im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau. Ein hoher Zahlenwert bedeutet eine hohe kognitive Belastung aufgrund der Aufgabenschwierigkeit.

Die messwiederholte ANOVA zum *ECL* ergibt nach Huynh-Feld-Korrektur einen signifikanten, marginalen Effekt zu den Messzeitpunkten ($F(2, 490) = 4.52$, $p[HF] = 0.012$, $\eta_g^2 < 0.02$). Der TukeyHSD-Test ergibt keine signifikanten Unterschiede zwischen den einzelnen Messzeitpunkten (MZIP1 - MZIP2: $p = 0.87$; MZIP1 - MZIP3: $p = 0.38$; MZIP2 - MZIP3: $p = 0.16$). Daraus lässt sich schließen, dass der gefundene Effekt sich aus den nicht-signifikanten Unterschieden zwischen allen Messpunkten ableitet. Abbildung 21 zeigt die Entwicklung des ECLs in den Vergleichsgruppen über die drei Messzeitpunkte hinweg.

Aufgrund der Post-Hoc-Poweranalyse zum gefundenen Effekt zeigen sich folgende Ergebnisse. Mit der kalkulierten Effektstärke von $f = 0.07$ und der Annahme eines Alpha-Fehlers mit einer Wahrscheinlichkeit von 0.05 (Cohen, 1988) ergibt sich für die Gesamtstichprobe von 248 Observationen eine Power von 0.77. Folglich ist die Teststärke unzureichend und ein Fehler zweiter Ordnung liegt mit einer Wahrscheinlichkeit von 23% vor (Cohen, 1988; Eid et al., 2015). Bei einer Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.05 und einer Power von 0.95 ergibt sich für die ermittelte Effektstärke eine Gesamtstichprobengröße von 510 Observationen, um einen signifikanten Unterschied aufzeigen zu können.

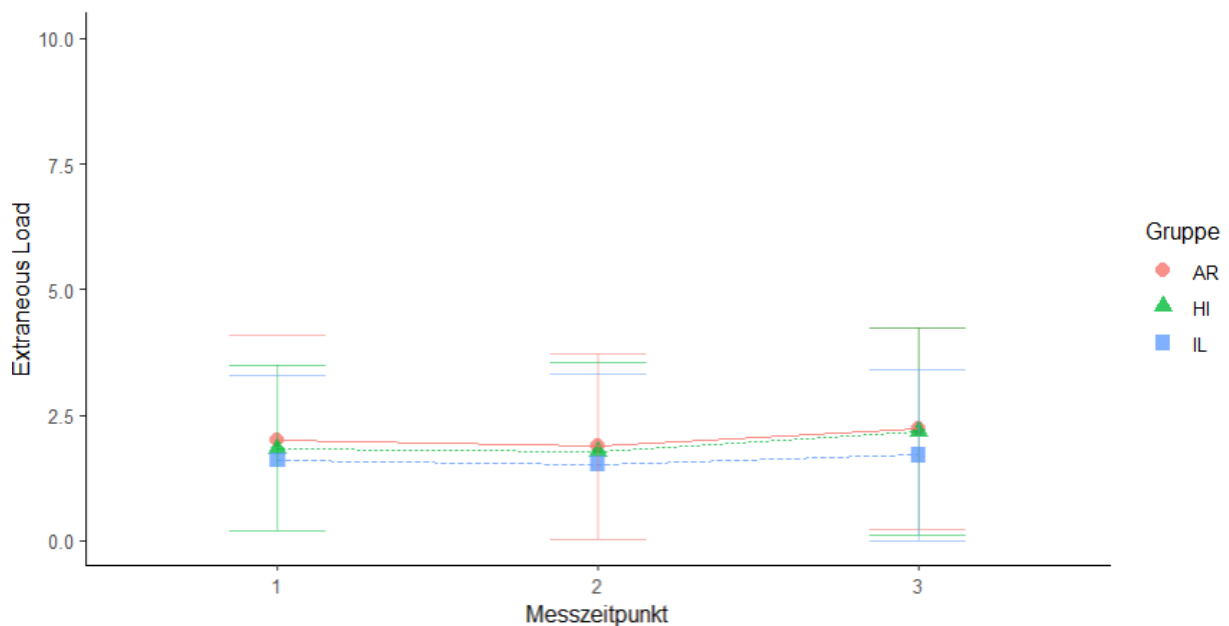


Abbildung 21: Entwicklung des Extraneous Cognitive Loads zu den drei Interventionszeitpunkten im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau. Ein hoher Zahlenwert bedeutet eine hohe kognitive Belastung aufgrund der Arbeitsmaterialien.

Situationalles Interesse

Das *situationale Interesse* wurde von der Gesamtstichprobe der Interventionsstudie im Mittel über alle drei Messzeitpunkte hinweg mit einem mittleren Wert von 4,5 (Skala von 1-7) bewertet. Die Analyse des Datensatzes auf Normalverteilung zeigt beim Konstrukt situationales Interesse sowohl im Histogramm zur Antwortverteilung (s. Abbildung 22) als im QQ-Plot zur Verteilung der studentisierten Residuen der messwiederholten einfaktoriellen ANOVA (s. Abbildung 23) keine gravierenden Abweichungen von der Normalverteilung. Da der Levene-Test für den vorliegenden Datensatz nicht signifikant ausfällt ($F(8, 735) = 1.49, p = 0.156$), kann Varianzhomogenität angenommen werden.

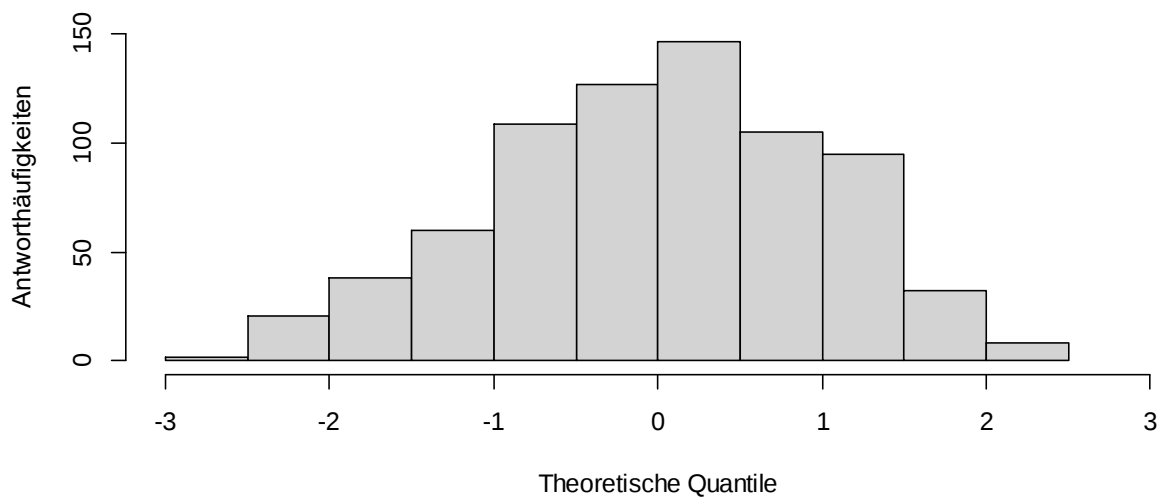


Abbildung 22: Histogramm zur Verteilung der Antworten der Gesamtstichprobe zum situationalen Interesse über die theoretischen ermittelten Quantile

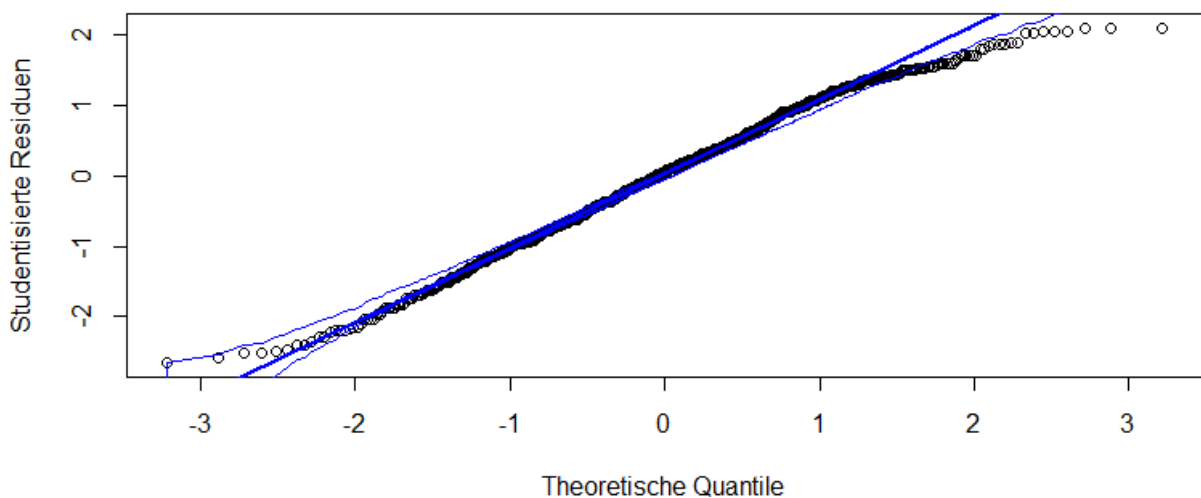


Abbildung 23: QQ-Plot zur Verteilung der studentisierten Residuen der messwiederholten einfaktoriellen ANOVA zum situationalen Interesse über die theoretischen ermittelten Quantile

Der Gruppenvergleich mittels ANOVA zum Messzeitpunkt 1 zeigt keine signifikanten Unterschiede auf (Levene-Test: $F(2, 245) = 1.72, p = 0.18$; ANOVA: $F(2, 245) = 1.83, p = 0.16, \eta_g^2 = 0.015$). Zwischen den einzelnen Gruppen wurden über den Interventionszeitraum hinweg keine signifikanten Unterschiede in ihrem situationalen Interesse festgestellt (Gruppe: $F(2, 245) = 1.44, p = 0.23, \eta_g^2 = 0.009$; Gruppe $\sim$ MZP: $F(4, 490) = 0.66, p = 0.62, \eta_g^2 = 0.001$). Die messwiederholte, einfaktorielle ANOVA mit Huynh-Feld-Korrektur zeigt einen signifikanten, marginalen Effekt ($p/[HF] < 0.001, \eta_g^2 < 0.02$) in Bezug auf die Messzeitpunkte. Der TukeyHSD-Test ergibt signifikante Unterschiede zwischen MZP1 und MZP2 ($p = 0.016$) sowie zwischen MZP1 und MZP3 ($p = 0.004$) auf. Dies zeigt, dass das situationale Interesse in den drei Vergleichsgruppen über die Messzeitpunkte hinweg gleich stark abgenommen hat (vgl. Abbildung 24). Tabelle 30 zeigt die deskriptiven Werte zum situationalen Interesse in den Vergleichsgruppen zu den drei Interventionszeitpunkten.

Die Post-Hoc-Poweranalyse zum gefundenen Effekt zwischen den Messzeitpunkten zeigt folgendes Ergebnis. Mit der kalkulierten Effektstärke von $f = 0.13$ und der Annahme eines Alpha-Fehlers mit einer Wahrscheinlichkeit von 0.01 (Cohen, 1988) ergibt sich für die Gesamtstichprobe von 248 Observationen eine Power von 1.0. Folglich gilt die Teststärke als ausreichend und ein Fehler zweiter Ordnung kann ausgeschlossen werden (Cohen, 1988; Eid et al., 2015). Bei einer Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.05 und einer Power von 0.95 ergibt sich für die ermittelte Effektstärke eine Gesamtstichprobengröße von 132 Observationen, um einen signifikanten Unterschied aufzeigen zu können.

Tabelle 30: Deskriptive Statistiken zum situationalen Interesse der einzelnen Experimentalgruppen und Messzeitpunkten (= MZP); AR = Augmented Reality, HI = Haptisch-Interaktiv, MW = Mittelwert, SD = Standardabweichung, min = erreichtes Minimum, max = erreichtes Maximum; Skala von 1-7

Gruppe	MZP	n	MW	SD	Median	min	max	Schiefe	Kurtosis
AR	1	77	4.69	1.19	4.8	1.6	7	-0.22	-0.2
AR	2	77	4.45	1.21	4.6	2	7	-0.04	-0.79
AR	3	77	4.27	1.4	4.2	1.6	7	0.09	-0.85
HI	1	77	4.94	1.37	5	1.6	7	-0.34	-0.62
HI	2	77	4.59	1.53	4.8	1	7	-0.49	-0.52
HI	3	77	4.5	1.5	4.6	1	7	-0.2	-0.94
IL	1	94	4.57	1.30	4.7	1.2	6.6	-0.55	-0.34
IL	2	94	4.16	1.37	4.4	1.4	7	-0.17	-0.68
IL	3	94	4.24	1.37	4.2	1	7	0.03	-0.52

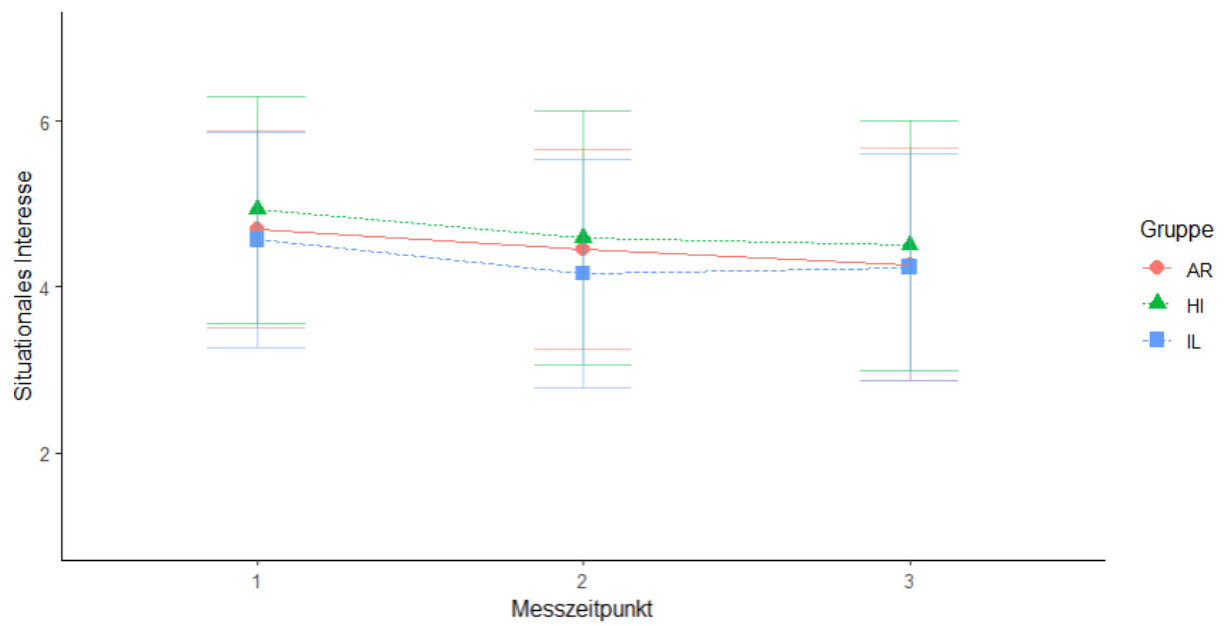


Abbildung 24: Entwicklung des situationalen Interesses zu den drei Interventionszeitpunkten im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau (Skala von 1-7). Ein hoher Zahlenwert bedeutet ein hohes Interesse an den Inhalten der jeweiligen Intervention.

Fachwissen

Die Analyse des Datensatzes auf Normalverteilung zeigt beim Konstrukt Fachwissen sowohl im Histogramm zur Antwortverteilung (s. Abbildung 25) als im QQ-Plot zur Verteilung der studentisierten Residuen der messwiederholten einfaktoriellen ANOVA (s. Abbildung 26) keine gravierenden Abweichungen von der Normalverteilung. Das signifikante Ergebnis des Levene-Tests ($F(5, 490) = 2.26, p = 0.048$) zeigt, dass für den vorliegenden Datensatz Varianzheterogenität anzunehmen ist. Für insgesamt 20 Fälle ($\triangleq 4\%$) wurde eine Cook'sche Distanz $< 4/496$ ermittelt. Daher wird der Umfang des Trimmens auf 4% festgelegt.

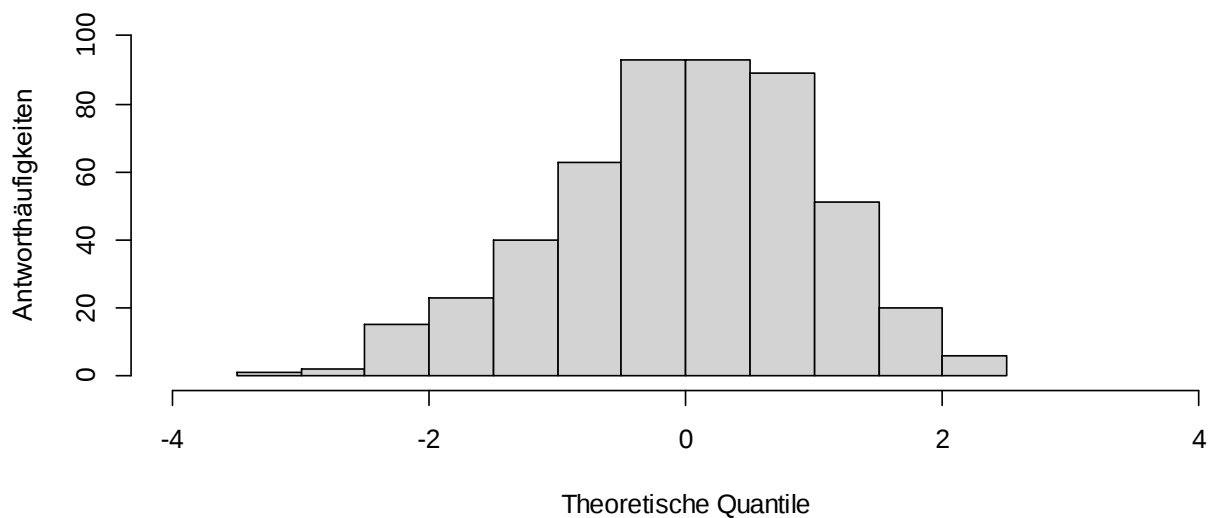


Abbildung 25: Histogramm zur Verteilung der Antworten der Gesamtstichprobe im Fachwissenstest über die theoretischen ermittelten Quantile

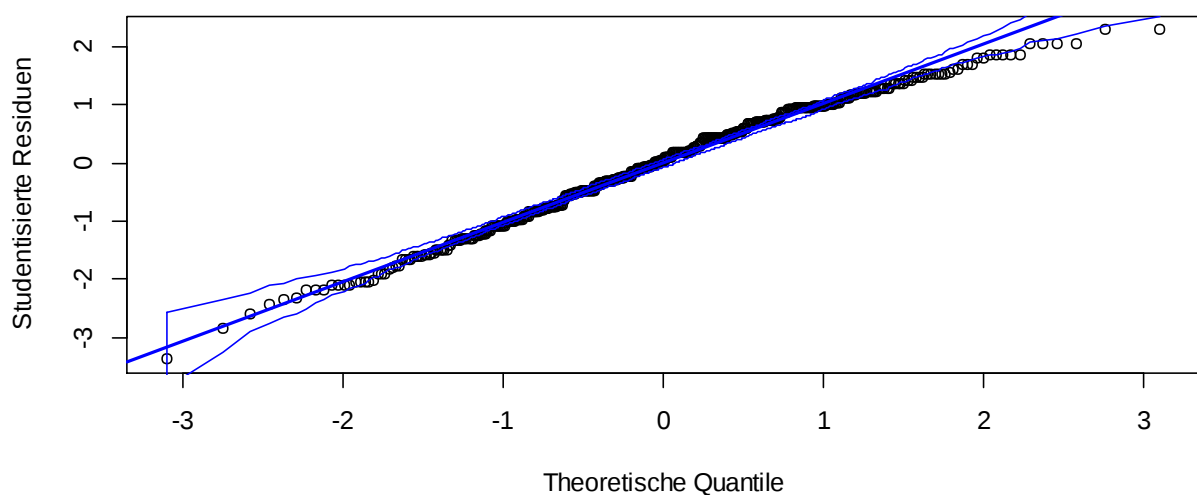


Abbildung 26: QQ-Plot zur Verteilung der studentisierten Residuen der messwiederholten einfaktoriellen ANOVA zum situationalen Interesse über die theoretischen ermittelten Quantile

Die einfaktorielle ANOVA zum Messzeitpunkt 0 ergibt einen signifikanten, kleinen Effekt in Bezug auf die Gruppenunterschiede (Levene-Test: $F(2, 245) = 14.77, p = 0.25$; ANOVA: $F(2, 245) = 4.07, p = 0.018, \eta_g^2 = 0.032$). Die Post-Hoc-Analyse mithilfe des TukeyHSD-Tests zeigt, dass dieser Effekt auf dem signifikanten Unterschied Gruppe IL zu den beiden anderen Gruppen AR und HI zustande kommt (IL-AR: $p = 0.03$; IL-HI: $p = 0.05$; HI-AR: $p = 0.98$). Der Welch-James-ADF-Test zum Faktor Fachwissen zeigt einen signifikanten, großen Effekt zwischen den Messzeitpunkten ($WJ(1, 400.3) = 135.236, p < 0.001, \hat{\delta}_j^{(R)} = 1.06$). Auch der Gruppenvergleich mittels der Daten zu beiden Messzeitpunkten zeigt ein signifikantes Ergebnis ($WJ(2, 281.3) = 6.604, p < 0.01$). Die Post-Hoc-Analyse über paarweise Kontraste der WJ-Statistiken zeigt, dass der Gruppenunterschied auf einen signifikanten, kleinen Effekt zwischen den Gruppen IL und HI zurückzuführen ist (IL-HI: $WJ(1, 283.1) = 10.382, p = 0.004, \hat{\delta}_j^{(R)} = 0.36$, AR-HI: $WJ(1, 282.0) = 1.626, p = 0.20, \hat{\delta}_j^{(R)} = 0.15$; IL-AR: $WJ(1, 283.3) = 3.283, p = 0.14, \hat{\delta}_j^{(R)} = 0.20$). Die Überprüfung auf Interaktionseffekte weist keinen signifikanten Effekt auf ($WJ(2, 281) = 1.777, p = 0.17$). Tabelle 31 zeigt die deskriptiven Statistiken zum Fachwissen in den Vergleichsgruppen zum Pre- (= MZP0) und Postzeitpunkt (= MZP4). Abbildung 27 zeigt die Entwicklung des Fachwissens in den Vergleichsgruppen von Pre zu Post. Zwischen den Messzeitpunkten MZP0 und MZP4 liegt eine starke Korrelation nach Pearson ($r = 0.63$) vor (Cohen, 1988).

Die Post-Hoc-Poweranalyse zum gefundenen, signifikanten Effekt zeigt folgendes Ergebnis. Mit der kalkulierten Effektstärke von $f = 0.53$ und der Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.01 (Cohen, 1988) ergibt sich für die Gesamtstichprobe von 248 Observationen eine Power von 1. Folglich gilt die Teststärke als ausreichend und ein Fehler zweiter Ordnung kann ausgeschlossen werden (Cohen, 1988; Eid et al., 2015). Bei einer Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.05 und einer Power von 0.95 ergibt sich für die ermittelte Effektstärke eine Gesamtstichprobengröße von 15 Observationen, um einen signifikanten Unterschied aufzeigen zu können.

