Der Einfluss von Datenkompetenz auf das Argumentieren beim Experimentieren

Entwicklung und Evaluation von Lernapps zur Förderung von Datenkompetenz

Dissertation

zur Erlangung des akademischen Grades

doctor philosophiae
(Dr. phil.)

eingereicht an der
Fakultät für Natur- und Sozialwissenschaften
an der Pädagogischen Hochschule Karlsruhe

von

Engin Kardaş

Gutachter:
1. Jun.-Prof. Dr. Tobias Ludwig
2. Prof. Dr. Burkhard Priemer

Zusammenfassung

Das Auswerten von Daten und die Begründung von wissenschaftlichen Hypothesen kann als zentraler Bestandteil von „scientific practice“ angesehen werden. Durch das naturwissenschaftliche Experiment können beide Aspekte adressiert werden: Lernende müssen Beobachtungen und Daten mit ihren Hypothesen verbinden. Dieser Prozess kann als Argumentieren aufgefasst werden und bedingt Datenkompetenz bei Lernenden.

Vor diesem Hintergrund hat die vorliegende Arbeit die Förderung von Datenkompetenz zum Ziel. Beim Experimentieren stellt der adäquate Umgang mit Daten und Messunsicherheiten einen wesentlichen Teil der Datenkompetenz dar. Zunächst wurde daher eine didaktische Strukturierung entwickelt, die die Datenkompetenz beim Experimentieren a) auf die für die Sekundarstufe I relevanten Aspekte reduziert und b) in die Kompetenzbereiche „Direktes Messen“, „Indirektes Messen“, „Grafische Auswertung“ und „Signifikanz“ modularisiert. Auf dieser Basis wurden zehn digitale Lernapps entwickelt, die durch eine Expertenstudie hinsichtlich der Passung zur Strukturierung und der fachlichen Richtigkeit evaluiert wurden. Die Wirksamkeit der Apps konnte in mehreren Interventionsstudien mit insgesamt $n_1 = 132$ Schüler:innen mit mittleren bis großen Effekten bezüglich des Kompetenzerwerbs empirisch nachgewiesen werden.

Zudem wurde untersucht, welchen Einfluss die Förderung der Datenkompetenz auf das Argumentieren beim Experimentieren besitzt. In einer längsschnittlichen Studie wurden den Schüler:innen vor und nach der Intervention mit Lernapps Messdaten zum Zusammenhang zwischen der Pendelmasse m und der Periodendauer T eines Fadenpendels gezeigt und ihre Hypothesen sowie die Zustimmung zu vier Argumentkategorien erfasst.

In der Studie mit $n_2 = 325$ Schüler:innen konnte gezeigt werden, dass der Erwerb von Datenkompetenz a) die Zustimmung zu daten- und evidenzbasierten Argumentkategorien erhöht, b) die Zustimmung zu intuitiven und auf Heuristiken basierenden Argumentkategorien verringert und c) positiv auf die Richtigkeit einer physikalischen Hypothese wirkt.

Die vorliegende Arbeit trägt dazu bei, die Wirkung von Datenkompetenz auf das Argumentieren beim Experimentieren zu verstehen: Durch kurze Interventionen zum Umgang mit Daten und Messunsicherheiten können die Argumentationsfähigkeit der Schüler:innen beim Experimentieren sowie der selbstständige Lernerfolg durch Experimentieren gefördert werden.

Abstract

Evaluation of Data and justifying hypotheses can be seen as central aspects of scientific practices. Using scientific experiments both aspects can be adressed: Students have to link data and observations with their hypotheses. This process can be seen as argumentation.

Against this background this work aims to foster data competencies. While experimenting an adequate handling of data and measurement uncertainties are substantial parts of data competencies. Hence, first a didactic structure was created that a) focusses on relevant aspects of those competencies for grade seven to ten and b) modularizes the data competencies into the four dimensions „direct measurement“, „indirect measurement“, „graphical analysis“ and „significance“. Based on this ten digital learning apps were created. The apps were evaluated by an expert study regarding their fitting to the structure and their technical correctness. The apps' efficacy could be shown empirically with medium sized and large effects in a longitudinal study with $n_1 = 132$ students.

Finally the effect of fostering data competencies on argumentation while experimenting was examined. In a longitudinal study students were shown measurement data regarding the relation between pendulum mass m and periodic length T and their hypotheses and approval to four argument categories were measured before and after the intervention with learning apps.

The study with $n_2 = 325$ students could show that improving data competencies leads to a a) higher approval of data- and evidence-based argument categories, b) lower approval of intuition- or heuristics-based argument categories and c) positive effect of the hypothesis' correctness.

The presented work contributes to understanding the effect of data competencies on argumentation while experimenting: Short and modular intervention on handling data and measurement uncertainties can foster argumentation skills and self-dependent learning success of students.