Tabelle 31: Deskriptive Statistiken zum Fachwissen der einzelnen Experimentalgruppen und Messzeitpunkten (= MZP); AR = Augmented Reality, HI = Haptisch-Interaktiv, MW = Mittelwert, SD = Standardabweichung, min = erreichtes Minimum, max = erreichtes Maximum; mögliches Maximum = 21

Gruppe	MZP	n	MW	SD	Median	min	max	Schiefe	Kurtosis
AR	0	77	9.86	3.64	9	4	18	0.62	-0.36
AR	4	77	15.1	4.11	16	6	21	-0.47	-0.89
HI	0	77	9.96	4.04	10	2	19	0.19	-0.85
HI	4	77	13.58	4.67	14	4	21	-0.24	-0.98
IL	0	94	11.31	3.61	12	3	18	-0.37	-0.51
IL	4	94	15.24	3.88	16	2	20	-1.15	1.09

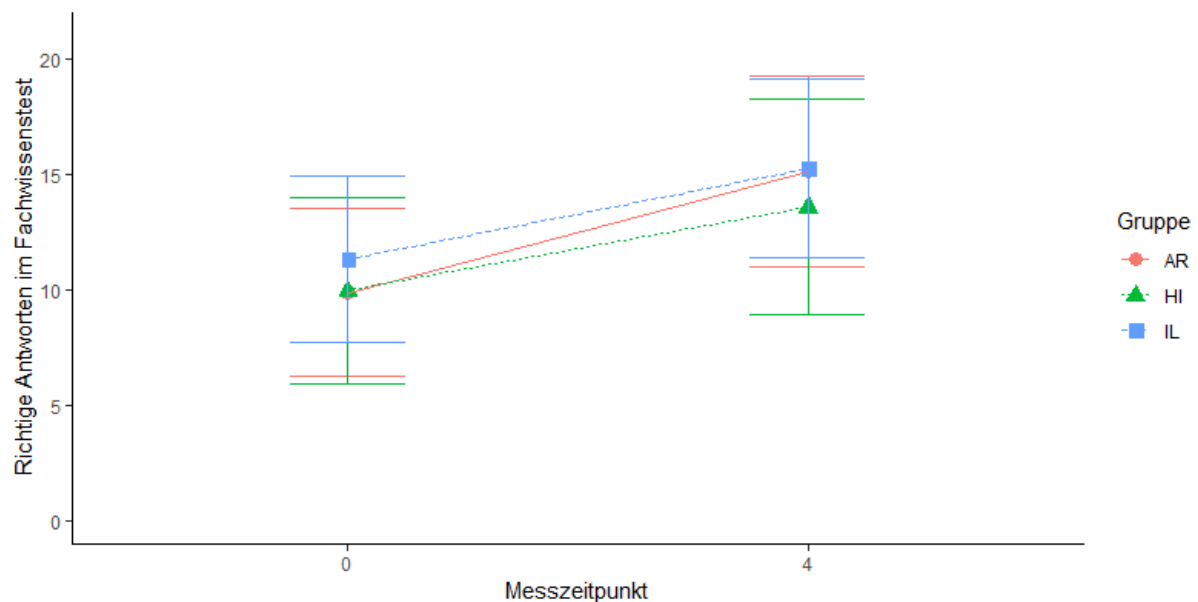


Abbildung 27: Entwicklung des Fachwissens zu den zu den Messzeitpunkten 0 (=pre) und 4 (=post) im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau.

Modellverständnis

Die Analyse des Datensatzes auf Normalverteilung zeigt beim Konstrukt Modellverständnis eine linksschiefe Verteilung im Histogramm (s. Abbildung 28), die jedoch als nicht problematisch anzusehen ist. Die Verteilung der studentisierten Residuen der messwiederholten einfaktoriellen ANOVA (s. Abbildung 29) zeigt im QQ-Plot ebenfalls Abweichungen von der Normalverteilung. Für 28 ($\cong 6\%$) Fälle wurde eine Cook'sche Distanz $< 4/496$ ermittelt. Daher wird der Datensatz für den messwiederholten Welch-James-ADF-Test um 6% getrimmt. Aufgrund des signifikanten Ergebnisses des Levene-Tests ($F(5, 490) = 10.37, p < 0.001$) ist für den vorliegenden Datensatz Varianzheterogenität anzunehmen.



Abbildung 28: Histogramm zur Verteilung der Antworten der Gesamtstichprobe zum Konstrukt Modellverständnis über die theoretischen ermittelten Quantile

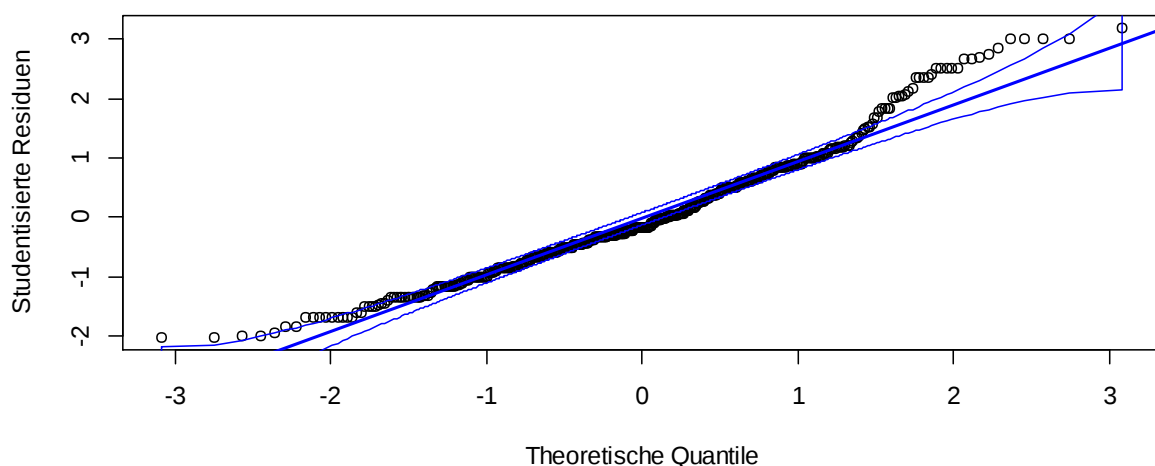


Abbildung 29: QQ-Plot zur Verteilung der studentisierten Residuen der messwiederholten einfaktoriellen ANOVA zum situationalen Interesse über die theoretischen ermittelten Quantile

Der Gruppenvergleich mittels einfaktorieller ANOVA zum Messzeitpunkt 0 zeigt keine signifikanten Unterschiede auf (Levene-Test: $F(2, 245) = 0.08, p = 0.92$; ANOVA: $F(2, 245) = 0.73, p = 0.48, \eta_g^2 = 0.006$). Der messwiederholte Welch-James-ADF-Test erbringt für den Gruppenvergleich ein signifikantes Ergebnis ($WJ(2, 230.4) = 3.540, p = 0.03$). Die Post-Hoc-Analyse mittels paarweiser Kontraste zeigt, dass dieser Effekt auf dem Unterschied zwischen den Gruppen AR und HI mit einer mittleren Effektstärke beruht (AR-HI: $WJ(1, 211.8) = 6.197, p = 0.04, \hat{\delta}_j^{(R)} = 0.59$; IL-AR: $WJ(1, 223.3) = 0.333, p = 0.56, \hat{\delta}_j^{(R)} = 0.13$; IL-HI: $WJ(1, 252.5) = 3.889, p = 0.10, \hat{\delta}_j^{(R)} = 0.44$). Die Prüfung der Lernwirksamkeit in Bezug auf das Modellverständnis im Pre-Post-Vergleich zeigt, dass sich die Gesamtstichprobe mit einem sehr großen Effekt signifikant verbessert hat (MZP0 – MZP4: $WJ(1, 339.1) = 21.17, p < 0.001, \hat{\delta}_j^{(R)} = 1.30$). In der Gruppe AR hat sich das Modellverständnis im Vergleich zur Gruppe HI mit einem großen Effekt signifikant stärker entwickelt, während es sich in der Gruppe IL mit einem mittleren Effekt im Vergleich zur Gruppe HI signifikant verbessert hat (Interaktion Gruppe ~ MZP: $WJ(2, 230.4) = 6.383, p = 0.002$; AR-HI: $WJ(1, 211.8) = 11.664, p = 0.002, \hat{\delta}_j^{(R)} = 0.81$; IL-AR: $WJ(1, 223.3) = 0.980, p = 0.32, \hat{\delta}_j^{(R)} = 0.23$; IL-HI: $WJ(1, 252.5) = 6.217, p = 0.03, \hat{\delta}_j^{(R)} = 0.56$). Tabelle 32 zeigt die deskriptiven Statistiken zum Modellverständnis in den Vergleichsgruppen zum Pre- (= MZP0) und Postzeitpunkt (= MZP4). Abbildung 30 zeigt die Entwicklung des Modellverständnisses in den Vergleichsgruppen von Pre zu Post.

Die Post-Hoc-Poweranalyse zum Effekt zwischen den Messzeitpunkten zeigt folgende Ergebnisse. Mit der kalkulierten Effektstärke von $f = 0.65$ und der Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.01 (Cohen, 1988) ergibt sich für die Gesamtstichprobe von 248 Observationen eine Power von 1. Folglich kann das Vorliegen eines Fehlers zweiter Ordnung ausgeschlossen und die Teststärke als ausreichend angesehen werden (Cohen, 1988; Eid et al., 2015). Bei einer Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.05 und Power von 0.95 ergibt sich für die ermittelte Effektstärke eine Gesamtstichprobengröße von 15 Observationen.

Die Ergebnisse der Post-Hoc-Poweranalyse zum Interaktionseffekt gestalten wie folgt. Mit der kalkulierten Effektstärke von $f = 0.14$ und der Annahme eines Alpha-Fehlers mit einer Wahrscheinlichkeit von 0.01 (Cohen, 1988) ergibt sich für die Gesamtstichprobe von 248 Observationen eine Power von 0.87. Folglich gilt die Teststärke als ausreichend und ein Fehler zweiter Ordnung kann mit einer Wahrscheinlichkeit von 13% ausgeschlossen werden (Cohen, 1988; Eid et al., 2015). Bei einer Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.05 und einer Power von 0.95 ergibt sich für die ermittelte Effektstärke eine Gesamtstichprobengröße von 237 Observationen.

Tabelle 32: Deskriptive Statistiken zum Modellverständnis der einzelnen Experimentalgruppen und Messzeitpunkten (= MZP); AR = Augmented Reality, HI = Haptisch-Interaktiv, MW = Mittelwert, SD = Standardabweichung, min = erreichtes Minimum, max = erreichtes Maximum; Skala von 1-5

Gruppe	MZP	n	MW	SD	Median	min	max	Schiefe	Kurtosis
AR	0	77	2.16	0.52	2.12	1.12	3.75	0.52	0.3
AR	4	77	2.75	0.93	2.62	1.25	5	0.59	-0.34
HI	0	77	2.23	0.54	2.12	1.12	3.75	0.27	-0.34
HI	4	77	2.25	0.8	2.12	1	4.5	0.7	0.01
IL	0	94	2.2	0.61	2.12	1	4.25	0.59	0.58
IL	4	94	2.62	0.95	2.5	1.12	5	0.68	-0.29

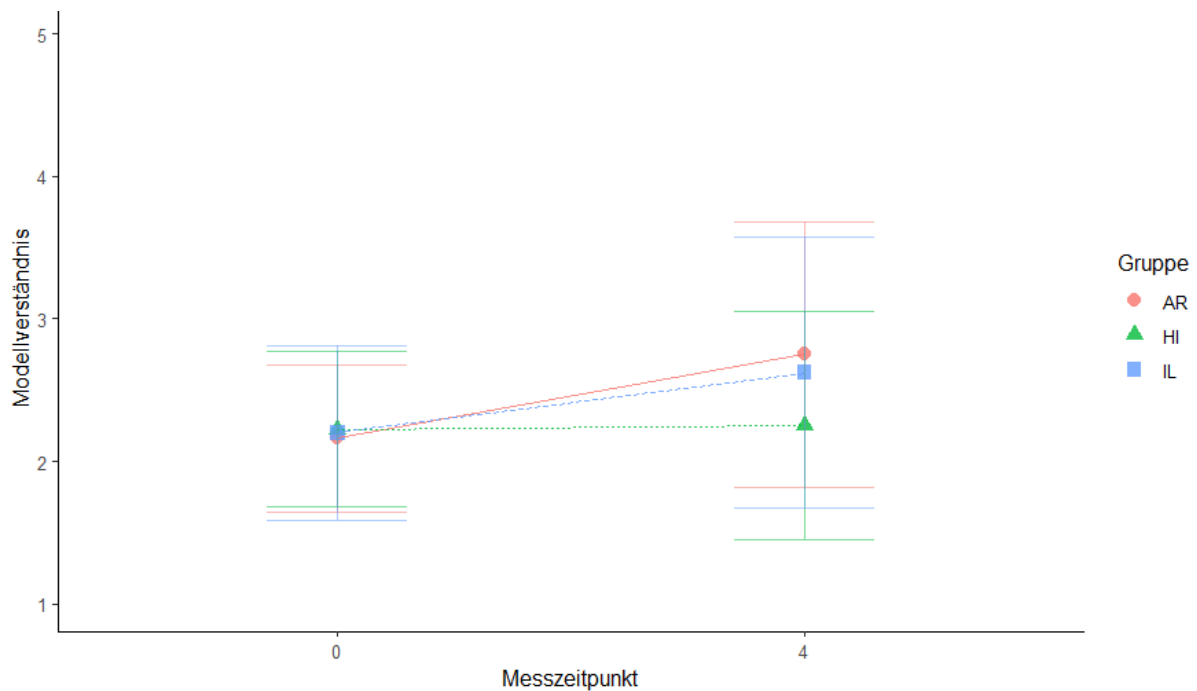


Abbildung 30: Entwicklung des Modellverständnisses zu den Messzeitpunkten 0 (=pre) und 4 (=post) im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau. Ein hoher Wert bedeutet ein Verständnis über Modelle, das dem wissenschaftlichen Konsens entspricht.

Der Einstichproben-t-Tests (Abbildung 31) zeigt mit einem mittleren Effekt ein signifikant niedrigeres Niveau im Modellverständnis der Gesamtstichprobe zum MZP4 (= Post) der vorliegenden Studie als in der Normierungsstudie ($t(247) = -4.33$, $p < 0.001$, $d = -0.28$, $n = 248$). Die Ergebnisse der Post-Hoc-Poweranalyse zum Interaktionseffekt gestalten wie folgt. Mit der kalkulierten Effektstärke von $d = 0.275$ und der Annahme eines Alpha-Fehlers mit einer Wahrscheinlichkeit von 0.05 (Cohen, 1988) ergibt sich für die Gesamtstichprobe von 248 Observationen eine Power von 0.99. Folglich gilt die Teststärke als ausreichend und ein Fehler zweiter Ordnung kann zu einer Wahrscheinlichkeit von 1% ausgeschlossen werden (Cohen, 1988; Eid et al., 2015). Bei einer Annahme einer Wahrscheinlichkeit eines Alpha-Fehlers von 0.05 und einer Power von 0.95 ergibt sich für die ermittelte Effektstärke eine Gesamtstichprobengröße von 174 Observationen.

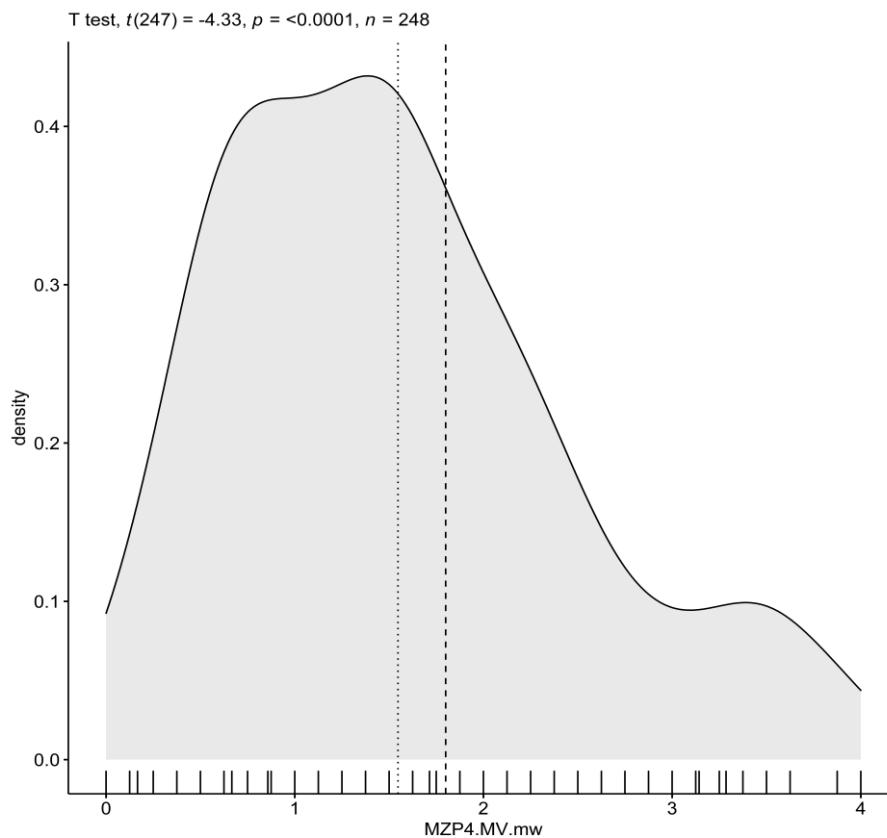


Abbildung 31: Einstichproben-t-Test zum Abgleich des Modellverständnisses der Gesamtstichprobe zum Post-Zeitpunkt (gepunktete Linie) mit der Normierung (gestrichelte Linie). Das Modellverständnis in der vorliegenden Intervention fällt mit einem kleinen Effekt signifikant geringer aus.

8.3.5 Diskussion zur Interventionsstudie

Die Diskussion der Interventionsstudie ist in mehrere Abschnitte unterteilt. Zunächst werden die Ergebnisse hypothesengeleitet reflektiert und interpretiert. Anschließend erfolgt die kritische Einordnung der Ergebnisse zu den theoretischen Hintergründen und dem aktuellen Forschungsstand. Nach der Bewertung der Konzeption werden die methodischen Ansätze der vorliegenden Interventionsstudie kritisiert und Optimierungsvorschläge aufgezeigt.

Inhaltliche Diskussion der Ergebnisse und Hypothesen

Die Hypothesen H3, H3.1 und H3.2 müssen aufgrund der vorliegenden Ergebnisse der inferenzstatistischen Untersuchungen zum **situationalen Interesse** als falsifiziert angesehen werden.

H3 Die Entwicklung des situationalen Interesses der einzelnen Gruppen (AR, HI & IL) unterscheidet sich über den Interventionszeitraum hinweg.

H3.1 Die Lernenden zeigen während der Intervention ein höheres situationales Interesse im Umgang mit den AR-Modellen als im Umgang mit den haptisch-interaktiven Modellen.

H3.2 Die Lernenden zeigen während der Intervention ein höheres situationales Interesse im Umgang mit den AR-Modellen als im Umgang mit den illustrativen Modellen.

Im Hinblick auf das situationale Interesse war zunächst für alle drei Vergleichsgruppen ein Neuheitseffekt zu erwarten (Theyßen, 2014). Dieser erwies sich in der vorliegenden Studie als marginal. Das erwartete erhöhte situationale Interesse über den Neuheitseffekt hinweg (Sofianidis, 2022) aufgrund eines gesteigerten Kompetenz- und Autonomieerlebens der Versuchspersonen im Umgang mit den digitalen (AR) Modellen (Hillmayr et al. 2020) ist jedoch nicht eingetreten. Für das gleich starke Auftreten des Neuheitseffekts in allen drei Vergleichsgruppen ergeben sich verschiedene, mögliche Ursachen. Zunächst ist davon auszugehen, dass alle Modelle der drei Modelltypen für die Schüler*innen neu waren. Daher ist ein Neuheitseffekt für alle drei Gruppen in gleicher Höhe zu erwarten. Eine alternative Ursache für die Ergebnisse dieser Studie könnte darin liegen, dass die Einflüsse der Gesamtkonzeption oder des eigenständigen Experimentierens überwiegen. So konnten die Schüler*innen in allen Gruppen über den gesamten Interventionszeitraum hinweg eigenständig Entscheidungen über ihren Lernprozess treffen und hatten somit ein vergleichbares Autonomieerleben (Ryan & Deci, 2002). Auch geben die gleich hohen Werte für System Usability und ECL in den Vergleichsgruppen Anhaltspunkte darüber, dass die Versuchspersonen aller drei Gruppen ein vergleichbares Kompetenzerleben hatten. Folglich bleiben die damit verbundenen Steigerungseffekte auf das situationale Interesse aus.

Ebenso ist anzunehmen, dass die Versuchspersonen den digitalen Modellen keinen höheren „intrinsic value“ im Sinne der Expectancy-Value Theory (Wigfield & Eccles, 2000) zugeschrieben haben. Eine mögliche Ursache hierfür ist, dass die Schülerinnen und Schüler es gewohnt sind, Tablets im Kontext von Schule und Unterricht regelmäßig zum Arbeiten zu nutzen. Sie erwarten daher keinen Unterhaltungswert durch den Einsatz von Tablets in der Schule und zeigen auch kein erhöhtes situationales Interesse am Lerngegenstand.

Die Hypothesen H4, H4.1 und H4.2 zum Konstrukt **Fachwissen** sind anhand der vorliegenden Ergebnisse der inferenzstatistischen Untersuchungen ebenfalls als falsifiziert anzusehen.

H4 Das Fachwissen der verschiedenen Gruppen (AR, HI, IL) entwickelt sich über die einzelnen Messzeitpunkte unterschiedlich stark.

*H4.1 Das Fachwissen der Schüler*innen in der Gruppe AR nimmt mehr zu als in der Gruppe HI.*

*H4.2 Das Fachwissen der Schüler*innen in der Gruppe AR nimmt mehr zu als das der IL.*

Das Ausbleiben eines Interaktionseffekts beim Faktor Fachwissen zugunsten der AR-Gruppe widerspricht den Ergebnissen der Metaanalyse von Höffler (2008), dass Animationen vor allem dann größere Lernerfolge erzielen, wenn deklaratives statt prozedural-problemlösendem Wissen geprüft wird. Die Modelle wurden zuvor zunächst auf ihre Qualität zum Erreichen der Lernziele von Expert*innen bewertet (Studie II). Zusätzlich wurden die Modelle hinsichtlich ihrer Eignung als Lernmaterial für die Zielgruppe durch die Überprüfung ihrer Usability und den Subskalen des CLs validiert. Daher ergeben sich aus lerntheoretischer Sicht keine empirisch belegbaren Gründe für einen Vorteil eines Modelltyps gegenüber den anderen. Die Vielzahl an empirischen Studien und Metaanalysen, die Vorteile von digitalen Animationen und AR zur Vermittlung komplexer Konzepte aufzeigen, weisen keine explizite Prüfung der Modellqualität auf (vgl. Abschnitt 3). Zudem unterscheiden sich die Vergleichsgruppen zumeist in weiteren Merkmalen als dem Modelltypus. Vor dem Hintergrund der Ergebnisse der vorliegenden Studie stellt sich die Frage, ob die Unterschiede in der Lernwirksamkeit der Modelle in der aktuellen Studienlage auf den Modelltyp an sich zurückgeführt werden können. Ross (1994) konstatiert, dass selbst randomisierte Medienvergleichsstudien, bei denen in den Vergleichsgruppen dieselbe Unterrichtsmethode verwendet wird, keine guten Untersuchungen zur Lernwirksamkeit von Bildungstechnologien darstellen. In solchen Studien sei das Ergebnis im Voraus klar (Buchner & Kerres, 2023; Ross, 1994), was zu einer Zunahme von Studien beiträgt, die keine signifikanten Unterschiede berichten (Reeves & Oh, 2017). Der Unterschied der Gruppen im *Fachwissenstest* zum Pre-Zeitpunkt ist kritisch zu diskutieren. Die Gruppe IL weist mit einem kleinen Effekt ein signifikant höheres *Fachwissen* auf als die beiden anderen

Experimentalgruppen. Da sich die Gruppen in allen restlichen abhängigen und Kontrollvariablen nicht unterscheiden und sich keine signifikanten Interaktionseffekte zum *Fachwissen* gezeigt haben, wird ein Profitieren der IL-Gruppe aufgrund des Matthäus-Effekts (Hattie et al., 2013) zurückgewiesen.

Die Hypothesen H5 und H5.1 können anhand der vorliegenden Ergebnisse der inferenzstatistischen Untersuchungen zum Konstrukt **Modellverständnis** verifiziert werden. Es konnte jedoch kein signifikanter Unterschied zwischen den Gruppen AR und IL festgestellt werden, woraufhin H5.2 falsifiziert wird.

H5 Das Modellverständnis der verschiedenen Gruppen (AR, HI, IL) entwickelt sich über die einzelnen Messzeitpunkte unterschiedlich stark.

*H5.1 Das Modellverständnis der Schüler*innen in der Gruppe AR nimmt mehr zu als in der Gruppe HI.*

*H5.2 Das Modellverständnis der Schüler*innen in der Gruppe AR nimmt mehr zu als in der Gruppe IL.*

Die vorteilhafte Entwicklung epistemologischer Fähigkeiten durch Animationen (Urhahne et al., 2009) bzw. AR (Bevan, 2023) im Vergleich zu den illustrativen Modellen konnte in dieser Studie nicht repliziert werden. Eine deutlichere Trennung zwischen Modell und Realität aufgrund der Animationen (Saborowski, 2000) kann ebenfalls nicht angenommen werden, da sich die Lernzuwächse zwischen den Gruppen AR und IL nicht signifikant unterscheiden. Die Fähigkeit zur Unterscheidung dieser beiden Ebenen könnte jedoch eine mögliche Ursache für die signifikant schlechtere Entwicklung in der Gruppe HI sein. In den abschließenden Besprechungen im Plenum kam es in der Gruppe HI vermehrt zu Verwechslungen von Experiment und Modell. Auch die Lehrkräfte gaben an, dass sie den Eindruck hatten, ihre Schüler*innen haben während des Arbeitens mit den Materialien die haptisch-interaktiven Modelle als Experimente bezeichnet. Für die eingesetzte Skala wird das Modellverständnis jedoch genau als diese Unterscheidung zwischen Modellen und Teilchen einerseits und Experimenten und Realität andererseits operationalisiert (Leisner 2005a). Dies liefert einen Erklärungsansatz, für den signifikant niedrigeren Zuwachs an Modellverständnis in der Gruppe HI im Vergleich zu den beiden anderen Gruppen. Um diese Frage zu klären, bedarf es jedoch weiterer qualitativer Forschung. Für die gefunden Effekte zum Modellverständnis besteht die Möglichkeit einer Beeinflussung der Messung aufgrund einer Regression zur Mitte (Barnett et al., 2005) von MZP0 zu MZP4. Da die Ergebnisse in den Untersuchungen von Mikelskis-Seifert (2002) und Leisner (2005a) ähnliche Tendenzen bei nahezu gleichen Ausgangspunkten zeigen, wird die Annahme einer Regression zur Mitte jedoch abgelehnt.

Diskussion der Konzeption

Das Modellverständnis nimmt im Chemieunterricht eine eher untergeordnete Rolle ein. In den Bildungsstandards und der fachdidaktischen Literatur wird zwar beschrieben, dass Schüler*innen die Fähigkeit entwickeln sollen, chemische Modelle zu verstehen, zu nutzen und zu bewerten (Eilks & Bindernagel, 2008; KMK, 2005). Eine gezielte Einordnung der Aussagekraft von chemischen Modellen im Sinne einer Trennung zwischen Realität und Modellvorstellung wird in den Bildungsstandards dennoch nicht vorgenommen. Auch gibt es nur wenige didaktische Konzepte, die diese Trennung vornehmen (Haas & Marohn, 2022; Lang, 2023). In der Studie von Lang (2023) zeigte sich, dass das explizite Thematisieren der Dimension Kenntnisse über Modell (=Modellverständnis) nicht zu einer Verbesserung der epistemologischen Fähigkeiten der Schüler*innen führte. Auch Halloun (2001) betont, dass epistemologische Schüler*innen -Vorstellungen sich nicht durch traditionelle Unterrichtsmethoden beeinflussen lassen. Die vorliegende Studie, die auf dem Conceptual-Change-Ansatz basiert, weist hingegen einen solchen signifikanten Effekt mit einer sehr großen Effektstärke nach. Laut TIMS-Studie III (Baumert et al., 2013) findet ein vollständiger Conceptual-Change jedoch häufig nicht statt. Eine Rekonstruktion der naiven Präkonzepte zu einem tiefen Verständnis naturwissenschaftlicher Themen gelingt selten (ebd.). Sogar bis hin zum Abitur weisen Schüler*innen Schwierigkeiten beim Bearbeiten von Aufgaben zur Rekonstruktion von alternativen Präkonzepten auf (ebd.). Folglich bedarf es weiterer Forschung, um qualitative Aussagen über die Entwicklung der Vorstellungen der Lernenden nach dem Arbeiten mit den einzelnen Modelltypen vornehmen zu können.

Die Entwicklung des Modellverständnisses ist in der vorliegenden Studie mit einem kleinen Effekt signifikant geringer als in der Untersuchung von Leisner (2005a). Gemessen an der aufgebrachten Time-On-Task ist die hier vorgestellte Intervention jedoch kaum als schlechter zu bewerten. In der Studie von Leisner (2005a) dauerte die Intervention acht bis zehn Wochen, wohingegen sich der Interventionszeitraum in der vorliegenden Studie auf lediglich 1,5 bis zwei Wochen beläuft.

Die von Lang (2023) durchgeführte Studie konnte eine Steigerung des Fachwissens und der Modellierungskompetenz nachweisen. Die Lernerfolge wurden auf das Anfertigen von Schüler*innen -Zeichnungen auf der Teilchenebene zur Deutung von Experimenten zurückgeführt (Lang, 2023). Diese Ergebnisse decken sich mit den in Studie III gefundenen, signifikanten Effekten. Der über alle drei Vergleichsgruppen gemittelte Lernzuwachs beträgt 20% von der möglichen Maximalpunktzahl im Fachwissenstest. Studien zu ähnlichen inhaltlichen Themen mit vergleichbarem Design zeigen eine Entwicklung von Pre zu Post von 7% (Peeters et al.,

2023) und 17% (Probst et al., 2022) von der möglichen Maximalpunktzahl. Folglich kann der in Studie III beobachtete Lernzuwachs als plausibel erachtet werden.

Die Gesamtstichprobe der Interventionsstudie hat die *System Usability* im Mittel über alle drei Messzeitpunkte mit 77 von 100 Punkten bewertet. Der Wert liegt damit im „besseren Bereich“ (Bangor et al., 2008). Da in der Pilotierungsstudie mit 83 Punkten ein vergleichbarer Wert ebenfalls im „besseren Bereich“ erzielt wurde, sind die Ergebnisse aus beiden Erhebungen als plausibel anzusehen. Das *situationale Interesse* wurde von der Gesamtstichprobe der Interventionsstudie im Mittel über alle drei Messzeitpunkte hinweg mit $MW = 4,49$ (Skala von 1-7) bewertet und plausibilisiert somit den Wert aus der Pilotierung der Materialien ($MW = 4,66$). Aufgrund dieser zufriedenstellenden Werte für die Handhabung der Arbeitsmaterialien lassen sich die Materialien als lernförderlich einstufen.

Die Durchschnittswerte der Subskalen des *Cognitive Loads* über alle Gruppen und Messzeitpunkte entsprechen dem Ideal einer gewünschten Verteilung ($ECL < ICL < GCL$). Der niedrige ECL-Wert zeigt an, dass der geringste Teil der kognitiven Kapazitäten für die Verarbeitung der Lernmaterialien aufgebracht wurde. Dies lässt den Schluss zu, dass die Materialien auf Basis eines fundierten Designs entwickelt wurden. Weiterhin spricht die Verteilung der Werte des *Cognitive Loads* auf die Subfacetten für eine besonders lernförderliche Lernumgebung mit ausreichend freier Kapazität, die zum Lernen aufgebracht werden kann (Paas et al., 2016; Moreno & Park, 2010).

Zusammenfassend lässt sich konstatieren, dass das vorliegende Unterrichtskonzept positive Ausprägungen der kognitiv-affektiven Variablen bei den Lernenden hervorruft. Die Ergebnisse auf der kognitiven Ebene sind überzeugend und weisen auf aktives und förderliches Lernen hin. Auch auf der affektiven Ebene zeigen die Lernenden zufriedenstellende Einschätzungen, sodass lernhinderliche Probleme aufgrund der Konzeption ausgeschlossen werden können. Dies unterstreicht die Aussagekraft der Interventionsstudie zum Anteil der Modelltypen an den gefundenen Effekten beim Gruppenvergleich.

Methodenkritik

Die vorliegende Forschungsarbeit intendiert, zur Überwindung der in der Fachliteratur vielfach konstatierten Diskrepanz zwischen Theorie und Praxis beizutragen. Mittels der Implementierung eines praxisorientierten Studiendesigns wird angestrebt, eine empirisch fundierte Verknüpfung zwischen Forschungsergebnissen und deren Umsetzung in der Bildungspraxis zu etablieren. Hierdurch soll ein Beitrag zur Integration wissenschaftlicher Erkenntnisse und bildungspraktischer Implikationen in die naturwissenschaftliche Bildungslandschaft geleistet werden. Die ausgeprägte Praxisorientierung des Forschungsansatzes impliziert jedoch potenzielle, methodologische Limitationen, welche die Generalisierbarkeit und externe Validität der gewonnenen Erkenntnisse möglicherweise einschränken. Zu Studie III ergeben sich sowohl Kritikpunkte zu den Methoden der Datenerhebung und der Datenauswertung als auch zur Interpretierbarkeit der Daten. Zunächst liegt ein quasi-experimentelles Untersuchungsdesign vor. Es wurde keine Randomisierung vorgenommen, sondern eine randomisierte Zuweisung zu den Vergleichsgruppen auf Klassenebene. Da die Teilnehmenden den Gruppen klassenweise zugewiesen wurden, besteht eine größere Gefahr, dass die Ergebnisse durch Störvariablen beeinflusst werden. Dies kann die interne Validität der Untersuchung beeinträchtigen (Cook et al., 1979). Zudem handelt es sich bei den drei Vergleichsgruppen jeweils um Experimentalgruppen, wodurch eine separate Kontrollgruppe fehlt (Bennett et al., 2007). Dadurch kann das Auftreten des „practice effects“ durch wiederholte Testung nicht ausgeschlossen werden (Calamia et al., 2012). Rost (2000) legitimiert den Verzicht auf eine Kontrollgruppe, wenn die Kontrollgruppe zu einem realitätsfremden und künstlichen Vergleich führt und demzufolge die Aussagekraft der Erkenntnisse mangels externer Validität eingeschränkt wird. Eine weitere Einschränkung der Studie III besteht darin, dass die Stichprobenziehung der Schülerinnen und Schüler auf Gelegenheitsstichproben statt Zufallsstichproben basiert. Zusätzlich ist hier die Freiwilligkeit zur Teilnahme an der Studie gegeben. Die Stichprobe ist somit nicht repräsentativ für die Gesamtpopulation, da die Teilnehmenden nicht zufällig ausgewählt wurden. Daher können die Ergebnisse der Untersuchung nicht ohne Weiteres auf die Gesamtpopulation übertragen werden (Döring & Bortz, 2016). Durch das beschriebene Zustandekommen der Stichproben wird die Generalisierbarkeit der Ergebnisse eingeschränkt (ebd.). Darüber hinaus ist die globale Generalisierbarkeit aufgrund der gezielten Ausrichtung der Lernumgebung auf die achte Klassenstufe der weiterführenden Schulen in Rheinland-Pfalz eingeschränkt. Die Stichprobengröße zu allen erhobenen Konstrukten kann auf Basis der Ergebnisse der Power-Analysen als ausreichend groß und angemessen ökonomisch betrachtet werden (Kraft, 2020). Zudem ist für den Stichprobenumfang anzumerken, dass die Untersuchung starken Einschränkungen durch die pandemiebedingten Schulschließungen unterlag. Dies erklärt auch den hohen Dropout in Studie III, da hier zu Beginn der Pandemie bei einer Covid-

Positivtestung auch die Sitznachbar*innen in Quarantäne verbleiben mussten und somit als Versuchspersonen wegfielen. Weiterhin erfolgte die Durchführung der Erhebungen in unterschiedlichen Settings, was zu variierenden externen Einflüssen führte. Diese Einflüsse resultierten in unvermeidbaren Variationen der Untersuchungsbedingungen, wie sie typischerweise in Feldstudien auftreten (Döring & Bortz, 2016). Folglich weichen in Studie III die Testbedingungen mit variierenden Klassenräumen von einem Laborsetting ab, auch wenn jede Intervention vom selben Testleiter durchgeführt wurde. Es kann nicht ausgeschlossen werden, dass die unterschiedlichen Lernorte einen Einfluss auf die Versuchspersonen hatten. Da die Untersuchungen jedoch stets in den gewohnten Chemiesälen der Lernenden stattfanden, wird dieser Einfluss als geringfügig eingeschätzt. Weiterhin war es aufgrund der praktischen Erfordernisse des Schulalltags nicht möglich, einen einheitlichen Zeitplan umzusetzen. In den meisten Fällen standen zwei Doppelstunden pro Woche zur Verfügung, während der Chemieunterricht an einer Schule in eine Doppelstunde und zwei Einzelstunden pro Woche aufgeteilt war. Da an dieser Schule jedoch drei Klassen beprobt wurden, die sich auf alle drei Vergleichsgruppen verteilt haben und die Gesamtzeitspanne der Untersuchung nicht variierte, wird dieser Störeinfluss auf die Gruppenvergleiche ebenfalls als gering eingeschätzt. Zwar konnte sichergestellt werden, dass während der Untersuchungsphase kein weiterer Chemieunterricht stattfand, dennoch kann nicht ausgeschlossen werden, dass bei den Lernenden zwischen den fünf bzw. sechs Untersuchungsterminen Lern- oder Vergessenseffekte in unterschiedlichem Ausmaß aufgetreten sind. Neben den genannten Limitationen bietet dieser praxisnahe Forschungsansatz jedoch den Vorteil, dass die Interventionen an die konkreten, schulischen Bedürfnisse angepasst sind und die Ergebnisse somit besser auf die Unterrichtspraxis übertragbar sind (Bernholt & Parchmann, 2021; Eilks & Ralle, 2002; Ralle & Di Fuccia, 2014). Zugunsten eines fairen Vergleichs wurden passgenaue Lernumgebungen für alle drei Vergleichsgruppen auf Basis valider Kriterien konzipiert (Theyßen, 2014). Dieses umfassendere Vorgehen bietet den Vorteil, dass sich die Vergleichsgruppen lediglich in der Wahl des Mediums unterscheiden und berücksichtigt damit die Kritik von Bétrancourt und Tversky (2000), Morrison et al. (2000) sowie von Buchner und Kerres (2023) gegenüber unkontrollierten und damit untauglichen Medienvergleichsstudien. In Bezug auf die Vergleichbarkeit der Gruppen hätte die Modellqualität der Modelltypen nach der Überarbeitung anhand der Expert*innenkritik noch einmal überprüft werden müssen. Die IL-Modelle haben sich vor der Überarbeitung mit einem kleinen Effekt signifikant von der Qualität der anderen Gruppen unterschieden. Retrospektiv erweist sich eine weitere Qualitätsprüfung nach der Überarbeitung aus zwei Gründen als redundant. Zum einen boten die Expert*innenkommentare ausreichend präzise Anhaltspunkte zu den validen Gütekriterien, um die Kritikpunkte an den illustrativen Modellen zu überarbeiten. Weiterhin zeigt die Befragung der Lernenden zu den Konstrukten *System Usability*,

Situationales Interesse und *Cognitive Load* während der Intervention, dass diese die Modelle als gleichwertig einstufen. Daher kann die Qualität der Modelle während der Interventionsstudie als vergleichbar angesehen werden.

Die Konstrukte *System Usability*, *Situationales Interesse*, *Fachinteresse Chemie*, *Fähigkeits-selbstkonzept* und die Subskalen des *Cognitive Loads* wurden mittels Fragebögen zur Selbsteinschätzung erhoben. Diese Art der Datenerhebung unterliegt generell diversen Arten der Verzerrung, da es den Befragten schwer fallen kann, sich selbst korrekt einzuschätzen (Paulhus & John, 1998). Beispielsweise kommt es aufgrund des Rückschaufehlers (=hindsight bias) zu Verzerrungen bei der Erinnerung an die eigene Leistung und den Schwierigkeitsgrad von Aufgaben (Roese & Vohs, 2012). Dadurch ist sowohl für die Bewertung der *System Usability* als auch der Subskalen des *Cognitive Loads* mit Verzerrungen zu rechnen. Zudem kann anhand der vorliegenden Befunde keine Aussage darüber getroffen werden, ob auf individueller Ebene in Einzelfällen die Kapazität des Arbeitsgedächtnisses aufgrund der Additivität der *Cognitive Load*-Typen überschritten wurde. Für eine solche Untersuchung wäre gemäß der neueren Interpretation der Additivität der Load-Typen nach Paas et al. (2003) eine Berücksichtigung der freien Kapazität an kognitiven Ressourcen der Probandinnen und Probanden erforderlich. Die Feststellung der freien Kapazität mithilfe von Fragebögen zur Selbsteinschätzung gilt jedoch als kaum realisierbar (ebd.). Weiterhin hängt die Bewertung des *Cognitive Loads* von der *aktuellen Motivation* der Befragten ab (Paas et al., 2005). In diesem Zusammenhang hätten neben dem situationalen Interesse noch die restlichen Subskalen des FAMs (Rheinberg et al., 2001) miterhoben werden können. Da die Subskala zum situationalen Interesse jedoch im gleichen Maße zur Aufklärung der Gesamtvarianz wie die Kumulation der restlichen Subskalen beiträgt (ebd.), wurde sich zugunsten der Testökonomie dagegen entschieden.

Weiterhin kann durch eine gezielte Analyse der inhaltlichen Aspekte der *System Usability* ein tieferes Verständnis der Benutzerfreundlichkeit der Lernumgebungen gewonnen werden. Lewis und Sauro (2009) schlagen in diesem Zusammenhang eine zweifaktorielle Modellierung der *System Usability* vor. In ihrem Ansatz grenzen sie zwei Items der Skala als den inhaltlichen Faktor „Learnable“ vom Faktor „Usable“ ab (Lewis & Sauro, 2009, S. 95). Diese Modellierung könnte wertvolle Einblicke in die lernrelevanten Aspekte der Usability liefern. Zusätzlich zur Betrachtung der Learnable-Komponente könnte auch die Analyse der *Content Usability* weitere Aufschlüsse über die Bedienerfreundlichkeit von Lernumgebungen geben.

Das eingesetzte Messinstrument zum Modellverständnis nach Leisner (2005a) ist trotz der guten Kennwerte kritisch zu betrachten. Die Antwortoption „habe keine Meinung“ wird als mittlere Ausprägung der Skala kodiert (ebd.). Dadurch erreichen Schüler*innen, die diese Option ankreuzen die Hälfte der möglichen Punkte der Skala. Dies führt dazu, dass Personen, die

tatsächlich keine Meinung haben, den gleichen Score erhalten wie diejenigen, die eine moderate Zustimmung oder Ablehnung ausdrücken. Dies kann zu einer künstlichen Normalisierung der Verteilung der Daten und damit zur Verzerrung des Skalenmittelwerts (Krosnick et al., 2009) und somit zum Verlust von Varianz führen. Dadurch erschweren sich die inferenzstatistische Analyse und weitere Interpretationen der Ergebnisse.

Alle eingesetzten Fragebogenskalen, mit Ausnahme der persönlichen Angaben und des Fachwissenstests, sind bereits validierte und publizierte Messinstrumente. Daher kann neben dem Gütekriterium der Validität auch die Objektivität für diese Skalen als erfüllt angesehen werden (Döring & Bortz, 2016). Da der Fachwissenstest nach den Kriterien gängiger Fachliteratur zur Fragebogenkonstruktion und Testtheorie (z.B. Bühner, 2006; Döring & Bortz, 2016; Moosbrugger & Kelava, 2012; Schmiemann & Lücken, 2014) konzipiert wurde und die Fragebogenmethode nach Döring und Bortz (2016) generell als objektives Verfahren einzustufen ist, kann auch für den Fachwissenstest Objektivität angenommen werden. Um finale Aussagen zur Objektivität treffen zu können, müsste das Messinstrument jedoch noch von weiteren Testleitern in anderen Kontexten erprobt werden. Da während der Testdurchführung kein weiteres Eingreifen der Testleitung über das Austeilen der Fragebögen hinaus erfolgte, kann diese ebenfalls als objektiv eingeschätzt werden.

Die Reliabilitäten der eingesetzten Skalen weisen alle Werte für Cronbachs $\alpha > 0.7$ und McDonald's $\omega > 0.7$ auf und liegen damit im akzeptablen Bereich (Dunn et al., 2014; George & Mallery, 2003). Auch die restlichen Kennwerte der Items und Skalen liegen alle, mit Ausnahme der Items CL4 und CL6 der ECL-Subskala im Normbereich. Die Items CL4 und CL6 weisen mit einer Itemschwierigkeit von $P_{\text{Dahl}} < 20$ auf und sind daher als zu schwer zu bewerten. Die Itemkennwerte für Trennschärfe und Schiefe und Kurtosis liegen im Normbereich. Daher wurden die Items nicht von der Skala ausgeschlossen und somit etwaige Bodeneffekte zugunsten einer gesteigerten Varianz relativiert.