Inhaltsverzeichnis

1 Einleitung 19

2 Theoretischer Hintergrund und Stand der Forschung 23
2.1 Das naturwissenschaftliche Argumentieren 23
2.1.1 Definition zentraler Begriffe 24
2.1.2 Argumentieren im naturwissenschaftlichen Unterricht 25
2.1.3 Überblick über den Stand der Forschung zur Förderung des Argumentierens im naturwissenschaftlichen Unterricht . . . 28
2.1.4 Analyse von Argumenten 35
2.1.5 Intrapersonelle Ebene des Argumentierens 38
2.2 Argumentieren beim Experimentieren 39
2.2.1 Warum ist das Experiment als Argumentationsanlass geeignet? 39
2.2.2 Nicht-hypothesenkonforme Daten als Ausgangspunkt 42
2.2.3 Modellierung von Argumentationsprozessen beim Experimentieren durch das Elaboration-Likelihood Model of Persuasion 44
2.2.4 Empirische Befundlage zum Argumentieren beim Experimentieren 45
2.3 Umgang mit Daten und Unsicherheiten im Physikunterricht 47
2.3.1 Grundlagen der Metrologie 47
2.3.2 Daten und Messunsicherheiten in bildungspolitischen Vorgaben 49
2.3.3 Was bedeutet adäquater Umgang mit experimentellen Daten und Messunsicherheiten? 50
2.3.4 Daten und Unsicherheiten beim Experimentieren 53
2.3.5 Fachdidaktische Arbeiten zu Messdaten im Physikunterricht 56
2.4 Digitale Medien im Physikunterricht 61
2.5 Zusammenfassung und Ableitung des Forschungsvorhabens 62

3 Entwicklungsziele und Forschungsfragen 65

4 Entwicklung der Lernapps zum Umgang mit Daten und Messunsicherheiten 69
4.1 Entwicklung 69
4.1.1 Didaktische Strukturierung 69
4.1.2 Validität der Passung zwischen didaktischer Struktur und Sachstrukturmodell 78
4.1.3 Methodische und didaktische Charakteristika der Lernapps . 79

4.2 Prototypische Darstellung der Apps 82
4.3 Zwischenfazit zur Entwicklung der digitalen Lernapps 88

5 Evaluation der Apps hinsichtlich der Wirksamkeit 89
5.1 Methode 89
5.1.1 Beschreibung der Studie 89
5.1.2 Operationalisierung der Kompetenz im Umgang mit Daten und Messunsicherheiten 90
5.1.3 Studiendesign 93
5.1.4 Ablauf der Studie 93
5.1.5 Pilotierung 94
5.1.6 Kurzbeschreibung der verwendeten statistischen Verfahren . 96
5.1.7 Stichprobe 99
5.1.8 Evaluation der Testinstrumente 101
5.2 Ergebnisse 102
5.2.1 Modellevaluation 104
5.2.2 Parameterschätzung 104
5.3 Diskussion 109

6 Wirkung der Datenkompetenz auf das Argumentieren und den Lernerfolg 113
6.1 Methode 113
6.1.1 Beschreibung der Experimentiersituation 113
6.1.2 Beschreibung der Intervention 115
6.1.3 Operationalisierung der Konstrukte 116
6.1.4 Studiendesign 119
6.1.5 Ablauf der Studie 119
6.1.6 Pilotierung 120
6.1.7 Beschreibung der verwendeten statistischen Verfahren 121
6.1.8 Stichprobe 124
6.1.9 Analyse der Messmodelle 127
6.2 Ergebnisse 132
6.2.1 Mittelwertstruktur 132
6.2.2 Beschreibung der latenten Faktoren 134
6.2.3 Latente Analysen 135
6.2.4 Wirkung der Datenkompetenz auf den Lernerfolg 144
6.2.5 Post-hoc-Analyse 145
6.2.6 Diskussion der Post-hoc-Analyse 156
6.3 Diskussion 157
6.3.1 Einfluss der Datenkompetenz auf die Zustimmung zu den Argumentkategorien 158
6.3.2 Einfluss der Datenkompetenz und der Art der Begründung auf den Lernerfolg 159
6.3.3 Methodische Limitationen 162

7 Bewertung des Forschungsvorhabens in der Gesamtschau 165
7.1 Übergreifende Diskussion aller drei Studien 165
7.2 Limitationen 167
7.3 Häufig geäußerte Kritik 169
7.3.1 Programmierte Instruktion 169
7.3.2 konkurrierende Modelle in der Auswertung 171

8 Fazit und Ausblick 173

Anhang 179

A Didaktische Ausarbeitungen zu den Inhalten und Konzepten der Lernapps 179
A.1 A Direktes Messen 179
A.1.1 A2 Bestimmung von Unsicherheiten 179
Konzept 179
Struktur 179
Didaktische Reduktion 181
Konzeptbildung nach Oser 181
Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene 183
A.1.2 A3 Bestimmung von Unsicherheiten 185
Konzept 185
Struktur 185
Didaktische Reduktion 186
Konzeptbildung nach Oser 187
Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene 189
A.2 B Indirektes Messen 191
A.2.1 B1 Addition von absoluten Unsicherheiten 191
Konzept 191
Struktur 191
Didaktische Reduktion 192
Konzeptaufbau nach Oser 193
Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene 195
A.2.2 B2 Addition von relativen Unsicherheiten 198
Konzept 198
Struktur 198
Didaktische Reduktion 199
Konzeptaufbau nach Oser 200
Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene 201