Zur Bewertung der Reliabilität der genutzten Messinstrumente werden die ermittelten internen Konsistenzen mit den Ergebnissen aus inhaltlich vergleichbaren, empirischen Normierungsstudien abgeglichen. Die jeweiligen α -Werte (Cronbach, 1951) sind in Tabelle 33 aufgeführt. Es liegen nur geringe Abweichungen zu Vorgängerstudien vor. Daher kann die allgemeine Testgüte als ausreichend und plausibel eingeschätzt werden.

Tabelle 33: Für die Konstrukte FSK = Chemiebezogenes Fähigkeitsselbstkonzept, FIC = Chemiebezogenes Fachinteresse, SI = Situationales Interesse, SU= System Usability, Intrinsic Cognitive Load, Extraneous Cognitive Load, Germane Cognitive Load, FW = Fachwissen und MV = Modellverständnis sind die Reliabilitäten nach Cronbach (α) der vorliegenden Interventionsstudie mit Kennwerten aus Normierungsstudien (α_{LIT}) in Beziehung gesetzt.

Konstrukt	α	α_{LIT}	Quelle
FSK	0.86	0.89	Engl, 2020
FIC	0.84	0.89	Engl, 2020
SI	0.85	0.71 – 0.90 0.74 – 0.80	Rheinberg et al., 2001 Cauet, 2016
SU	0.82	0.91 0.81	Bangor et al., 2008 Neff, 2023
ICL	0.93	0.86 – 0.91 0.80 – 0.90	Leppink et al., 2013 Neff, 2023
ECL	0.71	0.63 – 0.87 0.75 – 0.78	Leppink et al., 2013 Neff, 2023
GCL	0.86	0.94 – 0.97 0.87 – 0.93	Leppink et al., 2013 Neff, 2023
FW	0.81	0.73 0.63	Freyer, 2013 Engl, 2020
MV	0.80	0.77	Leisner, 2005a

Die Reliabilität des Fachwissenstests muss hierbei weiter differenziert betrachtet werden. Die Höhe der beiden internen Konsistenzmaße α und ω steigt stark mit der Anzahl der Items (Döring & Bortz, 2016). Da die Skala aus 21 Items besteht ist hier von einer Überschätzung der Reliabilität auszugehen. Mit Cronbachs $\alpha > 0.8$ und McDonald's $\omega > 0.8$ liegen die Werte im Bereich einer guten internen Konsistenz (Dunn et al., 2014; George & Mallery, 2003). Daher kann selbst bei einer Überschätzung der Reliabilität aufgrund der hohen Itemanzahl noch ein akzeptabler Wert angenommen werden kann.

Neben der Sicherstellung der Inhaltsvalidität durch eine Orientierung der Itemkonstruktion an den formulierten Lernzielen und empirischen Studien zu Präkonzepten erfolgte eine Überprüfung der Augenscheinvalidität durch drei weitere Fachdidaktiker aus der Arbeitsgruppe Chemiedidaktik der RPTU Landau. Als robustes Verfahren zur Prüfung der Konstruktvalidität erfolgte die Kombination aus einer explorativen Faktorenanalyse (EFA) gefolgt von einer konfirmatorischen Faktorenanalyse (CFA) (Brown, 2015). Während die EFA verwendet wurde, um die zugrunde liegende Struktur der relativ großen Menge an manifesten Variablen aufzudecken (Fabrigar & Wegener, 2011), wurde die gefundene Faktorenstruktur mittels CFA bestätigt und zu einem validierten Modell überarbeitet (Kline, 2023). Die zur Bewertung des Modells herangezogenen Fit-Indizes und erreichten Werte wurden bereits ausführlich in Kapitel 8.2

diskutiert. Kritisch ist hierbei noch zu erwähnen, dass die verwendeten Grenzwerte der Fit-Indizes für die CFA großzügig gewählt sind (Hu & Bentler, 1999). Das Konstrukt Fachwissen umfasst Wissen über sechs unterschiedliche Versuche und ist daher breit gestreut. Dementsprechend komplex gestaltet sich das zu überprüfende Modell zur Datenstruktur. Da der Test das Konstrukt Fachwissen noch möglichst breit messen soll, empfiehlt es sich jedoch die weniger strikten Grenzwerte für die Fit-Indizes anzunehmen (Hair et al., 2010; Kline, 2023; MacCallum et al., 1996; Marsh et al., 2004). Weiterhin birgt die Vorgehensweise die Validierung mittels EFA und CCFA anhand von abhängigen Daten durchzuführen potenzielle Probleme (Moosbrugger & Schermelleh-Engel, 2012). Aufgrund der zeitlichen Abhängigkeit der Schätzungen kann es zu Verzerrungen dieser kommen (Beaujean, 2014). Zudem wurde auf die Anwendung fortgeschrittener longitudinaler Analysemethoden, wie Latent State-Trait Modelle oder Latent Growth Curve Modelle verzichtet (Brown, 2015). Diese Methoden hätten eine präzisere Berücksichtigung der temporalen Interdependenz der Daten sowie eine formale Prüfung der Messinvarianz über die Zeit und damit der Generalisierbarkeit und Stabilität des Faktormodells ermöglicht (Liu et al., 2017). Der entwickelte Fachwissenstest wurde spezifisch konzipiert, um die Inhalte der in dieser Studie durchgeführten Intervention abzubilden. Das Instrument zielt darauf ab, die unmittelbaren Effekte der Intervention auf das fachspezifische Wissen der Teilnehmenden zu erfassen. Eine Übertragbarkeit auf andere Forschungskontexte oder eine Generalisierbarkeit über den spezifischen Anwendungsfall hinaus wurde bei der Konstruktion des Tests nicht angestrebt und ist folglich nicht intendiert. Die Konzeption des Fachwissenstests verfolgt das Ziel, eine valide Messung zu beiden Erhebungszeitpunkten (Pre und Post) zu gewährleisten. Die Itemkennwerte erwiesen sich zu beiden Erhebungszeitpunkten als adäquat, was auf eine konsistente Funktionsweise der einzelnen Testitems hindeutet. Darüber hinaus zeigte der Test zu beiden Messzeitpunkten eine hohe interne Konsistenz (Cronbachs $\alpha > .80$), was auf eine zuverlässige Erfassung des zugrundeliegenden Konstrukts schließen lässt. Darüber hinaus wurde eine signifikante Veränderung der Testwerte zwischen den Messzeitpunkten beobachtet, indiziert durch einen großen Effekt ($\hat{\delta}_j^{(R)} = 1.02$). Weiterhin dient die starke Korrelation nach Pearson ($r = 0.63$) zwischen den Messzeitpunkten Pre und Post als Anzeichen für hohe Kriteriums- (Carmines & Zeller, 1979) und Konstruktvalidität (Messick, 1995). Diese Befunde unterstützen die Annahme der Messinvarianz über die Zeit und stärken somit die Validität der Interpretation von Veränderungen in den Testwerten als tatsächliche Veränderungen im Fachwissen der Teilnehmer*innen.

9. Fazit und Ausblick

Relevanz der Arbeit für das Forschungsfeld

- Aufzeigen von Herausforderungen und Lösungsansätzen für Interventionsstudien zum Vergleich verschiedener Modelltypen in den Naturwissenschaftsdidaktiken
- Formulierung von Kriterien zur Bewertung der Qualität von Anschauungsmodellen in den Naturwissenschaftsdidaktiken
- Entwicklung und Validierung eines Testinstruments zur Bewertung der Qualität von Anschauungsmodellen in den Naturwissenschaftsdidaktiken
- Erarbeitung konkreter Umsetzungsbeispiele für die kriteriengeleitete Konzeption lernförderlicher Modelle
- Generierung und Bereitstellung lernförderlicher und praxisnaher Lernumgebungen für Schule und Lernforschung
- Erörterung von inferenzstatistischen Hinweisen auf die unterschiedliche Lernwirksamkeit der Modelltypen auf das Konstrukt Modellverständnis
- Darstellung von relevanten Kenn- und Referenzwerten
- Kreierung von Grundlagen für weitere Forschung zum Zusammenhang von Expert*innen- bzw. Noviz*innen-Niveau und dem Lernerfolg der Modelltypen sowie dem Einfluss von multiplen, externen Repräsentationen auf das Modellverständnis

Ausgehend vom Ziel den Einfluss unterschiedlicher Modelltypen auf das Modellverständnis, das Fachwissen und das situationale Interesse von Schülerinnen und Schülern im Chemieunterricht zu untersuchen, wurden in diesem Forschungsprojekt verschiedene empirische Untersuchungen durchgeführt. Das vorliegende Forschungsprojekt präsentiert neue Erkenntnisse bezüglich der Auswirkungen von AR-Modellen im Vergleich zu haptisch-interaktiven und illustriativen Modellen auf die Lernprozesse im Chemieunterricht. Um valide Aussagen treffen zu können, wurden in Hinblick auf das Studiendesign Herausforderungen und Lösungsansätze für Medienvergleichsstudien aufgezeigt: Es sollten qualitativ gleichwertige Modelle in speziell für Vergleichszwecke entwickelten Unterrichtseinheiten eingesetzt werden. Weiterhin gilt es, den Neuheitseffekt zu berücksichtigen sowie geeignete Erhebungsinstrumente sowohl für die abhängigen als auch für die zu kontrollierenden Variablen zu verwenden.

In **Studie I** wurden zur Überprüfung der Qualität der Modelltypen Gütekriterien für Anschauungsmodelle formuliert und empirisch getestet, um somit eine valide Basis für die anschließenden Untersuchungen zu schaffen. Durch die qualitative und quantitative Validierung dieser Kriterien wurde ein Messinstrument entwickelt, das als Grundlage für die objektive Beurteilung der Modellqualität von Anschauungsmodellen dient.

Mithilfe dieses Messinstruments wurden in **Studie II** die drei Modelltypen – AR, HI und IL – auf Basis der entwickelten Qualitätskriterien durch Expert*innenratings systematisch bewertet. Mit der anschließenden Anpassung der Modelle wurde die Gleichwertigkeit der Modellqualität gewährleistet. Infolge dessen kann die Forschungslücke hinsichtlich eines fairen Vergleichs dieser Modelltypen als geschlossen betrachtet werden. Die empirisch fundierte Sicherstellung der

Modellqualität in den Vergleichsgruppen stellt ein Alleinstellungsmerkmal der vorliegenden Arbeit dar und trägt wesentlich zur Erhöhung der internen Validität bei. Diese methodische Innovation, Modelle anhand objektiver Qualitätskriterien empirisch zu bewerten, stellt einen bedeutenden Fortschritt in der naturwissenschaftsdidaktischen Forschung zu Lehr-Lern-Modellen dar und bildet das Fundament für die abschließende Interventionsstudie.

In der abschließenden Interventionsstudie (**Studie III**) wurde der Einfluss dieser qualitativ bewerteten Modelltypen auf das kognitive und affektive Lernen von Schüler*innen untersucht. Insbesondere der Einfluss des Modelltyps auf das Modellverständnis bietet hierbei ein Novum in der naturwissenschaftsdidaktischen Community. Hierzu liefern die inferenzstatistischen Auswertungen der Interventionsstudie Hinweise auf die unterschiedliche Lernwirksamkeit der Modelltypen. Die Ergebnisse zeigen, dass AR- und IL-Modelle signifikant zur Verbesserung des Modellverständnisses der Schüler*innen beitragen und die Fähigkeit zur Differenzierung zwischen Modell und Realität stärken. Dies ist insbesondere im Chemieunterricht von zentraler Bedeutung, da Schüler*innen oft Schwierigkeiten haben, abstrakte Modelle als vereinfachte Darstellungen der Realität zu verstehen. Aus diesen Ergebnissen ergibt sich das Forschungsdesiderat, die Ursachen dafür zu untersuchen, warum haptisch-interaktive Modelle im Vergleich zu den anderen Modelltypen ein geringeres Potenzial zur Förderung des Modellverständnisses aufweisen.

Die gewonnenen Erkenntnisse leisten einen Beitrag zur Normierung zukünftiger Untersuchungen in Form der Bereitstellung eines Messinstruments zur Qualitätssicherung der Vergleichsgruppen sowie von Kenn- und Referenzwerten für die untersuchten Konstrukte. Darüber hinaus wurden praxisnahe, lernwirksame Unterrichtseinheiten zum Wissen über das Teilchenmodell im Kontext von Löse- und Diffusionsprozessen entwickelt. Die im Projekt erstellten Lernmaterialien und Apps sind über die Einbindung in die Online-Plattform *zenodo.org* als Open Educational Resources (OER) der Öffentlichkeit zur Verfügung gestellt. In Kooperation mit dem Pädagogischen Landesinstitut Rheinland-Pfalz könnte die Lerneinheit mit Lehrkräften im Rahmen von Lehrkräftefortbildungen für die Verwendung im Unterricht diskutiert und weiterentwickelt werden. Die Produkte und Erkenntnisse der Dissertation haben bereits Eingang in die universitäre Lehrkräfteausbildung gefunden. So werden sie im Modul „Grundlagen der Chemie“ im Studiengang B.Ed. Biologie sowie in den Modulen „Aktuelle Themen des Chemieunterrichts“ und „Anorganische Chemie für Fortgeschrittene“ M.Ed Chemie bereits intensiv genutzt und reflektiert. Ein weiterer Transfer in andere Module des B.Ed. Chemie ist in Vorbereitung. Weiterhin liefern die konzipierten Lernumgebungen einen Beitrag, um Grundlagen für den Bedarf an prozessbezogener Forschung (Krüger et al., 2018; Louca & Zacharia, 2012; Nicolaou & Constantinou, 2014) zum Problemlöseverhalten während der Anwendung der Modellmethode zu schaffen (Gogolin & Krüger, 2018; Krell et al., 2017; Orsenne, 2016). Durch

diese qualitative Prozessforschung könnte die aus den Ergebnissen dieser Arbeit abgeleitete Frage beantworten werden, warum die HI-Modelle im Vergleich zu den AR- und IL-Modellen zu einem geringeren Zuwachs im Modellverständnis führten. Ferner besteht Forschungsbedarf dazu, wie das Fachwissen, das Wissen über bestimmte Modelle und das epistemologische Wissen zu Modellen hinsichtlich ihrer Anwendung auf Probleme zusammenhängen (Krüger et al., 2018). Weiterführende Analysen der in dieser Arbeit gewonnenen Daten, können zur Klärung dieses Desiderats beitragen. Höffler (2008) empfiehlt zur Klärung solcher Fragestellungen im Rahmen von Medienvergleichsstudien auf Basis des Vorwissens einen Mediansplit zur Trennung zwischen Noviz*innen und Expert*innen. Mittels einer zweifaktoriellen ANOVA wird anschließend der Effekt der beiden unabhängigen Variablen Vorwissen und Modelltyp auf den Lernerfolg als abhängige Variable untersucht. Weitere mögliche Auswertungen auf Basis der vorliegenden Daten, wären die Prüfung auf Moderationseffekte der untersuchten Konstrukte untereinander mittels Regressionsanalysen. Ebenso können die vorliegende Studie und der Datensatz noch durch eine weitere Experimentalgruppe erweitert werden, die im Sinne von multiplen externen Repräsentationen alle drei Modelltypen simultan verwendet. So sehen Wu und Puntambekar (2012) die simultane Nutzung unterschiedlicher Medien im Vergleich zum Arbeiten mit nur einem Medientyp unter Berücksichtigung der Theorie der kognitiven Flexibilität als vielversprechenden Forschungsansatz. Abschließend ist anzumerken, dass die Berücksichtigung weiterer Kovariaten während der Datenerhebung und -auswertung zusätzliche Gewissheit über die Vergleichbarkeit der Gruppen und der Belastbarkeit der Ergebnisse liefern würde. In Bezug auf die Kovariaten wäre zunächst das Selbstkonzept im Umgang mit Medien zu nennen. Darüber hinaus könnte die Erhebung der Einstellungen der Befragten zu den Modelltypen Aufschluss über etwaige Vorurteile in Bezug auf die Modelltypen geben.

10. Verzeichnisse

10.1 Abbildungsverzeichnis

Abbildung 1: Urteil über das Modellsein (verändert nach Mahr, 2008a; Upmeyer zu Belzen & Krüger, 2010)	7
Abbildung 2: Schema zum "Denken in Modellen" (nach Steinbuch, 1977)	13
Abbildung 3: Schema der Modell-Original-Subjekt-Relation (nach Kircher, 1995)	13
Abbildung 4: erweitertes M-O-S-Schema zu Modellen zu Struktur der Materie (nach Saborowski, 2000)	15
Abbildung 5: Themengebiete der Interventionen aus dem Rheinland-Pfälzischen Lehrplan für Chemie in Klassenstufe 8. (das Thema Osmose wurde aus zeitlichen Gründen von der Intervention ausgeschlossen und ist daher ausgegraut)	81
Abbildung 6: Ergebnissicherung nach dem Experiment zur Temperaturabhängigkeit der Diffusion. Es liegen optische Marker zur Symbolisierung der drei Ebenen - Blauer Kasten und Erlenmeyerkolben: „Reale Welt der Experimente und Phänomene“; Grüner Kasten und Wolke: „Ausgedachte Modellwelt“; Oranger Kasten und Glühbirne „Modellkritik“	85
Abbildung 7: Beispiel zum Einstieg in die erste Doppelstunde zum Thema Brownsche Bewegung. Zu sehen sind der grüne Kasten, die grün hinterlegte Seitenzahl und die grüne Wolke als optische Marker für die Modellebene.	87
Abbildung 8: Versuchsanleitung zum Versuch Diffusion	91
Abbildung 9: Nutzung des AR-Modells zur Bildung der Hydrathülle - oben links: Augmented Solutions-Karte als optischer Marker; Mitte: Tablet mit dem AR-Modell zur Teilchendarstellung der Bildung der Hydrathülle	94
Abbildung 10: Arbeitsmaterial und -auftrag zum AR-Modell zur Bildung der Hydrathülle	95
Abbildung 11: Eines von vier Legebildern zur Darstellung der Bildung der Hydrathülle. Zu erkennen sind die blauen Wasserteilchen, die gelben Salzteilchen sowie die roten Pfeile zur Symbolisierung der Teilchenbewegung und weißen Highlightings der relevanten Stellen.	96
Abbildung 12: Arbeitsmaterial und -auftrag zum illustrativen Modell zur Bildung der Hydrathülle	97
Abbildung 13: Arbeitsmaterial und -auftrag zum haptisch-interaktiven Modell zur Bildung der Hydrathülle	99
Abbildung 14: Gemittelte Gesamtgüte sowie die fünf Gütekriterien der drei Modelltypen IL = Illustratives Modell, HI = Haptisch-interaktives Modell und AR = Augmented Reality Modell; Wertebereich: 0 = „stimme nicht zu“, 1 = „stimme eher nicht zu“, 2 = „stimme eher zu“, 3 = „stimme zu“; * = $p < 0.05$ (Przywarra et al., 2021, S. 33)	119
Abbildung 15: Ablauf der Entwicklung der Materialien und Messinstrumente sowie deren Analyse (RSP1 = Klasse 1 an der RealschulePlus, GYM1 = Klasse 1 am Gymnasium, LIX = Lesbarkeitsindex, RIX = Regenburger Index, MV = Modellverständnis, SI = situationales Interesse, SU = System Usability, CL = Cognitive Load, KF = Kognitive Fähigkeiten, FI = Fachinteresse, FSK = Fähigkeitsselbstkonzept, FW = Fachwissen)	125
Abbildung 16: Beispiel-Item des Fachwissenstests, dessen Antwortmöglichkeiten auf existierenden Präkonzepten basieren	126

Abbildung 17: QQ-Plots der Verteilung der studentisierten Residuen auf die theoretischen Quantile der Indizes gSMOG, RIX und LIX: Die Darstellungen weisen keine gravierenden Abweichungen von der Normalverteilung auf.....	134
Abbildung 18: Personenbezogene Mittelwerte der Skalen Situationales Interesse, Intrinsic Cognitive Load, Extraneous Cognitive Load, Germane Cognitive Load und System Usability.....	136
Abbildung 19: Screeplot zur Parallelanalyse zum Fachwissenstest (FW3) nach Ausschluss der Items mit unzureichenden Itemkennwerten. Der Datenverlauf der empirischen Faktoren in blau wird mit dem Verlauf der simulierten Daten in rot verglichen. a) stellt die Parallelanalyse zu den 24 Items verbleibenden Items nach der Itemanalyse dar. b) zeigt die Parallelanalyse zu den 22 Items verbleibenden Items nach der explorativen Faktorenanalyse	146
Abbildung 20: Entwicklung des Intrinsic Cognitive Loads zu den drei Interventionszeitpunkten im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau. Ein hoher Zahlenwert bedeutet eine hohe kognitive Belastung aufgrund der Aufgabenschwierigkeit.....	158
Abbildung 21: Entwicklung des Extraneous Cognitive Loads zu den drei Interventionszeitpunkten im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau. Ein hoher Zahlenwert bedeutet eine hohe kognitive Belastung aufgrund der Arbeitsmaterialien.....	159
Abbildung 22: Histogramm zur Verteilung der Antworten der Gesamtstichprobe zum situationalem Interesse über die theoretischen ermittelten Quantile.....	160
Abbildung 23: QQ-Plot zur Verteilung der studentisierten Residuen der messwiederholten einfaktoriellen ANOVA zum situationalen Interesse über die theoretischen ermittelten Quantile.....	160
Abbildung 24: Entwicklung des situationalen Interesses zu den drei Interventionszeitpunkten im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau (Skala von 1-7). Ein hoher Zahlenwert bedeutet ein hohes Interesse an den Inhalten der jeweiligen Intervention.....	162
Abbildung 25: Histogramm zur Verteilung der Antworten der Gesamtstichprobe im Fachwissenstest über die theoretischen ermittelten Quantile	163
Abbildung 26: QQ-Plot zur Verteilung der studentisierten Residuen der messwiederholten einfaktoriellen ANOVA zum situationalen Interesse über die theoretischen ermittelten Quantile.....	163
Abbildung 27: Entwicklung des Fachwissens zu den drei Interventionszeitpunkten im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau.	165
Abbildung 28: Histogramm zur Verteilung der Antworten der Gesamtstichprobe zum Konstrukt Modellverständnis über die theoretischen ermittelten Quantile	166
Abbildung 29: QQ-Plot zur Verteilung der studentisierten Residuen der messwiederholten einfaktoriellen ANOVA zum situationalen Interesse über die theoretischen ermittelten Quantile.....	166
Abbildung 30: Entwicklung des Modellverständnisses zu den drei Interventionszeitpunkten im Vergleich der Gruppen AR in Rot, HI in Grün und IL in Blau. Ein hoher Wert bedeutet ein Verständnis über Modelle, das dem wissenschaftlichen Konsens entspricht.....	168
Abbildung 31: Einstichproben-t-Test zum Abgleich des Modellverständnisses der Gesamtstichprobe zum Post-Zeitpunkt (gepunktete Linie) mit der Normierung (gestrichelte Linie). Das Modellverständnis in der vorliegenden Intervention fällt mit einem kleinen Effekt signifikant geringer aus.	169

10.2 Tabellenverzeichnis

Tabelle 1: Kriterien und Konsequenzen zum Verhältnis zwischen Modell und Original nach Fischer et al., 1995, verändert nach Upmeier zu Belzen & Krüger, 2010, S.41	11
Tabelle 2: Kriterien zur Klassifikation von Modellen (verändert nach Upmeier zu Belzen, 2013; erweitert nach Keller, 1977; Christen, 1990; Eschenhagen, Kattmann, & Rodi, 2008; Fleige et al., 2012; Meisert, 2014).....	26
Tabelle 3: Positionierung der Dimensionen der nationalen und internationalen Kompetenzmodelle zur Modellkompetenz in Bezug auf Fachwissen und Erkenntnisgewinnung. Hellgrau unterlegt: Kompetenzmodell auf das sich als theoretische Grundlage in dieser Arbeit gestützt wird. Dunkelgrau unterlegt: Modellverständnis als epistemologische Dimension.	37
Tabelle 4: Aspekte von Modellen im Kontext von Modellverständnis im internationalen Vergleich (erweitert nach Krell et al., 2014)	38
Tabelle 5: Übersicht über den Forschungsstand zu Medienvergleichsstudien im Kontext von naturwissenschaftlichem Lernen (Anim = Animationen, IL = Illustrationen, AR = Augmented Reality, Hap = Haptisch, ECL = Extraneous Cognitive Load, selbstg. = selbstgesteuert, fremdg. = fremdgesteuert, PPP= PowerPointPräsentation, AC = Anorganische Chemie, OC = Organische Chemie)	68
Tabelle 6: Übersicht der für die Intervention entwickelten Versuche	89
Tabelle 7: Übersicht zu den Modellen der drei Modelltypen zu den Phänomenen der Intervention.	93
Tabelle 8: Ergebnisse der Kodierung des Probematerials aus dem ersten Zyklus nach Proband*innen, Kodiereinheiten und Kodierenden sortiert.....	106
Tabelle 9: Ergebnisse der Kodierung des zweiten Zyklus (Anmerkung: Die Ziffer vor dem Punkt gibt die Kategorie an, die Ziffer hinter dem Punkt gibt die Subkategorie an; Abweichungen sind fett markiert).....	109
Tabelle 10: Interraterreliabilität für die Items der jeweiligen Gütekriterien	109
Tabelle 11: Kennwerte für die Intraraterreliabilitäten zu den Gütekriterien	112
Tabelle 12: Deskriptive Kennwerte der Items und Gesamtskala des BeQuAn.....	117
Tabelle 13: Ergebnisse der inferenzstatistischen Auswertungen mit dem Kruskal Wallis Test (Przywarra et al., 2021, S. 32)	118
Tabelle 14: Bewertung der einzelnen Gütekriterien sowie der Gesamtgüte (Przywarra et al., 2021, S. 33)	118
Tabelle 15: Anzahl der Kommentare zu den Modelltypen aufgelistet nach den Gütekriterien (die Prozentzahl zu den einzelnen Kriterien bezieht sich immer auf den Anteil an der Gesamtanzahl [N = 180] der Kommentare)	120
Tabelle 16: Kategorien zu den Kritikpunkten der drei Modelltypen.....	121
Tabelle 17: Ablauf der Datenerhebung während der Pilotierung.....	125
Tabelle 18: Kategoriensystem zur qualitativen Auswertung von Adressatengerechtigkeit und Verständlichkeit der Materialien auf Basis der Videodateien	130
Tabelle 19: Deskriptive Statistiken für die Aufgaben, Versuchsanleitungen und Lückentexte. Für die Indizes Lesbarkeitsindex (LIX), Regensburger Index (RIX) und Simple Measure of Gobbledygook – german (gSMOG) sind die deskriptiven Kennwerte MW = Mittelwert,	

Median, SD = Standardabweichung, Schiefe, Kurtosis und Anzahl der Beobachtungen n angegeben. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert. 133

Tabelle 20: Für die Indizes Lesbarkeitsindex (LIX) sowie der Regensburger Index (RIX) und der Simple Measure of Gobbledygook – german (gSMOG) sind die deskriptiven Kennwerte MW = Mittelwert, SD = Standardabweichung, die Testvoraussetzungen Normalverteilung und Varianzhomogenität sowie die Ergebnisse der Unterschiedsprüfungen mit Effektstärken. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert. 134

Tabelle 21: Liste der Codes zu den identifizierten Schwierigkeiten während der Durchführung der Lerneinheiten; die Anzahl der kodierten Segmente steht in den Klammern. 135

Tabelle 22: Für die Konstrukte Situationales Interesse, Cognitive Load und System Usability sind die Itemkennwerte MW = Mittelwert, Median SD = Standardabweichung, Schiefe, Kurtosis, d = Schwierigkeitsindex nach Dahl, r_{it} = Korrigierte Trennschärfe und die Reliabilität nach Cronbach α angegeben; n entspricht der Anzahl an messwiederholten Datenpunkten, die Zahl in der Klammer zu n entspricht der tatsächlichen Anzahl an Personen. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert. 136

Tabelle 23: Für die Skalen SI = Situationales Interesse, ICL = Intrinsic Cognitive Load, ECL = Extraneous Cognitive Load, GCL = Germane Cognitive Load, SU = System Usability, FIC = Chemiebezogenes Fachinteresse, FSK = Chemiebezogenes Fähigkeitsselbstkonzept, MV1 = Modellverständnis Version 1, MV2 = Modellverständnis Version 2, FW3 = Fachwissen Version 3 und FW4 = Fachwissen Version 4 sind die Itemkennwerte r_{it} = Korrigierte Trennschärfe, P_{Dahl} = Schwierigkeitsindex nach Dahl, α = Reliabilität nach Cronbach, ω = Reliabilität nach McDonald, Schiefe, Kurtosis, Varianzaufklärung, $n_{effektiv}$ = effektive Stichprobengröße und der Dropout der Items auf Grundlage der Itemanalyse angegeben. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert. 145

Tabelle 24: Vergleich der zwei Messmodelle zur Fachwissensskala anhand des χ^2 -Tests mit zugehörigen p-Werten im Verhältnis des χ^2 -Werts mit der dreifachen Anzahl der Freiheitsgrade df, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Residual. Auffälligkeiten und Abweichungen der Kennwerte sind fett markiert. Alle Werte wurden mit dem robusten DWLS-Schätzer und NLMINB-Optimierungsmethode berechnet. 147

Tabelle 25: Anpassung der Itemformulierung. Die Formulierung vor der Überarbeitung ist durchgestrichen. 148

Tabelle 26: Methoden und Stichprobe der Untersuchungen auf Gruppenunterschiede 149

Tabelle 27: Einteilung der Klassen in die Experimentallgruppen. Den teilnehmenden Schulen wurde ein Farbcode zugeordnet. Der Lehrkräfte-Code ist kursiv dargestellt. 149

Tabelle 28: Ablauf der Datenerhebung mittels Fragebögen während der Interventionsstudie. Grau unterlegt sind die Konstrukte, die Pre und Post erhoben wurden; weiß unterlegt: Konstrukte, die begleitend zu den Interventionszeitpunkten erhoben wurden. 150

Tabelle 29: Unterschiede der Kovariablen zwischen den Experimentalgruppen zum Messzeitpunkt 0 (Pre) mit den Levene-Teststatistiken zur Analyse mittels einfaktorieller ANOVA und der zugehörigen Effektgröße. AR = Augmented Reality, HI = Haptisch-Interaktiv, MW = Mittelwert, SD = Standardabweichung, OR = Odds Ratio. 157

Tabelle 30: Deskriptive Statistiken zum situationalen Interesse der einzelnen Experimentalgruppen und Messzeitpunkten (= MZP); AR = Augmented Reality, HI = Haptisch-Interaktiv, MW = Mittelwert, SD = Standardabweichung, min = erreichtes Minimum, max = erreichtes Maximum; Skala von 1-7. 161

Tabelle 31: Deskriptive Statistiken zum Fachwissen der einzelnen Experimentalgruppen und Messzeitpunkten (= MZP); AR = Augmented Reality, HI = Haptisch-Interaktiv, MW = Mittelwert, SD = Standardabweichung, min = erreichtes Minimum, max = erreichtes Maximum; mögliches Maximum = 21.....	165
Tabelle 32: Deskriptive Statistiken zum Modellverständnis der einzelnen Experimentalgruppen und Messzeitpunkten (= MZP); AR = Augmented Reality, HI = Haptisch-Interaktiv, MW = Mittelwert, SD = Standardabweichung, min = erreichtes Minimum, max = erreichtes Maximum; Skala von 1-5.....	168
Tabelle 33: Für die Konstrukte FSK = Chemiebezogenes Fähigkeitsselbstkonzept, FIC = Chemiebezogenes Fachinteresse, SI = Situationales Interesse, SU= System Usability, Intrinsic Cognitive Load, Extraneous Cognitive Load, Germane Cognitive Load, FW = Fachwissen und MV = Modellverständnis sind die Reliabilitäten nach Cronbach ($=\alpha$) der vorliegenden Interventionsstudie mit Kennwerten aus Normierungsstudien ($=\alpha_{LIT}$) in Beziehung gesetzt.	179

10.3 Literaturverzeichnis

- Abel, G. (2008). Modelle und Wirklichkeit. In U. Dirks, & E. Knobloch, Modelle (S. 31-46). Frankfurt am Main: Peter Lang Verlag.
- Abraham, M., Varghese, V. & Tang, H. (2010). Using molecular representations to aid student understanding of stereochemical concepts. *Journal of Chemical Education*, 87(12), 1425-1429.
- Ackerman, P. L., Beier, M. E. & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological bulletin*, 131(1), 30.
- Ali, N. A., Sadiq, M. H., Albabawat, A. A. & Salah, R. M. (2022). Methods and Applications of Augmented Reality in Education: A Review. In 2022 International Conference on Computer Science and Software Engineering (CSASE) (S. 175–181). IEEE.
- Andersson, B. (1990). Pupils' Conceptions of Matter and its Transformations. *Studies in Science Education*, 18:1, S. 53-85.
- Ardac, D. & Akaygun, S. (2004). Effectiveness of multimedia-based instruction that emphasizes molecular representations on students' understanding of chemical change. *Journal of Research in Science Teaching*, 41(4), 317–337.
- Ardito, C., Costabile, M. F., Marsico, M. de, Lanzilotti, R., Levialdi, S., Roselli, T. & Rossano, V. (2006). An approach to usability evaluation of e-learning applications. *Universal access in the information society*, 4, 270–283.
- Astuti, F. N., Suranto, S. & Masykuri, M. (2019). Augmented reality for teaching science: Students' problem solving skill, motivation, and learning outcomes. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, 5(2), 305–312.
- Aydeniz, M., & Kotowski, E. L. (2012). What Do Middle and High School Students Know About the Particulate Nature of Matter After Instruction? Implications for Practice. *School Science and Mathematics*, 112(2), S. 59-65.
- Bader, H. J. & Lühken, A. (2018). Anforderungen an ein Schulexperiment. *Konkrete Fachdidaktik Chemie. Grundlagen für das Lernen und Lehren im Chemieunterricht*. Sommer, K., Wambach-Laiher, J., Pfeifer, P. (Hrsg.). Aulis, Seelze, 462–466.
- Bamberger, R. (2006). Erfolgreiche Leseerziehung. Theorie und Praxis. München: Domino.
- Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2), 191–215.
- Bangor, A., Kortum, P. T. & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human-Computer Interaction*, 24(6), 574–594.
- Barak, M. & Dori, Y. J. (2005). Enhancing undergraduate students' chemistry understanding through project-based learning in an IT environment. *Science Education*, 89(1), 117–139.
- Barak, M. (2013). Making the unseen seen: Integrating 3D molecular visualizations in elementary, high school, and higher education. In *Pedagogic roles of animations and simulations in chemistry courses*. American Chemical Society, S. 273-291.
- Barak, M., & Dori, Y. J. (2005). Enhancing undergraduate students' chemistry understanding through project-based learning in an IT environment. *Science Education*, 89(1), S. 117-139.
- Barke, H.-D. & Harsch, G. (2001a). *Chemiedidaktik Heute: Lernprozesse in Theorie und Praxis*. Springer Berlin Heidelberg.
- Barke, H.-D. & Harsch, G. (2001b). Der „Horror vacui“ in den Vorstellungen zum Teilchenkonzept. In *Chemiedidaktik Heute* (S. 223–252). Springer, Berlin, Heidelberg.

- Barke, H.-D. (2011). Der „Horror vacui“ in den Vorstellungen vom Teilchenkonzept. In H.-D. H. Barke, *Chemiedidaktik kompakt. Lernprozesse in Theorie und Praxis* (S. 221-244). Berlin, Heidelberg: Springer.
- Barke, H.-D., Harsch, G., Marohn, A. & Krees, S. (2015). *Chemiedidaktik kompakt: Lernprozesse in Theorie und Praxis* (2. Aufl.). Lehrbuch. Springer Berlin Heidelberg; Imprint: Springer Spektrum.
- Barker, V. (2000). *Beyond Appearances: Students' misconceptions about basic chemical ideas*. London: Royal Society of Chemistry.
- Barnett, A. G., van der Pols, J. C. & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International journal of epidemiology*, 34(1), 215–220.
- Barrett, T. J., Stull, A. T., Hsu, T. M., & Hegarty, M. (2015). Constrained interactivity for relating multiple representations in science: When virtual is better than real. *Computers & Education* (81), S. 69–81.
- Baudson, T. G. & Preckel, F. (2016). mini-q: Intelligenzscreening in drei Minuten. *Diagnostica*, 62(3), 182–197.
- Baumann, P. (2006). *Erkenntnistheorie*. Stuttgart: Verlag Metzler.
- Baumert, J., Bos, W. & Lehmann, R. H. (2013). TIMSS/III Dritte Internationale Mathematik und Naturwissenschaftsstudie—Mathematische und naturwissenschaftliche Bildung am Ende der Schul Laufbahn: Band 1 Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Springer-Verlag.
- Baumert, J.; Lehmann, R.; Lehrke, M.; Schmitz, B.; Clausen, M.; Hosenfeld, I.; Köller, O. & Neubrand, J. (2013). TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Opladen: Leske und Budrich.
- Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge.
- Beerenwinkel, A., & Parchmann, I. (2008). Metadiskussionen über Modelle - Historische Aspekte als Impuls. *Praxis der Naturwissenschaften - Chemie in der Schule*, 4, S. 57.
- Behmke, D., Kerven, D., Lutz, R., Paredes, J., Pennington, R., Brannock, E., Deiters, M., Rose, J. & Stevens, K. (2018). Augmented reality chemistry: Transforming 2-D molecular representations into Interactive 3-D structures. In *Proceedings of the Interdisciplinary STEM Teaching and Learning Conference*. 2 (1), 4-12.
- Bennett, J., Lubben, F. & Hogarth, S. (2007). Bringing science to life: A synthesis of the re-search evidence on the effects of context-based and STS approaches to science teaching. *Science Education*, 91(3), 347–370.
- Berk, R. A. & Freedman, D. A. (2003). Statistical Assumptions. *Punishment and social control*, 235.
- Bernholt, S. & Parchmann, I. (2021). Unterrichtsforschung unter fachlichen Perspektiven: Naturwissenschaften. In *Handbuch Schulforschung* (S. 1–24). Springer.
- Bétrancourt, M., & Tversky, B. (2000). Effect of computer animation on users' performance: a review/(Effet de l'animation sur les performances des utilisateurs: une synthèse). *Le travail humain*, 63(4), S. 311.
- Bevan, D. (2023). The Role of Augmented Reality in Enhancing Learning Outcomes in Science Education. *Journal of Informatics Education and Research*, 3(1).
- Birkholz, J., & Elster, D. (2014). Ermittlung von Wissenschaftsverständnis durch Reflexionscafés im basci Schülerlabor. *Erkenntnisweg Biologiedidaktik*, S. 149-163.

- Bittorf, R., Hallier, S., Busch, S., & Sieve, B. (2017). Modellieren mit Linsen und Kichererbsen. Diffusionsvorgänge auf der Teilchenebene visualisieren. *Naturwissenschaften im Unterricht - Chemie*, 28 (160), S. 12-15.
- Blanca, M. J., Alarcón Postigo, R., Arnau Gras, J., Bono Cabré, R. & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, 2017, vol. 29, num. 4, p. 552-557.
- Blankenburg, J. & Scheersoi, A. (2018). Interesse und Interessenentwicklung. In *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 245–259). Springer, Berlin, Heidelberg.
- Blankenburg, J. S., Broman, K., & Parchmann, I. (2016). Interesse an naturwissenschaftlichen Tätigkeiten – Das RIASEC+N Modell. In C. Maurer, Authentizität und Lernen - das Fach in der Fachdidaktik. Gesellschaft für Didaktik der Chemie und Physik (GDChP). Jahrestagung 2015. (S. 476-478). Regensburg: Gesellschaft für Didaktik der Chemie und Physik (GDChP).
- Bogner, A., & Menz, B. (2002). Das theoriegenerierende Experteninterview. Erkenntnisinteresse, Wissensformen, Interaktion. In A. Bogner, B. Littig, & B. Menz, *Das Experteninterview* (S. 33-70). Wiesbaden: Springer Fachmedien.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Springer-Lehrbuch. Springer.
- Brademann, T. (1997). Grundsätze und Beispiele zur Arbeit mit physikalischen Denkmodellen im Physikunterricht der Sekundarstufe I. *Physik in der Schule*, 6, 207–210.
- Brook, A., Briggs, H., & Driver, R. (1984). Aspects of students' understanding of the particulate nature of matter. Leeds: Centre for Studies in Science and Mathematics Education, University of Leeds.
- Brooke, J. (1996). SUS - A quick and dirty usability scale 189(194). *Usability Evaluation in Industry*, 4–7.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.
- Brückmann, M. & Duit, R. (2014). Videobasierte Analyse unterrichtlicher Sachstrukturen. *Methoden in der naturwissenschaftsdidaktischen Forschung*, 189–201.
- Brunner, M., Keller, U., Dierendonck, C., Reichert, M., Ugen, S., Fischbach, A. & Martin, R. (2010). The structure of academic self-concepts revisited: The nested Marsh/Shavelson model. *Journal of Educational Psychology*, 102(4), 964.
- Bruun, A., Law, E. L.-C., Heintz, M. & Alkly, L. H. A. (2016). Understanding the relationship between frustration and the severity of usability problems: What can psychophysiological data (not) tell us? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.
- Buchner, J. & Kerres, M. (2023). Media comparison studies dominate comparative research on augmented reality in education. *Computers & Education*, 104711.
- Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2. Aufl.). PS Psychologie. Pearson Studium.
- Burgoyne, A. P., Mashburn, C. A., Tsukahara, J. S., Hambrick, D. Z. & Engle, R. W. (2023). Understanding the relationship between rationality and intelligence: a latent-variable approach. *Thinking & Reasoning*, 29(1), 1–42.
- Busker, M. (2014). Entwicklung eines Fragebogens zur Untersuchungen des Fachinteresses. *Methoden in der naturwissenschaftsdidaktischen Forschung*, 269–281.
- Calamia, M., Markon, K. & Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, 26(4), 543–570.

- Cao, W. & Yu, Z. (2023). The impact of augmented reality on student attitudes, motivation, and learning achievements — a meta-analysis (2016-2023). *Humanities and Social Sciences Communications*, 10(1), 1–12.
- Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). “An experiment is when you try it and see if it works”: a study of grade 7 students’ understanding of the construction of scientific knowledge. *International Journal of Science Education* 11 (Special Issue), S. 514-529.
- Carmines, E. & Zeller, R. (1979). *Reliability and Validity Assessment*. SAGE Publications, Inc.
- Caspari, I., Weber-Peukert, G. & Graulich, N. (2018). Der Einsatz von Modellen zum Erkenntnisgewinn / Eine Unterrichtseinheit zur Förderung der Modellkompetenz im Kontext „Batterie“ unter explizitem Einbezug von Schülervorstellungen. *CHEMKON*, 25(1), 23–34.
- Cauet, E. (2016). *Testen wir relevantes Wissen? Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten*. Berlin: Logos Verlag Berlin GmbH.
- Chambers, S. K. & Andre, T. (1997). Gender, prior knowledge, interest, and experience in electricity and conceptual change text manipulations in learning about direct current. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 34(2), 107–123.
- Chang, H.-Y., Binali, T., Liang, J.-C., Chiou, G.-L., Cheng, K.-H., Lee, S. W.-Y. & Tsai, C.-C. (2022). Ten years of augmented reality in education: A meta-analysis of (quasi-) experimental studies to investigate the impact. *Computers & Education*, 191, 104641.
- Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. sage.
- Chen, S.-C., Hsiao, M.-S. & She, H.-C. (2015). The effects of static versus dynamic 3D representations on 10th grade students’ atomic orbital mental model construction: Evidence from eye movement behaviors. *Computers in Human Behavior*, 53, 169–180.
- Chen, Y. C. (2006). A study of comparing the use of augmented reality and physical models in chemistry education. *Proceedings of the 2006 ACM international conference on Virtual reality continuum and its applications*, 369-372.
- Cheng, K.-H. (2018). Surveying Students’ Conceptions of Learning Science by Augmented Reality and their Scientific Epistemic Beliefs. *EURASIA Journal of Mathematics, Science and Technology Education*, 14(4).
- Chiu, M.-H. & Lin, J.-W. (2019). Modeling competence in science education. *Disciplinary and Interdisciplinary Science Education Research*, 1(1), 1–11.
- Christen, H. R. (1990). *Chemieunterricht. Eine praxisorientierte Didaktik*. Basel, Boston, Berlin: Birkhäuser.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4), 284.
- Clement, J. (2000). Model based learning as a key research area for science education. *International Journal of Science Education*, 22(9), 1041–1053.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Edition) [s.n.].
- Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.
- Cook, R. D. & Weisberg, S. (1982). Criticism and influence analysis in regression. *Sociological methodology*, 13, 313–361.

- Cook, T. D., Campbell, D. T. & Day, A. (1979). Quasi-experimentation: Design & analysis issues for field settings (Bd. 351). Houghton Mifflin Boston.
- Cordova, D. I. & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88(4), 715.
- Crawford, B. A., & Cullin, M. (2005). Dynamic assessments of preservice teachers' knowledge of models and modelling. In K. Boersma, M. Goedhart, O. de Jong, & H. Eijkelhof, Research and the quality of science education (S. 309-323). Dordrecht: Springer.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Daniels, Z. (2008). Entwicklung schulischer Interessen im Jugendalter. Waxmann Verlag.
- De Jong, O. & Taber, K. S. (2013). Teaching and learning the many faces of chemistry. In *Handbook of research on science education* (S. 631–652). Routledge.
- Deary, I. J., Strand, S., Smith, P. & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1), 13–21.
- Deci, E. L. & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, 19(2), 109–134.
- Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik*(39), 223–238.
- Di Serio, Á., Ibáñez, M. B. & Kloos, C. D. (2013). Impact of an augmented reality system on students' motivation for a visual art course. *Computers & Education*, 68, 586–596.
- Dickhäuser, O. (2001). Computernutzung und Geschlecht. Waxmann Verlag.
- Dickhäuser, O. (2006). Fähigkeitsselbstkonzepte: Entstehung, Auswirkung, Förderung (1/2). Verlag Hans Huber, Hogrefe AG. *Zeitschrift für pädagogische Psychologie*, 20.
- Digel, S. (2020). Messung von Modellierungskompetenz in Physik: Theoretische Herleitung und empirische Prüfung eines Kompetenzmodells physikspezifischer Modellierungskompetenz. LOGOS Verlag BERLIN.
- DIN EN (2018-03). ISO 9241-11:2018-03.
- DIN EN (2020-05). ISO 9241-110:2020-05.
- Dirks, U., & Knobloch, E. (2008). Modelle. Probleme und Perspektiven. In U. Dirks, & E. Knobloch, Modelle (S. 9-28). Frankfurt am Main: Peter Lang Verlag.
- Döring, N. & Bortz, J. (2016). Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Springer. Berlin Heidelberg.
- Dresing, T., & Pehl, T. (2015). Praxisbuch Interview, Transkription & Analyse. Anleitungen und Regelsysteme für qualitativ Forschende. 6. Auflage. Marburg: Quelle: www.audiotranskription.de/praxisbuch (Datum des Downloads: 20.10.2019).
- Driver, R., Guesne, E., & Tiberghien, A. (1985). Children's Ideas in Science. Milton Keynes: Open University Press.
- Duit, R. (1995). Zur Rolle der konstruktivistischen Sichtweise in der naturwissenschaftsdidaktischen Lehr- und Lernforschung. *Zeitschrift für Pädagogik*, 41 (6), 905-923.
- Duit, R. (1995a). Conceptual Change in Science Education. In M. Carretero, W. Schnotz, & S. Vosniadou, Conceptual Change. London: Lawrence Earlbaum.

- Dündar, H. & Akçayır, M. (2014). Implementing tablet PCs in schools: Students' attitudes and opinions. *Computers in Human Behavior*, 32, 40–46.
- Dunn, T. J., Baguley, T. & Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *British journal of psychology*. (London, England), 105(3), 399–412.
- Eccles, J. S. & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, 61, 101859.
- Eckstein, P. P. (2012). *Angewandte Statistik mit SPSS: Praktische Einführung für Wirtschaftswissenschaftler* (7. Aufl.). SpringerLink: Bücher. Gabler Verlag.
- Edelmann, W. (2003). Intrinsische und extrinsische Motivation. *Grundschule*, 4, 30–32.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2015). *Statistik und Forschungsmethoden: Mit Online-Materialien* (5., korrigierte Auflage). Beltz.
- Eilks, I. & Bindernagel, J. A. (2008). Modelle und Modelldenken im Chemieunterricht. MNU-Dokumentation-15. Fachleitertagung Chemie.
- Eilks, I. & Ralle, B. (2002). Partizipative Fachdidaktische Aktionsforschung. Ein Modell für eine begründete und praxisnahe curriculare Entwicklungsforschung in der Chemiedidaktik. In CHEMKON: Forum für Unterricht und Didaktik. Symposium im Rahmen der Tagung von Wiley Online Library.
- Elford, D., Lancaster, S. J. & Jones, G. A. (2022). Exploring the effect of augmented reality on cognitive load, attitude, spatial ability, and stereochemical perception. *Journal of Science Education and Technology*, 31(3), 322–339.
- Elster, D. (2005). Was macht naturwissenschaftlichen Unterricht für Mädchen und für Buben interessant? In *Begabungsförderung durch Geschlechtssensibilität in Mathematik, Naturwissenschaften und Technik* (S. 95-103). Wien: Bundesministerium für Bildung, Wissenschaft und Kultur, Abteilung für Gender Mainstreaming und geschlechtsspezifische Bildungsfragen (Z/12) sowie Abteilung für Schulwissenschaften und Kulturpädagogik (I/4)/Referat für Begabungs- und Kreativitätsförderung.
- Elster, D. (2007). Zur Diskussion gestellt. Interessante und weniger interessante Kontexte für das Lernen von Naturwissenschaften. *Mathematische und Naturwissenschaftliche Unterricht*, 60 (4), S. 243-248.
- Engl, A. (2020). *CHEMIE PUR – Unterrichten in der Natur: Entwicklung und Evaluation eines kontextorientierten Unterrichtskonzepts im Bereich Outdoor Education zur Änderung der Einstellung zu „Chemie und Natur“*. LOGOS Verlag BERLIN.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge: MA: MIT Press.
- Eschenhagen, D., Kattmann, U., & Rodi, D. (2008). *Fachdidaktik Biologie*. Köln: Aulis Verlag Deubner.
- Etikan, I., Musa, S. A. & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1), 1–4.
- Everitt, B. S. & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics* (4th ed.). Cambridge University Press.
- Fabrigar, L. R. & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University Press.
- Feldon, D. F., Callan, G., Juth, S. & Jeong, S. (2019). Cognitive load as motivational cost. *Educational Psychology Review*, 31, 319–337.

- Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. Sage publications.
- Finkel, D., Reynolds, C. A., McArdle, J. J. & Pedersen, N. L. (2007). Age changes in processing speed as a leading indicator of cognitive aging. *Psychology and aging*, 22(3), 558.
- Finney, S. J. & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. *Structural equation modeling: A second course*, 10(6), 269–314.
- Fischer, M., Grube, G. & Reisin, F.-M. (1995). KIT REPORT 125 Abbild oder Konstruktion-Modellierungsperspektiven in der Informatik.
- Fischler, H., & Lichtfeldt, M. (1997). Teilchen und Atome. Modellbildung im Unterricht. *Naturwissenschaften im Unterricht - Physik*, 180-184.
- Fleige, J., Seegers, A., Upmeyer zu Belzen, A., & Krüger, D. (2012). Förderung von Modellkompetenz im Biologieunterricht. *Der Mathematische und Naturwissenschaftliche Unterricht*, 65(1), 19, 7-10.
- Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3), 613–619.
- Flick, U. (2017). *Qualitative Sozialforschung: Eine Einführung* (8. Aufl.). Rororo Rowohlt's Enzyklopädie: Bd. 55694. Rowohlt Taschenbuch Verlag. Hamburg.
- Flint, A. (2014). Vom didaktischen Konzept zur Unterrichtseinheit. In D. Krüger, I. Parchmann, & H. Schecker, *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 57-66). Berlin, Heidelberg: Springer.
- Forero, C. G., Maydeu-Olivares, A. & Gallardo-Pujol, D. (2009). Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(4), 625–641.
- Fred Paas, Juhani E. Tuovinen, Huib Tabbers & Pascal W. M. Van Gerven. (2016). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. In *Cognitive Load Theory* (S. 63–71). Routledge.
- Freyer, K. (2013). Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemesterstudierender im Fach Chemie (Bd. 156). Logos Verlag Berlin GmbH.
- Funke, J., & Spering, M. (2006). Methoden der Denk- und Problemlöseforschung. In J. Funke, *Enzyklopädie der Psychologie. Denken und Problemlösen. Kognition*, Bd C/II/8 (S. 647-726). Göttingen: Hogrefe.
- Gabel, D. (1999). Improving teaching and learning through chemistry education research: A look to the future. *Journal of Chemical education*, 76(4), S. 548-554.
- Gardner, P. L. (1998). The development of males' and females' interests in science and technology. *Seon Conference on Interest and Gender*.
- Garner, R., Brown, R., Sanders, S. & Menke, D. J. (1992). Seductive details" and learning from text. The role of interest in learning and development, 239–254.
- Garzón, J. & Acevedo, J. (2019). Meta-analysis of the impact of Augmented Reality on students' learning gains. *Educational Research Review*, 27, 244–260.
- Garzón, J., Pavón, J. & Baldiris, S. (2019). Systematic review and meta-analysis of augmented reality in educational settings. *Virtual Reality*, 23(4), 447–459.
- Gehlen, C. (2016). *Kompetenzstruktur naturwissenschaftlicher Erkenntnisgewinnung im Fach Chemie*. Berlin: Logos.
- Gentner, D. & Stevens, A. L. (2014). *Mental models*. Psychology Press.

- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170.
- George, D. & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference; 11.0 update (4th ed.)*. Allyn & Bacon.
- Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese*, 172(2), 269–281.
- Gilbert, J. K. & Justi, R. (2016). *Modelling-based teaching in science education (Bd. 9)*. Springer.
- Gilbert, J. K. (2004). Models and modelling: Routes to more authentic science education. *International Journal of Science and Mathematics Education*, 2, 115–130.
- Gilbert, S. W. (1991). Model building and a definition of science. *Journal of Research in Science Teaching*, 28(1), S. 73-79.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago: Aldine.
- Gläser, J. & Laudel, G. (2012). *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen*. Lehrbuch VS, Verl. für Sozialwiss.
- Gogolin, S. & Krüger, D. (2018). Students' understanding of the nature and purpose of models. *Journal of Research in Science Teaching*, 55(9), 1313–1338.
- Gouvea, J. & Passmore, C. (2017). 'Models of' versus 'Models for'. *Science & Education*, 26(1-2), 49–63.
- Graf, E. (2002). Modelle im Chemieunterricht. *Naturwissenschaften im Unterricht: Chemie*, 67, S. 4-6.
- Greca, I. M. & Moreira, M. A. (2000). Mental models, conceptual models, and modelling. *International Journal of Science Education*, 22(1), 1–11.
- Greca, I. M. & Moreira, M. A. (2002). Mental, physical, and mathematical models in the teaching and learning of physics. *Science Education*, 86(1), 106–121.
- Gropengießer, H. & Marohn, A. (2018). Schülervorstellungen und Conceptual Change. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 49–67). Springer Berlin Heidelberg.
- Gropengießer, H., & Kattmann, U. (2013). Didaktische Rekonstruktion. In H. Gropengießer, U. Harms, & U. Kattmann, *Fachdidaktik Biologie* (S. 16-23). Hallbergmoos: Aulis Verlag.
- Grosslight, L., Unger, C., Jay, E., & Smith, C. (1991). Understanding Models and their Use in Science: Conceptions of Middle and High School Students and Experts. *Journal of Research in Science Teaching*, 28, S. 799-822.
- Gruber, H. & Stamouli, E. (2015). Intelligenz und Vorwissen. In E. Wild & J. Möller (Hrsg.), *Springer-Lehrbuch. Pädagogische Psychologie* (S. 25–44). Springer Berlin Heidelberg.
- Günther, J. (2006). *Lehrerfortbildung über die Natur der Naturwissenschaften: Studien über das Wissenschaftsverständnis von Grundschullehrkräften*. Logos-Verlag.
- Haas, J.-B. & Marohn, A. (2022). The teaching concept chem:LEVEL – promoting technical language on the basis of the Johnstone triangle. *CHEMKON*, 29(S1), 213–217.
- Hair, J., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham, R. L. (2010). *Multivariate Data Analysis*. 6th (ed.) Upper Saddle River NJ. Prentice-Hall.
- Halloun, I. (2001). *Student views about science: a comparative survey*. Beirut: Lebanon: Educational Research Centre, Lebanese University.

- Hammann, M. & Asshoff, R. (2023). Schülervorstellungen im Biologieunterricht: Ursachen für Lernschwierigkeiten. Klett / Kallmeyer.
- Hammerich, P. (1998). Confronting Students' Conceptions of the Nature of Science with cooperative Controversy. In W. F. McComas, *The Nature of science in Science Education. Rationales and Strategies* (S. 127-136). New York, Boston, Dordrecht, London, Moscow: Kluwer Academic Publishers.
- Hansen-Schirra, S. & Gutermuth, S. (2019). Empirische Überprüfung von Verständlichkeit. Eds. C. Maaß, and I. Rink. *Handbuch Barrierefreie Kommunikation*. Berlin: Frank & Timme, 163–182.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M. & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology*, 92(2), 316–330.
- Harp, S. F., & Mayer, R. E. (1997). The role of interest in learning from scientific text and illustrations: On the distinction between emotional interest and cognitive interest. *Journal of educational psychology*, 89 (1), S. 92.
- Harrison, A. G., & Treagust, D. F. (1996). Secondary Students' Mental Models of Atoms and Molecules. Implications for Teaching Chemistry. *Science Education*, 80(5), 509-534.
- Harrison, A. G., & Treagust, D. F. (2000). A typology of school science models. *International Journal of Science Education* 22 (9), 1011-1026.
- Harter, S. (1981). A new self-report scale of intrinsic versus extrinsic orientation in the classroom: Motivational and informational components. *Developmental psychology*, 17(3), 300.
- Härtig, H., Ostermann, A., Ropohl, M., Schwanewedel, J., Kampschulte, L. & Lindmeier, A. (2021). Gibt es einen fachspezifischen Medieneinsatz im naturwissenschaftlichen Fachunterricht? Ergebnisse einer Fragebogenerhebung. *Zeitschrift für Didaktik der Naturwissenschaften*, 27(1), 139–154.
- Hattie, J., Beywl, W. & Zierer, K. (2013). *Lernen sichtbar machen*. Schneider-Verl. Hohengehren.
- Hauke, J. & Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *QUAGEO*, 30(2), 87–93.
- Hayes, A. F. & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But... *Communication Methods and Measures*, 14(1), 1–24.
- Held, C., Gottfried, M., Vosgerau, G. V. & Knauff, M. (2006). *Mental models and the mind: current developments in Cognitive Psychology, Neuroscience and Philosophy of Mind*. Elsevier.
- Henze, I., van Driel, J., & Verloop, N. (2007). Science teachers' knowledge about teaching models and modelling in the context of a new syllabus on public understanding of science. *Research in Science Education*, 37, S. 99-122.
- Hidi, S. & Renninger, K. A. (2006). The Four-Phase Model of Interest Development. *Educational Psychologist*, 41(2), 111–127.
- Hidi, S., Renninger, K. A. & Krapp, A. (2004). Interest, a motivational variable that combines affective and cognitive functioning. *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development*, 89, 115.
- Hillmayr, D., Ziernwald, L. Reinhold, F., Hofer, S. I., & Reiss, K. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education*, 153, S. 103897.

- Hodson, D. (1992). In search for meaningful relationship: An exploration of some issues relating to integration in science and science education. *International Journal of Science Education* 14, S. 541-562.
- 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.
- Höffler, T. N. & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, 17(6), 722–738.
- Höffler, T. N. (2008). Lernen mit dynamischen Visualisierungen: Metaanalyse und experimentelle Untersuchungen zu einem naturwissenschaftlichen Lerninhalt. Essen: Universität Duisburg-Essen. [Elektronische Dissertation].
- Höffler, T. N., & Schwartz, R. N. (2011). Effects of pacing and cognitive style across dynamic and non-dynamic representations. *Computers & Education*, 57(2), S. 1716-1726.
- Hoffmann, L., & Lehrke, M. (1986). Die Untersuchung über Schülerinteressen in Physik und Technik (32). *Zeitschrift für Pädagogik*, S. 189-204.
- Hoffmann, L., Häußler, P. & Lehrke, M. (1998). Die IPN-Interessenstudie Physik. IPN.
- Hopf, C. (1995). Befragungsverfahren. In U. Flick, E. v. Kardorff, H. Keupp, L. v. Rosenstiel, & S. Wolff, *Handbuch Qualitative Sozialforschung*. 2. Auflage (S. 177-188). Weinheim: Beltz Psychologie Verlags Union.
- Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, 57(5), 253.
- Hornecker, E., & Dünser, A. (2007). Supporting Early Literacy with Augmented Books - Experiences with an Exploratory Study. *Informatik 2007 – Informatik trifft Logistik – (1)*, S. 555-559.
- Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Huynh, H. & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of educational statistics*, 1(1), 69–82.
- Ibáñez, M. B., Di Serio, Á., Villarán, D., & Kloos, C. D. (2014). Experimenting with electromagnetism using augmented reality: Impact on flow student experience and educational effectiveness. *Computers & Education*, 71, S. 1-13.
- International Test Commission. (2017). The ITC guidelines for translating and adapting tests (Second Edition).
- Jansen, M., Schroeders, U. & Stanat, P. (2013). Motivationale Schülermerkmale in Mathematik und den Naturwissenschaften. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *Empirische Erziehungswissenschaft 2013/14. IQB-Ländervergleich 2012* (S. 348–365). Waxmann Verlag.
- Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, 4(1), 71–115.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.
- Johnson-Laird, P. N. (2005). Mental models and thought. *The Cambridge handbook of thinking and reasoning*, 185–208.
- Johnstone, A. H. (1999). The nature of chemistry. *Education in chemistry*, 36(2), 45–47.

- Jong, J.-P., Chiu, M.-H. & Chung, S.-L. (2015). The use of modeling-based text to improve students' modeling competencies. *Science Education*, 99(5), 986–1018.
- Jonkisz, E., Moosbrugger, H., & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger, & K. A., *Testtheorie und Fragebogenkonstruktion* (S. 27-72). Berlin, Heidelberg: Springer-Verlag.
- Jung, W. (1977). Fachdidaktische Aspekte von Theoriebegriffen. *chimica didacticae* 3(4), S. 153-168.
- Justi, R. S., & Gilbert, J. K. (2002). Modelling, teachers' views on the nature of modelling, and implications for the education of modellers. *International Journal of Science Education*, 24:4, S. 369-387.
- Justi, R. S., & Gilbert, J. K. (2003). Teacher's views on the nature of models. *International Journal of Science Education*, 25, S. 1369-1386.
- Kalemkuş, J. & Kalemkuş, F. (2022). Effect of the use of augmented reality applications on academic achievement of student in science education: meta analysis review. *Interactive Learning Environments*, 1–18.
- Kalyuga, S. (2010). Schema Acquisition and Sources of Cognitive Load. In J. Plass, R. Moreno, & R. Brünken, *Cognitive load theory* (S. 48-64). New York: Cambridge University Press.
- Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of Expertise and Instructional Design. *Human Factors*, 40(1), S. 1-17.
- Kane, M. J., Hambrick, D. Z. & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005).
- Karapanos, M., Becker, C. & Christophel, E. (2018). Die Bedeutung der Usability für das Lernen mit digitalen Medien. *Medienpädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, 36–57.
- Karplus, R. (1996). Chemical phenomena in elementary school science. *Journal of Chemical Education*, 43(5), S. 267-269.
- Kelava, A. & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Springer-Lehrbuch. Testtheorie und Fragebogenkonstruktion* (S. 75–102). Springer Berlin Heidelberg.
- Keller, G. (1977). *Über das Denken in Modellen*. Frankfurt am Main, Aarau: Verlag Moritz Diesterweg / Otto Salle Verlag; Verlag Sauerländer AG.
- Keller, S., Habig, S., & Rumann, S. (2020). Fostering internal mental model-construction through Augmented Reality while learning Organic Chemistry. In S. Frodsham, *European Science Education Research Association. Virtual Doctoral Network 2020. Book of Synopses* (S. 349-356). Oxford.
- Kircher, E. (1977). Einige erkenntnistheoretische und wissenschaftstheoretische Auffassungen zur Fachdidaktik. *Chimica Didactica* 3, 61-79.
- Kircher, E. (1995). *Studien zur Physikdidaktik: Erkenntnis- und wissenschaftstheoretische Grundlagen*. Kiel: IPN.
- Kircher, E. (2015). Modellbegriff und Modellbildung in der Physikdidaktik. In E. Kicher, R. Gierwidz, & P. Häußler, *Physikdidaktik. Theorie und Praxis* (S. 783-875). Berlin Heidelberg: Springer Spektrum.
- Kircher, E., & Dittmer, A. (2004). Lehren und Lernen über die Natur der Naturwissenschaften - ein Überblick. In C. Hößler, D. Höttecke, & E. Kircher, *Lehren und Lernen über die Natur der Naturwissenschaften* (S. 2-22). Baltmannsweiler: Schneider Verlag.

- Kirman Bilgin, A., Demircioğlu Yürükel, F., & Yiğit, N. (2017). The Effect of a Developed REACT Strategy on the Conceptual Understanding of Students: "Particulate Nature of Matter". *Journal of Turkish Science Education (TUSED)*, 14 (2), 65-81.
- Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in psychology* 8:1997, 1-17.
- Klieme, E., Funke, J., Leutner, D., Reimann, P., & Wirth, J. (2001). Problemlösen als fächerübergreifende Kompetenz. Konzeption und erste Resultate aus einer Schulleistungsstudie. *Zeitschrift für Pädagogik* 47, 179-200.
- Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.
- Köhler, K., & Meisert, A. (2012). Welche Erkenntnismethoden sind für den Biologieunterricht relevant? In U. Spörhase, *Biologie Didaktik. Praxishandbuch für die Sekundarstufe I und II* (S. 130-151). Berlin: Cornelsen.
- Köller, O., Daniels, Z., Schnabel, K. U. & Baumert, J. (2000). Kurswahlen von Mädchen und Jungen im Fach Mathematik: zur Rolle von fachspezifischem Selbstkonzept und Interesse. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology*.
- Köller, O., Trautwein, U., Lüdtke, O. & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe. *Zeitschrift für pädagogische Psychologie*, 20(1/2), 27–39.
- Konrad, K. (2010). Lautes Denken. In G. Mey, & K. (. Mruck, *Handbuch Qualitative Forschung in der Psychologie* (S. 476-490). Wiesbaden: VS Verlag.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2), 155-163.
- Korakakis, G., Pavlatou, E. A., Palyvos, J. A. & Spyrellis, N. (2009). 3D visualization types in multimedia applications for science learning: A case study for 8th grade students in Greece. *Computers & Education*, 52(2), 390–401.
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher*, 49(4), 241–253.
- Krapp, A. & Lewalter, D. (2001). Development of interests and interest-based motivational orientations: A longitudinal study in vocational school and work settings. *Motivation in learning contexts: Theoretical advances and methodological implications*, 209–232.
- Krapp, A. & Prenzel, M. (2011). Research on Interest in Science: Theories, methods, and findings. *International Journal of Science Education*, 33(1), 27–50.
- Krapp, A. & Ryan, R. M. (2002). Selbstwirksamkeit und Lernmotivation. In D. Hopf & M. Jerusalem (Hrsg.), *Zeitschrift für Pädagogik Beiheft: Bd. 44. Selbstwirksamkeit und Motivationsprozesse in Bildungsinstitutionen* (Bd. 44, S. 54–82). Beltz.
- Krapp, A. (1992). Interesse, Lernen und Leistung. *Neue Forschungsansätze in der Pädagogischen Psychologie. Zeitschrift für Pädagogik*, 38(5), 747–770.
- Krapp, A. (1992). Interesse, Lernen und Leistung. *Neue Forschungsansätze in der Pädagogischen Psychologie. Zeitschrift für Pädagogik*, 38(5), S. 747-770.
- Krapp, A. (1998). Entwicklung und Förderung von Interessen im Unterricht. *Psychologie in Erziehung und Unterricht (PEU)*, 45(3), 186–203.
- Krapp, A. (2002a). An educational-psychological theory of interest and its relation to self-determination theory. In R. M. Ryan & E. L. Deci (Hrsg.), *Handbook of self-determination research* (S. 405–427). University of Rochester Press.

- Krapp, A. (2002b). Structural and dynamic aspects of interest development: theoretical considerations from an ontogenetic perspective. *Learning and Instruction*, 12(4), 383–409.
- Krapp, A. (2005). Basic needs and the development of interest and intrinsic motivational orientations. *Learning and Instruction*, 15(5), 381–395.
- Krapp, A., & Prenzel, M. (2011). Research on Interest in Science: Theories, methods, and findings. *International Journal of Science Education*, 33 (1), S. 27–50.
- Krell, M., Upmeyer zu Belzen, A., & Krüger, D. (2014). Student's levels of understanding models and modelling in biology: Global or aspect-dependent? *Research in Science Education*, 44, S. 109–132.
- Krell, M., Upmeyer zu Belzen, A., & Krüger, D. (2016). Modellkompetenz im Biologieunterricht. In A. Sandmann, & P. Schmiemann, *Biologiedidaktische Forschung: Schwerpunkte und Forschungsstände* (S. 83-106). Berlin: Logos Verlag.
- Krell, M., Walzer, C., Hergert, S. & Krüger, D. (2017). Development and application of a category system to describe pre-service science teachers' activities in the process of scientific modelling. *Research in science education*, 49, 1319–1345.
- Krosnick, J., Presser, S. & Building, A.-S. (2009). Question and Questionnaire Design. *Handbook of Survey Research*.
- Krüger, D. & Upmeyer zu Belzen, A. (2021). Kompetenzmodell der Modellierkompetenz – Die Rolle abduktiven Schließens beim Modellieren. *Zeitschrift für Didaktik der Naturwissenschaften*, 27(1), 127–137.
- Krüger, D. (2007). Die Conceptual Change-Theorie. In D. Krüger & H. Vogt, *Theorien in der Biologiedidaktischen Forschung - Ein Handbuch für Lehramtssudenten und Doktoranden* (S. 81-92). Berlin, Heidelberg: Springer.
- Krüger, D., Kauertz, A. & Upmeyer zu Belzen, A. (2018). Modelle und das Modellieren in den Naturwissenschaften. *Theorien in der naturwissenschaftsdidaktischen Forschung*, 141–157.
- Kuckartz, U. (2016). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*. 3., überarbeitete Auflage. Weinheim: Beltz Juventa.
- Kuckartz, U., Dresing, T., Rädiker, S. & Stefer, C. (2008). *Qualitative Evaluation. Der Einstieg in die Praxis*. Wiesbaden: VS-Verlag.
- Kultusministerkonferenz [KMK] (2005). Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004. Von kmk.org: https://www.kmk.org/fileadmin/Daten/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Chemie.pdf abgerufen
- Kultusministerkonferenz [KMK] (2024). Weiterentwickelte Bildungsstandards in den Naturwissenschaften für das Fach Chemie (MSA) (Beschluss der KMK vom 16.12.2004 i.d.F. vom 13.06.2024). Berlin: Sekretariat der Kultusministerkonferenz.
- Lang, V. (2023). Förderung der Modellbildungskompetenz im Chemieunterricht durch Anwendung eines Modellierungsprozesses unter Integration von Differenzierungs- und Digitalisierungsaspekten. *Saarländische Universitäts- und Landesbibliothek*.
- Ledesma, R. D., Ferrando, P. J., Trógo, M. A., Poó, F. M., Tosi, J. D. & Castro, C. (2021). Exploratory factor analysis in transportation research: Current practices and recommendations. *Transportation Research Part F: Traffic Psychology and Behaviour*, 78, 340–352.

- Lehnert, H.-J., & Köhler, K. (2012). Welche Medien werden im Biologieunterricht genutzt? In U. Spörhase, *Biologiedidaktik. Praxishandbuch für die Sekundarstufe I und II* (S. 152-174). Berlin: Cornelsen.
- Lehrer, R. & Schauble, L. (2015). The development of scientific thinking. *Handbook of child psychology and developmental science*, 2(7), 671–714.
- Lehrer, R., Horvath, J. & Schauble, L. (1994). Developing Model-Based Reasoning. *Interactive Learning Environments*, 4(3), 218–232.
- Lehrl, S., Zipp, A., Schwarzfischer, C. & Eissing, G. (2016). Kurztest für allgemeine Intelligenz (KAI). <https://doi.org/10.17877/DE290R-17112>
- Leisner, A. (2005a). *Entwicklung von Modellkompetenz im Physikunterricht*. Berlin: Logos Verlag.
- Leisner, A. (2005b). *Modellkompetenz im Physikunterricht*. LLF-Berichte/Universität Potsdam, Zentrum für Lehrerbildung, 20, S. 35-49.
- Leisner-Bodenthin. (2006). Zur Entwicklung von Modellkompetenz im Physikunterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, 12, S. 91–108.
- Lenhard, W. & Lenhard, A. (2011). Berechnung des Lesbarkeitsindex LIX nach Björnson. <https://doi.org/10.13140/RG.2.1.1512.3447>
- Leppink, J., Paas, F., van der Vleuten, C. P. M., van Gog, T. & van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior research methods*, 45(4), 1058–1072.
- Lewalter, D. (1997a). Kognitive Informationsverarbeitung beim Lernen mit computerpräsentierten statischen und dynamischen Illustrationen. *Unterrichtswissenschaft*, 25(3), 207–222.
- Lewalter, D. (1997b). Lernen mit Bildern und Animationen: Studie zum Einfluss von Lernermerkmalen auf die Effektivität von Illustrationen. Waxmann.
- Lewis, J. R. & Sauro, J. (2009). The factor structure of the system usability scale. In *Human Centered Design: First International Conference, HCD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009 Proceedings 1. Symposium im Rahmen der Tagung von Springer*.
- Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological methods*, 21(3), 369–387.
- Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Psychologie Verlags Union.
- Little, R. J. A. & Rubin, D. B. (2019). *Statistical analysis with missing data* (Bd. 793). John Wiley & Sons.
- Liu, Q., Ma, J., Yu, S., Wang, Q. & Xu, S. (2022). Effects of an Augmented Reality-Based Chemistry Experiential Application on Student Knowledge Gains, Learning Motivation, and Technology Perception. *Journal of Science Education and Technology*. Advance online publication.
- Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R. & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological methods*, 22(3), 486.
- Louca, L. T. & Zacharia, Z. C. (2012). Modeling-based learning in science education: cognitive, metacognitive, social, material and epistemological contributions. *Educational Review*, 64(4), 471–492.
- Luhmann, M. (2020). *R für Einsteiger: Einführung in die Statistik-Software für die Sozialwissenschaften: mit Online-Material* (5., überarbeitete Auflage). Beltz.
- MacCallum, R. C., Browne, M. W. & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, 1(2), 130.

- Mahr, B. (2008a). Ein Modell des Modellseins - Ein Beitrag zur Aufklärung des Modellbegriffs. In U. Dirks, & E. Knobloch, Modelle. Berlin: Peter Lang Verlag, 187-220
- Mahr, B. (2008b). Cargo. Zum Verhältnis von Bild und Modell. Wilhelm Fink Verlag, München.
- Mahr, B. (2015). Modelle und ihre Befragbarkeit Grundlagen einer allgemeinen Modelltheorie. Erwägen Wissen Ethik, 26(3).
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. Journal of Educational Psychology, 79(3), 280.
- Marsh, H. W. (1990). The structure of academic self-concept: The Marsh/Shavelson model. Journal of Educational Psychology, 82(4), 623.
- Marsh, H. W., Hau, K.-T. & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Structural equation modeling, 11(3), 320–341.
- Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S. & Davier, M. von (2013). Why item parcels are (almost) never appropriate: two wrongs do not make a right--camouflaging misspecification with item parcels in CFA models. Psychological methods, 18(3), 257–284.
- Marshall, M. N. (1996). Sampling for qualitative research. Family Practice, 13, S. 522-525.
- Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger, & H. (.) Vogt, Theorien in der biologiedidaktischen Forschung (S. 177-186). Berlin, Heidelberg: Springer.
- Mayer, R. E. & Gallini, J. K. (1990). When is an illustration worth ten thousand words? Journal of Educational Psychology, 82(4), 715–726.
- Mayer, R. E. & Pilegard, C. (2005). Principles for managing essential processing in multimedia learning: Segmenting, pretraining, and modality principles. The Cambridge handbook of multimedia learning, 169–182.
- Mayer, R. E. (2009). Multimedia learning (Second edition). Cambridge University Press.
- Mayer, R. E. (2014a). Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. The Cambridge handbook of multimedia learning, 16, 345–370.
- Mayer, R. E. (Hg.). (2014). Cambridge handbooks in psychology. The Cambridge handbook of multimedia learning. Cambridge University Press.
- Mayer, R. E., & Moreno, R. (2003). Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning. In J. L. Plass, R. Moreno, & R. Brünken, Cognitive Load Theory (S. 131-152). Cambridge, NY: Cambridge University Press.
- Mayer, R. E., Hegarty, M., Mayer, S., & Campbell, J. (2005). When static media promote active learning: Annotated illustrations versus narrated animations in multimedia instruction. Journal of Experimental Psychology: Applied, 11(4), 256-265.
- Mayes, J. T. & Fowler, C. J. (1999). Learning technology and usability: a framework for understanding courseware. Interacting with computers, 11(5), 485–497.
- Mayring, P. & Fenzl, T. (2019). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Hrsg.), Handbuch Methoden der empirischen Sozialforschung (S. 633–648). Springer Fach-medien Wiesbaden.
- Mayring, P. (2010). Qualitative Inhaltsanalyse. Beltz Verlagsgruppe.
- Mayring, P. (2019). Qualitative Inhaltsanalyse – Abgrenzungen, Spielarten, Weiterentwicklungen. Advance online publication (Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, Vol 20, No 3 (2019): Qualitative Content Analysis I).

- Mazzuco, A., Krassmann, A. L., Reategui, E. & Gomes, R. S. (2022). A systematic review of augmented reality in chemistry education. *Review of Education*, 10(1), Artikel e3325.
- McComas, W. F., Clough, M. P. & Almazroa, H. (1998). The Role and Character of the Nature of Science in Science Education. In W. F. McComas (Hrsg.), *Science and technology education library: Bd. 5. The nature of science in science education: Rationales and strategies* (S. 3–39). New York, Bosten, Dordrecht, London, Moscow: Kluwer Academic.
- Medienpädagogischer Forschungsverbund Südwest (2022). JIM-Studie 2022. Jugend, Information, Medien: Basisuntersuchung zum Medienumgang 12- bis 19- jähriger.
- Meiselwitz, G. & Sadera, W. A. (2008). Investigating the connection between usability and learning outcomes in online learning environments. *Journal of Online Learning and Teaching*, 4(2), 9.
- Meisert, A. (2008). Vom Modellwissen zum Modellverständnis - Elemente einer umfassenden Modellkompetenz und deren Fundierung durch lernerseitige Kriterien zur Klassifikation von Modellen. *Zeitschrift für Didaktik der Naturwissenschaften*, 14, 243-261.
- Meisert, A. (2009). Modelle in der Biologie: Wie lässt sich im Unterricht ein Verständnis für ihre Bedeutung fördern? *Der mathematische und naturwissenschaftliche Unterricht* 62 (7), 424–430.
- Meisert, A. (2014). Mit Modellen arbeiten. In S. S. Holger Weitzel, *Biologie unterrichten: planen, durchführen, reflektieren* (S. 105-117). Berlin: Cornelsen Schulverlage.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- Meyling, H. (1990). *Wissenschaftstheorie im Physikunterricht der Gymnasialen Oberstufe. Das wissenschaftstheoretische Vorverständnis und der Versuch seiner Veränderung durch expliziten wissenschaftstheoretischen Unterricht.* . Bremen: Universität Bremen.
- Mierdel, J. & Bogner, F. X. (2019). Comparing the Use of Two Different Model Approaches on Students' Understanding of DNA Models. *Education Sciences*, 9(2), 115.
- Mikelskis-Seifert, S. & Leisner, A. (2003). Das Denken in Modellen fördern. Ein Unterrichtsbeispiel zur Entwicklung von Teilchenvorstellungen. *Naturwissenschaften im Unterricht, Physik*, 71, 32–34.
- Mikelskis-Seifert, S. (2002). Die Entwicklung von Metakzepten zur Teilchenvorstellung bei Schülern. Untersuchung eines Unterrichts über Modelle mithilfe eines Systems multipler Repräsentationsebenen. Berlin: Logos.
- Miller, C. J., Smith, S. N. & Pugatch, M. (2020). Experimental and quasi-experimental designs in implementation research. *Psychiatry research*, 283, 112452.
- Ministerium für Bildung, Wissenschaft, Weiterbildung und Kultur. (2014). *Lehrpläne für die naturwissenschaftlichen Fächer für die weiterführenden Schulen in Rheinland-Pfalz: Biologie Chemie Physik. Klassenstufen 7 bis 9/10.*
- Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology*, 85(3), 424–436.
- Mittelstraß, J. (2005). Anmerkungen zum Modellbegriff. In B.-B. A. *Wissenschaften, Modelle des Denkens: Streitgespräch in der Wissenschaftlichen Sitzung der Versammlung der Berlin-Brandenburgischen Akademie der Wissenschaften am 12. Dezember 2003* (S. 65-67). Berlin: Oktoberdruck.
- Moosbrugger, H. & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion: Mit 66 Abbildung und 41 Tabellen* (2. Aufl.). Springer-Lehrbuch. Springer.

- Moosbrugger, H. & Schermelleh-Engel, K. (2012). Exploratorische (EFA) und Konfirmatorische Faktorenanalyse (CFA). In H. Moosbrugger & A. Kelava (Hrsg.), Springer-Lehrbuch. Test-theorie und Fragebogenkonstruktion (S. 325–343). Springer Berlin Heidelberg.
- Moreno, R. (2006). When worked examples don't work: Is cognitive load theory at an impasse? *Learning and Instruction*, 16, S. 170-181.
- Moreno, R. & Park, B. (2010). Cognitive load theory: historical development and relation to other theories. In J. Plass, R. Moreno, & R. Brünken, *Cognitive load theory* (S. 9-28). New York: Cambridge University Press.
- Morrison, J. B., Tversky, B. & Betrancourt, M. (2000). Animation: Does it facilitate learning. In *AAAI spring symposium on smart graphics* (Vol. 5359).
- Myers, D. G. (2014). *Psychologie* (3rd ed.). Springer Berlin Heidelberg; Imprint: Springer.
- National Research Council [NRC]. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, D.C.: The National Academies Press.
- Neff, S. (2023). Transfer digitaler Innovationen in die Schulpraxis: Eine explorative Untersuchung zur Förderung der Implementation, vol. 365. LOGOS Verlag BERLIN.
- Nerdel, C. (2017). *Grundlagen der Naturwissenschaftsdidaktik*. Berlin: Springer.
- Nettelbeck, T. (2011). Basic processes of intelligence. In R. J. Sternberg, & S. B. Kaufman, *The Cambridge handbook of intelligence* (S. 371–393). New York, NY: Cambridge University Press.
- Nicolaou, C. T. & Constantinou, C. P. (2014). *Educational Research Review*.
- Niegemann, H. M., Domagk, S., Hessel, S., Hein, A., Hupfer, M. & Zobel, A. (2008). *Kompodium multimediales Lernen*. Springer Science & Business Media.
- Nikmehr, N. & Doroodchi, M. (2008). New paradigm in evaluating usability of E-learning system. In *2008 International Conference on Innovations in Information Technology* (S. 347–351). IEEE.
- Nitz, S. & Fechner, S. (2018). Mentale Modelle. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 69–86). Springer Berlin Heidelberg.
- Nokelainen, P., Miettinen, M., Kurhila, J., Flore´en, P., & Tirri, H. (2005). A shared document-based annotation tool to support learner-centred collaborative learning. *British Journal of Educational Technology*, 36(5), 757-770.
- Norman, D. A. (2014). Some observations on mental models. In *Mental models* (S. 15–22). Psychology Press.
- Oh, P. S., & Oh, S. J. (2011). What Teachers of Science Need to Know about Models: An overview. *International Journal of Science Education*, 33(8), 1109-1130.
- Olejnik, S. & Algina, J. (2003). Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological methods*, 8(4), 434–447.
- Oliva, J. M., Del Mar Aragón, M. & Cuesta, J. (2015). The Competence of Modelling in Learning Chemical Change: A Study with Secondary School Students. *International Journal of Science and Mathematics Education*, 13(4), 751–791.
- Orsenne, J. (2016). Aktivierung von Schülervorstellungen zu Modellen durch praktische Tätigkeiten der Modellbildung.
- Osborne, R. J., & Cosgrove, M. M. (1983). Children's conceptions of the changes of state of water. *Journal of research in science teaching*, 20 (9), 825-838.

- Ostermann, A., Härtig, H., Kampschulte, R., Lindmeier, A., Ropohl, M., & Schwanewedel, J. (2019). Wie nutzen MINT-Lehrkräfte Medien? Erste Ergebnisse einer Befragung. In C. Maurer (Hg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe*. Regensburg: Gesellschaft für Didaktik der Chemie und Physik (GDChP).
- Özmen, H. (2011a). Turkish Primary students' conceptions about the particulate nature of matter. *International journal of Environmental & Science Education*, 6(1), 99-121.
- Özmen, H. (2011b). Effect of animation enhanced conceptual change texts on 6th grade students' understanding of the particulate nature of matter and transformation during phase changes. *Computers & Education* 57, 1114–1126.
- Paas, F. & van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load (Nr. 2). Elsevier. *Learning and Instruction*, 16.
- Paas, F. G., & Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review*, 6(4), 351-371.
- Paas, F. G., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educ. Psychol.* 38, 1–4. doi: 10.1207/S15326985EP3801_1.
- Paas, F. G., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional science* 32(1/2), 1-8.
- Paas, F. G., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1), 63-71.
- Paas, F. G., van Gog, T., & Sweller, J. (2010). Cognitive load theory: new conceptualizations, specifications, and integrated research. *Educ. Psychol. Rev.* 22, 115–121. doi: 10.1007/s10648-010-9133-8.
- Paas, F., Tuovinen, J. E., van Merriënboer, J. J. G. & Aubteen Darabi, A. (2005). A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. *Educational technology research and development*, 53, 25–34.
- Pagano, R. R. (2013). *Understanding statistics in the behavioral sciences* (10th ed.). Wadsworth Cengage Learning.
- Parchmann, I., & Schwarzer, S. (2016). Kann man Atome sehen? Naturwissenschaften im Unterricht - Chemie, 27 (153), 15-17.
- Passmore, C., Gouvea, J. S. & Giere, R. (2014). Models in science and in learning science: Focusing scientific practice on sense-making. *International handbook of research in history, philosophy and science teaching*, 1171–1202.
- Paulhus, D. L. & John, O. P. (1998). Egoistic and Moralistic Biases in Self-Perception: The Interplay of Self-Deceptive Styles With Basic Traits and Motives. *Journal of Personality*, 66(6), 1025–1060.
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley, *The role of constructs in psychological and educational measurement* (S. 49-69). Mahwah NY: Erlbaum.
- Pawek, C. (2009). Schülerlabore als interessefördernde außerschulische Lernumgebungen für Schülerinnen und Schüler aus der Mittel- und Oberstufe.
- Peeters, H., Habig, S. & Fechner, S. (2023). Does Augmented Reality Help to Understand Chemical Phenomena during Hands-On Experiments?—Implications for Cognitive Load and Learning. *Multi-modal Technologies and Interaction*, 7(2), 9.

- Pfundt, H. (1981). Das Atom-letztes Teilungsstück oder erster Aufbaustein. Zu den Vorstellungen, die sich Schüler vom Aufbau der Stoffe machen. *Chimica Didactica*, 7, 75-94.
- Pitschke. (2011). Gute Modelle – Wie die Qualität von Unternehmensmodellen definiert und gemessen werden kann. Version 2.0. BCS White Paper-Reihe „Modellieren als ingenieurtechnische Disziplin“, 1-11.
- Prediger, S., Link, M., Hinz, R., Hußmann, S., Thiele, J. & Ralle, B. (2012). Lehr-Lernprozesse initiieren und erforschen-fachdidaktische Entwicklungsforschung im Dortmunder Modell. *Der mathematische und naturwissenschaftliche Unterricht*, 65(8), 452–457.
- Prendergast, L. A. (2008). Trimming influential observations for improved single-index model estimated sufficient summary plots. *Computational Statistics & Data Analysis*, 52(12), 5319–5327.
- Prieto, T., Watson, J. R., & Dillon, J. S. (1992). Pupils' understanding of combustion. *Research in Science Education* (22), S. 331-340.
- Probst, C., Fetzer, D., Lukas, S. & Huwer, J. (2022). Effekte von Augmented Reality (AR) zur Visualisierung eines dynamischen Teilchenmodells-virtuelle Modelle zum Anfassen. *CHEMKON*, 4(29), 164–170.
- Przywarra, T. & Risch, B. (2021). Didaktik: Kugeln, Bilder oder Augmented Reality? *Nachrichten aus der Chemie*, 69(11), 12–15.
- Przywarra, T. & Risch, B. (2022). Intervention studies to compare different model types: challenges and possible solutions. *CHEMKON*, 29(S1), 250–254.
- Przywarra, T., Engl, A. & Risch, B. (2021). Entwicklung eines Messinstruments zur Qualitätsbestimmung von (digitalen) Anschauungsmodellen. *Empirische Pädagogik (EP)*, 35(1), 19–37.
- Radu, I. (2014). Augmented reality in education: a meta-review and cross-media analysis. *Personal and Ubiquitous Computing*, 18(6), 1533–1543.
- Ralle, B. & Di Fuccia, D.-S. (2014). Aktionsforschung als Teil fachdidaktischer Entwicklungsforschung. *Methoden in der naturwissenschaftsdidaktischen Forschung*, 43–55.
- Rasch, B., Frieze, M., Hofmann, W. & Naumann, E. (2014a). *Quantitative Methoden 2*. Springer Berlin Heidelberg.
- Rasch, B., Frieze, M., Hofmann, W. & Naumann, E. (2014b). *Quantitative Methoden. Band 2* (4. Auflage): SPSS-Ergänzung. Springer-Verlag.
- Rau, M. A. (2017). Conditions for the effectiveness of multiple visual representations in enhancing STEM learning. *Educational Psychology Review*, 29, 717–761.
- Rauh, O., & Stickel, E. (1997). *Konzeptuelle Datenmodellierung*. Stuttgart, Leipzig: B.G. Teubner Verlagsgesellschaft.
- Reeves, T. C. & Oh, E. G. (2017). The goals and methods of educational technology research over a quarter century (1989–2014). *Educational technology research and development*, 65(2), 325–339.
- Rehfus, W. (. (2003). *Handbuch Philosophie*. Göttingen: Vandenhoeck & Ruprecht.
- Reith, M., & Nehring, A. (2020). Scientific reasoning and views on the nature of scientific inquiry: testing a new framework to understand and model epistemic cognition in science. *International Journal of Science Education*, S. 1-26.
- Renninger, K. A. & Hidi, S. (2002). Student Interest and Achievement. *Development of Achievement Motivation*, 173–195.

- Renninger, K. A. & Hidi, S. (2011). Revisiting the Conceptualization, Measurement, and Generation of Interest. *Educational Psychologist*, 46(3), 168–184.
- Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. *Intrinsic and Extrinsic Motivation*, 373–404.
- Renninger, K. A., Ewen, L. & Lasher, A. K. (2002). Individual interest as context in expository text and mathematical word problems. *Learning and Instruction*, 12(4), 467–490.
- Renninger, K. A., Sinclair, N., Hand, V. M., Stohl, H., Alejandre, S. & Underwood, J. S. (2004). Students' interest for and work with applet-enhanced word problems. In C. A. Steinkuehler, Y. Kafai, W. A. Sandoval & N. Enyedy (Hrsg.), *Proceedings of the 6th international conference on Learning sciences*.
- Renström, L., Andersson, B., & Marton, F. (1990). Students' Conceptions of Matter. *Journal of Educational Psychology* 82(3), S. 555-569.
- Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen. *Diagnostica* 47(2), S. 57-66.
- Riaz, M. (2004). Helping children to understand particulate nature of matter. *Alberta Science Education Journal*, 36(2), S. 56-59.
- Riese, J. & Reinhold, P. (2014). Entwicklung eines Leistungstests für fachdidaktisches Wissen. *Methoden in der naturwissenschaftsdidaktischen Forschung*, 257–267.
- Ritchey, T. (2012). Outline for a morphology of modelling methods. *Acta Morphologica Generalis AMG Vol*, 1(1), 1012.
- Roese, N. J. & Vohs, K. D. (2012). Hindsight bias. *Perspectives on psychological science*, 7(5), 411–426.
- Roos, J. & Schöler, H. (2013). Veränderung des Fähigkeitsselbstkonzeptes durch den Übergang. *Transitionen in der Bildungsbiographie: Der Übergang vom Primar- zum Sekundarbereich*, 121–152.
- Ross, S. M. (1994). Delivery trucks or groceries? More food for thought on whether media (will, may, can't) influence learning. *Educational technology research and development*, 42(2), 5–6.
- Rost, J. (2000). Allgemeine Standards für die Evaluationsforschung. *Evaluation psychologischer Interventionsmaßnahmen: Standards und Kriterien*, 130–140.
- Rumelhart, D. E. (1981). "Schemata: the building blocks of cognition". In J. T. Guthrie, *Comprehension and Teaching: Research Reviews* (S. 3-26). Newark: International Reading Assn.
- Rupp, C. (2007). *Requirements-Engineering und -Management: Professionelle, iterative Anforderungsanalyse für die Praxis*. München: Carl Hanser Verlag.
- Ryan, R. M. & Deci, E. L. (Hg.). (2002). *Handbook of self-determination research*. University of Rochester Press.
- Saborowski, J. (2000). *Computervisualisierung und Modelldenken: konzeptionelle Grundlagen und fachdidaktische Konsequenzen für den Chemieunterricht*. Köln: Jörg Saborowski Verlag.
- Sadoski, M. (2001). Resolving the Effects of Concreteness on Interest, Comprehension, and Learning Important Ideas From Text. *Educational Psychology Review*, 13(3), 263–281.
- Sandmann, A. (2014). Lautes Denken - die Analyse von Denk-, Lern- und Problemlöseprozessen. In D. Krüger, I. Parchmann, & H. Schecker, *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 179-188). Berlin Heidelberg: Springer Spektrum.
- Sanger, M. (2000). Using Particulate Drawings to Determine and Improve Students' Conceptions of Pure Substances and Mixtures. *Journal of Chemical Education* 77, (12), 762-766.

- Sanger, M. (2009). Computer Animations of Chemical Processes at the Molecular Level. In N. J. Pienta, M. M. Cooper, & T. J. Greenbowe, *Chemists' guide to effective teaching*, Volume II (S. 198–211). New Jersey: Pearson Prentice Hall.
- Sasaki, T. (2003). Recipient orientation in verbal report protocols: Methodological issues in concurrent think-aloud. *Second Language Studies*, 22(1), 1–54.
- Sawilowsky, S. S. (2009). New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, 8(2), 597–599.
- Schermelleh-Engel, K. & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Springer-Lehrbuch. Testtheorie und Fragebogenkonstruktion* (S. 119–141). Springer Berlin Heidelberg.
- Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, 8(2), 23–74.
- Schiefele, U. (2008). Lernmotivation und Interesse. *Handbuch der pädagogischen Psychologie*, 10, 38–49.
- Schiefele, U. (2009). Situational and Individual Interest. In *Handbook of Motivation at School* (S. 211–236). Routledge.
- Schmider, E., Ziegler, M., Danay, E., Beyer, L. & Bühner, M. (2010). Is it really robust? *Methodology*.
- Schmiemann, P. & Lücken, M. (2014). Validität-Misst mein Test, was er soll? *Methoden in der naturwissenschaftsdidaktischen Forschung*, 107–118.
- Schreier, M. (2014). Varianten qualitativer Inhaltsanalyse: Ein Wegweiser im Dickicht der Begrifflichkeiten. Advance online publication (Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, Vol 15, No 1 (2014)).
- Schwarz, C. V. & Gwekwerere, Y. N. (2007). Using a guided inquiry and modeling instructional framework (EIMA) to support preservice K-8 science teaching. *Science Education*, 91(1), 158–186.
- Schwarz, C., Reiser, B., Davis, E., Kenyon, L., Acher, A., Fortus, D., & al., e. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, 46, S. 632–654.
- Seidel, T. (2003). Technischer Bericht zur Videostudie "Lehr-Lern-Prozesse im Physikunterricht"; BI-QUA. IPN.
- Sekretariat der Ständigen Konferenz der Kultusminister. (2020). Bildungsstandards im Fach Chemie für die Allgemeine Hochschulreife (Beschluss der Kultusministerkonferenz vom 18.06.2020).
- Shavelson, R. J., Hubner, J. J. & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, 46(3), 407–441.
- Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. *American Journal of Applied Mathematics and Statistics*, 9(1), 4–11.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.
- Sieve, B., Graulich, N., Caspari, I., & Bittorf, R. (2017). Chemische Vorgänge als Prozesse erfassen. *Naturwissenschaften im Unterricht - Chemie*, 28 (160), S. 2–7.
- Silva, M., Bermúdez, K. & Caro, K. (2023). Effect of an augmented reality app on academic achievement, motivation, and technology acceptance of university students of a chemistry course. *Computers & Education: X Reality*, 2, 100022.

- Smith, C., Maclin, D., Houghton, C., & Hennessey, M. (2000). Sixth grade students' epistemologies of science: the impact of school science experiences on epistemological development. *Cognition and Instruction*, 18, S. 349–422.
- Sodian, B., Thoermer, C., Kircher, E., Grygier, P., & Günther, J. (2002). Vermittlung von Wissenschaftsverständnis in der Grundschule. In M. [.] Prenzel, & J. [.] Doll, *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen* (S. 192–206). Weinheim: Beltz.
- Sofianidis, A. (2022). Why Do Students Prefer Augmented Reality: A Mixed-Method Study on Preschool Teacher Students' Perceptions on Self-Assessment AR Quizzes in Science Education. *Education Sciences*, 12(5), 329.
- Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Springer: Wien, New-York.
- Steinbuch, K. (1977). Denken in Modellen. In G. Trommer, *Denken in Modellen*. Braunschweig: Westermann Anlage.
- Steiner, G. (2007). Lernen und Wissenserwerb. In A. Krapp, & B. Weidenmann, *Pädagogische Psychologie* (S. 137–202). Weinheim: Beltz Psychologie Verlags Union.
- Stern, E. (2001). Intelligence, prior knowledge, and learning. *International encyclopedia of the social and behavioral sciences*, 11(6), 7670–7674.
- Stiensmeier-Pelster, J. & Schöne, C. (2008). Fähigkeitsselbstkonzept. *Handbuch der pädagogischen Psychologie*, 62–73.
- Stiller, K. (2003). Lernstrategien und Lernerfolg beim computerbasierten Wissenserwerb. *Psychologie in Erziehung und Unterricht* 50, S. 258–269.
- Stull, A. T. & Hegarty, M. (2016). Model manipulation and learning: Fostering representational competence with virtual and concrete models. *Journal of Educational Psychology*, 108(4), 509–527.
- Sumfleth, E. (1992). Schülervorstellungen im Chemieunterricht. *MNU*, 45(7), 410–414.
- Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O. & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, 30(3), 261–288.
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr.* 4, S. 295–312. doi: 10.1016/0959-4752(94)90003-5.
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educ. Psychol. Rev.* 22, S. 123–138. doi: 10.1007/s10648-010-9128-5.
- Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology: General*, 119, S. 176–192.
- Sweller, J., van Meriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, S. 251–296. doi: 10.1023/A:1022193728205.
- Tamayo Alzate, O. E. & Sanmartí Puig, N. (2007). High-school Students' Conceptual Evolution of the Respiration Concept from the Perspective of Giere's Cognitive Science Model. *International Journal of Science Education*, 29(2), 215–248.
- Tarng, W., Tseng, Y.-C. & Ou, K.-L. (2022). Application of Augmented Reality for Learning Material Structures and Chemical Equilibrium in High School Chemistry. *Systems*, 10(5), 141.
- Tekedere, H. & Göke, H. (2016). Examining the effectiveness of augmented reality applications in education: A meta-analysis. *International Journal of Environmental and Science Education*, 11(16), 9469–9481.

- Tepner, O. & Dollny, S. (2014). Entwicklung eines Testverfahrens zur Analyse fachdidaktischen Wissens. *Methoden in der naturwissenschaftsdidaktischen Forschung*, 311–323.
- Terzer, E. & Upmeier zu Belzen, A. (2007). Naturwissenschaftliche Erkenntnisgewinnung durch Modelle – Modellverständnis als Grundlage für Modellkompetenz. *Zeitschrift für Didaktik der Biologie (ZDB) - Biologie Lehren und Lernen*, 16, 33–56 (33-56 Seiten / Zeitschrift für Didaktik der Biologie (ZDB) - Biologie Lehren und Lernen, Bd. 16 (2007)).
- Terzer, E. (2013). Modellkompetenz im Kontext Biologieunterricht. <https://doi.org/10.18452/16651>
- Terzer, E., & Upmeier zu Belzen, A. (2008). Naturwissenschaftliche Erkenntnisgewinnung durch Modelle – Modellverständnis als Grundlage für Modellkompetenz. *Zeitschrift für Didaktik Der Biologie (ZDB) - Biologie Lehren Und Lernen* 16(1), S. 33-56.
- Theyßen, H. (2014). Methodik von Vergleichsstudien zur Wirkung von Unterrichtsmedien. *Methoden in der naturwissenschaftsdidaktischen Forschung*, 67–79.
- Thomas, J., Struckmeier, S., & Sieve, B. (2017). „Molekulares Sieben 2.0“ – vom Kontinuum zum Diskontinuum mit molekularer Küche. *Chemkon* 3/2017, S. 142–145.
- Thomsen, T., Lessing, N., Greve, W. & Dresbach, S. (2018). Selbstkonzept und Selbstwert. *Entwicklungspsychologie des Jugendalters*, 91–111.
- Thuseethan, S., Achchuthan, S. & Kuhanesan, S. (2014). Usability evaluation of learning management systems in Sri Lankan universities. *arXiv preprint arXiv:1412.0197*.
- Tracy, J. P. & Albers, M. J. (2006). Measuring cognitive load to test the usability of web sites. In *Annual Conference-society for technical communication*.
- Treagust, D., Chittleborough, G., & Mamiala, T. (2002). Student's understanding of the role of scientific models in learning science. *International Journal of Science Education*, 24, S. 357-368.
- Trizano-Hermosilla, I. & Alvarado, J. M. (2016). Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. *Frontiers in psychology*, 7, 769.
- Tschiersch, A., Krug, M., Huwer, J. & Banerji, A. (2021). Augmented Reality in chemistry education – an overview. *CHEMKON*, 28(6), 241–244.
- Tselios, N., Avouris, N. & Komis, V. (2008). Towards a unified usability evaluation approach for learning software environments: Issues and challenges. *Education and Information Technologies Journal*, 13(1), 55–76.
- Tullis, T. S. & Stetson, J. N. (2004). A comparison of questionnaires for assessing website usability. *usability professionals association (UPA) 2004 conference*. Minneapolis, Minnesota.
- Tversky, B., Morrison, J. B. & Bétrancourt, M. (2002). Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4), 247–262.
- Upmeier zu Belzen, A. (2013). Unterrichten mit Modellen. In H. Gropengießer, U. Harms, & K. U. Fachdidaktik Biologie (S. 325-334). Aulis-Verlag: Halbergmoos.
- Upmeier zu Belzen, A., & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften* (16), S. 41-57.
- Urhahne, D., Nick, S. & Schanze, S. (2009). The effect of three-dimensional simulations on the understanding of chemical structures and their properties. *Research in science education*, 39, 495–513.

- Urhahne, D., Prenzel, M., Davier, M. von, Senkbeil, M. & Bleschke, M. (2000). Computereinsatz im naturwissenschaftlichen Unterricht-Ein Überblick über die pädagogisch-psychologischen Grundlagen und ihre Anwendung. *Zeitschrift für Didaktik der Naturwissenschaften*, 6, 157–186.
- van Driel, J. H. & Verloop, N. (2002). Experienced teachers' knowledge of teaching and learning of models and modelling in science education. *International Journal of Science Education*, 24(12), 1255–1272.
- van Driel, J. H., & Verloop, N. (1999). Teachers' knowledge of models and modelling in science. *International Journal of Science Education*, 21(11), 1141-1153.
- van Gog, T. (2014). The signaling (or cueing) principle in multimedia learning. *The Cambridge handbook of multimedia learning*, 263–278.
- van Merriënboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17, 147-177.
- van Merriënboer, J. J. G., Schuurman, J. G., Croock, M. B. M. de & Paas, F. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction*, 12(1), 11–37.
- Villacorta, P. J. (2017). The welchADF Package for Robust Hypothesis Testing in Unbalanced Multivariate Mixed Models with Heteroscedastic and Non-normal Data. *R J*, 9(2), 309.
- Vlachogianni, P. & Tselios, N. (2022). Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review. *Journal of Research on Technology in Education*, 54(3), 392–409.
- von Aufschnaiter, C., & von Aufschnaiter, S. (2005). Über den Zusammenhang von Handeln, Wahrnehmen und Denken. In R. Voß, & (Hrsg.), *Unterricht aus konstruktivistischer Sicht* (2. Auflage) (S. S. 234-248). Weinheim: Beltz.
- Weinert, F. E. & Helmke, A. (1996). *Der gute Lehrer: Person, Funktion oder Fiktion?* Beltz Verlag.
- Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert, *Leistungsmessungen in Schulen* (S. 17-31). Weinheim: Beltz.
- Welch, B. L. (1951). On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika*, 38(3/4), 330.
- Wellnitz, N., & Mayer, J. (2012). Beobachten, Vergleichen und Experimentieren: Wege der Erkenntnisgewinnung [Observing, Comparing, and Experimentation: Ways of Inquiry]. In U. Harms, & F. X. Bogner, *Lehr- und Lernforschung in der Biologiedidaktik. Band 5. Didaktik der Biologie – Standortbestimmung und Perspektiven. Internationale Tagung der Fachsektion Didaktik der Biologie im VBIO, Bayreuth 2010*. Innsbruck: Studien Verlag.
- Weßnigk, S. (2012). *Kooperatives Arbeiten an industrienahen außerschulischen Lernorten*. Doctoral dissertation. Kiel.
- Wigfield, A. & Eccles, J. S. (2000). Expectancy-Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, 25(1), 68–81.
- Wild, J. & Pissarek, M. (2016). *RATTE Regensburger Analysetool für Texte: Dokumentation*.
- Willems, A. S. (2018). *Lernmotivation und Interesse*. Aulis; Seelze.
- Williamson, V. (2011). Teaching chemistry with visualizations: What's the research evidence? In D. Bunce, *Investigating classroom myths through research on teaching and learning* (S. 65-81). Washington, DC: American Chemical Society.

- Williamson, V. (2015). What is the research evidence for using visualization techniques in the chemistry classroom? How should these techniques be implemented? *LUMAT: International Journal on Math, Science and Technology Education*, 3(4), 545-555.
- Williamson, V. M. & Abraham, M. R. (1995). The effects of computer animation on the particulate mental models of college chemistry students. *Journal of Research in Science Teaching*, 32(5), 521–534.
- Williamson, V. M., & José, T. J. (2009). Using visualization techniques in chemistry teaching. In N. J. Pienta, M. M. Cooper, & T. J. Greenbowe, *Chemists' guide to effective teaching*, Volume II (S. 71-88). New Jersey: Pearson Prentice Hall.
- Winnenburg, W. (1993). Tafelbilder-Lernoptimierung durch wahrnehmungspsychologische adäquate Gestaltung. In G. Otto, *Unterrichtsmedien Jahresheft XI* (S. 6-9). Seelze: Friedrich Verlag.
- Wirtz, M. (2004). Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und reduziert werden kann [On the problem of missing data: How to identify and reduce the impact of missing data on findings of data analysis]. *Die Rehabilitation*, 43(2), 109–115.
- Wirtz, M. A. & Caspar, F. (2002). Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. Hogrefe Verl. für Psychologie.
- Wu, H.-K. & Puntambekar, S. (2012). Pedagogical affordances of multiple external representations in scientific processes. *Journal of Science Education and Technology*, 21, 754–767.
- Wu, J.-H. & Yuan, Y. (2003). Improving searching and reading performance: the effect of highlighting and text color coding. *Information & Management*, 40(7), 617–637.
- Xu, W.-W., Su, C.-Y., Hu, Y. & Chen, C.-H. (2022). Exploring the Effectiveness and Moderators of Augmented Reality on Science Learning: A Meta-analysis. *Journal of Science Education and Technology*, 31(5), 621–637.
- Yan, F., & Talanquer, V. (2015). Students' Ideas about How and Why Chemical Reactions Happen: Mapping the conceptual landscape. *International Journal of Science Education* 37 (18), S. 3066–3092.
- Yu, Z. (2023). Meta-analyses of effects of augmented reality on educational outcomes over a decade. *Interactive Learning Environments*, 1–15.
- Yürük, N. (2007). The effect of supplementing instruction with conceptual change texts on students' conceptions of electrochemical cells. *Journal of Science Education and Technology*, 16, 515–523.
- Ziemek, H. P., Keiner, K. H. & Mayer, J. (2005). Problemlöseprozesse von Schülern der Biologie im naturwissenschaftlichen Unterricht-Ergebnisse qualitativer Studien. *Lehr- und Lernforschung in der Biologiedidaktik*, 2, 29–40.

11. Anhang

Der Anhang dieser Arbeit wurde digital auf der Plattform [zenodo.org](https://doi.org/10.5281/zenodo.14186392) unter <https://doi.org/10.5281/zenodo.14186392> veröffentlicht. Die Dateien stehen durch Klicken auf die entsprechende Zeile zum Direktdownload bereit.

11.1 Lernmaterialien

[Anhang 1 Lernmaterialien zu Tag 1 Augmented Reality, pilotiert](#)

[Anhang 2 Lernmaterialien zu Tag 2 Augmented Reality, pilotiert](#)

[Anhang 3 Lernmaterialien zu Tag 3 Augmented Reality, pilotiert](#)

[Anhang 4 Lernmaterialien zu Tag 1 haptisch-interaktiv, pilotiert](#)

[Anhang 5 Lernmaterialien zu Tag 2 haptisch-interaktiv, pilotiert](#)

[Anhang 6 Lernmaterialien zu Tag 3 haptisch-interaktiv, pilotiert](#)

[Anhang 7 Lernmaterialien zu Tag 1 illustrativ, pilotiert](#)

[Anhang 8 Lernmaterialien zu Tag 2 illustrativ, pilotiert](#)

[Anhang 9 Lernmaterialien zu Tag 3 illustrativ, pilotiert](#)

11.2 Fragebögen

[Anhang 10 Fragebogen Pre- und Posttest, pilotiert](#)

[Anhang 11 Zwischenfragebogen, pilotiert](#)

11.3 Ergänzende methodische Inhalte

[Anhang 12 Deskriptive Statistiken der Kovariablen in der Studie III](#)

[Anhang 13 Ergebnisse zu den Unterschiedsprüfungen der Kovariablen in der Studie III](#)

[Anhang 14 Link zum Video zur Handhabung der Modellbewertung](#)

[Anhang 15 Transkriptionsregeln zu Studie I](#)

[Anhang 16 Erstes Kategoriensystem zu Studie I](#)

[Anhang 17 Kodierleitfaden zu Studie I](#)

[Anhang 18 Ausdifferenziertes Kategoriensystem zu Studie I](#)

[Anhang 19 Transkripte zu Studie I](#)

[Anhang 20 APK-Dateien zur Installation der AR-Apps](#)

11.4 Darstellung des Eigenanteils und Eidesstattliche Erklärung

„Die zu dieser Dissertation zugehörigen Materialien der Lerneinheiten habe ich selbst auf Basis chemiedidaktischer Grundlagen entworfen und evaluiert. Alle empirischen Daten wurden eigenverantwortlich erhoben und ausgewertet. Alle Publikationen, die im Rahmen dieses Forschungsprojekts entstanden sind, wurden von mir als Erstautor konzipiert, verfasst und unter Zusammenarbeit mit meinen Mitautoren kritisch überarbeitet.“

Hiermit erkläre ich, dass ich die vorgelegte Dissertation selbstständig verfasst habe. Alle wörtlich oder inhaltlich übernommenen Stellen habe ich als solche gekennzeichnet. Ich versichere außerdem, dass ich die beigefügte Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und, dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.“

Landau den 26.11.2024

Tobias Bier

11.5 Danksagung

An dieser Stelle möchte ich all jenen danken, die zum Gelingen dieser Arbeit beigetragen haben.

Mein besonderer Dank gilt zunächst Björn Risch, Sandra Nitz und Markus Emden für ihre Tätigkeit als Gutachter*innen sowie Dirk Felzmann für den Vorsitz der Prüfungskommission.

Ein ganz besonderer und aufrichtiger Dank gilt Herrn Prof. Dr. Björn Risch. Ohne seine Unterstützung, seine Offenheit für neue Ideen und nicht zuletzt die Möglichkeit, unter seiner Betreuung promovieren zu dürfen, wäre diese Arbeit nicht entstanden. Sein Vertrauen in meine wissenschaftliche Arbeit, seine stetige Ermutigung und seine konstruktive Begleitung durch alle Phasen der Dissertation haben maßgeblich zu deren Erfolg beigetragen.

Ein herzlicher Dank geht an Helma Ederer, Susanne Brückmann, Doris Burgard und Anja Hanß, die mit großem Engagement die Vorbereitung der Chemikalien und Materialien übernommen haben.

Ebenso danke ich den Hilfswissenschaftler*innen Viktorija Ovcharova, Isabel Nicole Zachert, Simeon Jasper Wallrath, Johanna Listi, Julia Bier und Felix Burkhardt für ihre tatkräftige Unterstützung bei der Durchführung der Experimente.

Dem Institut für naturwissenschaftliche Bildung (InB), der Graduiertenakademie Bildung-Mensch-Umwelt, dem Projekt DiAmant sowie allen Mitgliedern der Arbeitsgruppe Chemiedidaktik danke ich für den bereichernden wissenschaftlichen Austausch.

Mein aufrichtiger Dank gilt außerdem den Beraterinnen für naturwissenschaftlichen Unterricht, die mit ihrer Expertise maßgeblich zum Gelingen beigetragen haben, sowie den zahlreichen Schüler*innen, die als Proband*innen an der Studie teilgenommen haben.

Ein besonderer Dank gebührt Alexander Engl für seinen unermüdlichen Einsatz während des gesamten Projekts.

Abschließend danke ich meiner Frau Julia von Herzen – für ihren Rückhalt, ihre Geduld und ihre kreative Bastelhilfe. Ohne sie wäre vieles nicht möglich gewesen.