A.3 C Grafische Auswertung 203
A.3.1 C1 Unsicherheitsflächen 203
Konzept 203
Struktur 203
Didaktische Reduktion 205
Konzeptaufbau nach Oser 206
Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene 207
A.3.2 C2 Ausgleichsgerade 209
Konzept 209
Struktur 209
Didaktische Reduktion 210
Konzeptaufbau nach Oser 211
Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene 212
A.3.3 C3 Drei-Geraden-Methode 216
Konzept 216
Inhaltlicher Schwerpunkt 216
Didaktische Reduktion 217
Konzeptbildung nach Oser 218
Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene 219
A.4 D Signifikanz 222
A.4.1 D2 Boxplots 222
Konzept 222
Inhaltlicher Schwerpunkt 222
Didaktische Reduktion 223
Konzeptaufbau nach Oser 225
Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene 226
B.1 Itemselektion 229
B.2 Reliabilität der Testinstrumente 231

C Fragebogen zur Evaluation der Wirksamkeit der Lernumgebung 233
C.1 Direktes Messen 233
C.2 Indirektes Messen 235
C.3 Grafische Auswertung 238
C.4 Signifikanz 242

D Fragebogen zu Studie 3 245

E Ausführungen zu den statistischen Methoden 249
E.1 Konvergenz der MCMC-Chains 249

E.2 Item Response Theorie 250
E.2.1 Analyse des Rasch-Modells für die Prä- und Posterhebung . . 250
Passung des Modells 252
Item Characteristic Curves (ICCs) 252
Wright-Maps 258
E.2.2 Messinvarianz 258
E.2.3 Analyse des Rasch-Modells für beide Erhebungszeitpunkte . 262
E.3 Latente Modellierung 265
E.3.1 Evaluation der Passung zwischen Daten und Modellen 266
hierarchische Modelle 268
E.3.2 Schätzmethoden 269
E.3.3 Messmodelle 270
Single-Item Indikator 272
E.3.4 Messinvarianz 272
E.4 Strukturgleichungsmodelle 280
E.4.1 Globales Messmodell 280
E.4.2 Strukturgleichungsmodell zur Untersuchung von FF2 281
E.4.3 Strukturgleichungsmodell zur Untersuchung von FF3 283
E.4.4 Strukturgleichungsmodell zur Untersuchung von FF4 284

Literatur 285

Abbildungsverzeichnis

4.1 Startseite der Lernapps 83
4.2 Einleitung der Apps 84
4.3 Darstellung einer Simulation von Messwerten 85
4.4 Visualisierung eines vollständiges Messergebnisses 86
4.5 Darstellung der Hilfe und der Lösungsanforderung 87

5.1 Darstellung der überarbeiteten Items zur Evaluation der Wirksamkeit 91
5.2 Darstellung der Mittelwertdifferenz und Effektstärke für die Apps A2 und B1 107
5.3 Darstellung Mittelwertdifferenz und Effektstärke für die Apps C1 und D1 108

6.1 Beispielitem zur Erfassung der Datenkompetenz 117
6.2 Übersicht zum Ablauf der Studie. 120
6.3 Darstellung der Zustimmung zu den Argumentkategorien 125

E.1 Traceplots der für Parameterschätzungen vewendeten MCMC-Chains 251
E.2 Darstellung der ICC für den Pretest 254
E.3 Darstellung der ICC für den Posttest 255
E.4 Wrightmap der Raschanalyse des *stacked* Datensatzes. 257
E.5 Darstellung aller ICCs Kurven für den Prä- (links) und Posterhebungszeitraum (rechts). 259
E.6 Vergleich der Itemschwierigkeiten zwischen den Testzeitpunkten . 259
E.7 Gegenüberstellung der geschätzten Schüler:innenfähigkeiten. . . . 260
E.8 Darstellung des ursprünglich ermittelten und verankerten Schwierigkeitsgrad für den Posttest. 261
E.9 Darstellung der ICCs für die Items S1 bis S9 des *stacked*-Modells 263
E.10 Ergebnisse der Rasch-Analyse des *Stacked*-Datensatzes. 264
E.11 Darstellung der verwendeten Messmodelle 280
E.12 Darstellung der Strukturgleichungsmodelle zur Untersuchung von Forschungsfrage 2 281
E.13 Latent-change-score-Modell zur Untersuchung von Forschungsfrage 2 282

E.14 Darstellung der Strukturgleichungsmodelle zur Untersuchung von Forschungsfrage 3 283
E.15 Darstellung der Strukturgleichungsmodelle zur Untersuchung von Forschungsfrage 4 284

Tabellenverzeichnis

2.1 Datenanalysekompetenz aus internationalen Bildungsplänen . . . 53

4.1 Übersicht der entwickelten didaktischen Struktur 70
4.2 Darstellung des Kompetenzbereichs Direktes Messen, der dazugehörigen Kompetenz und der Aufteilung in Konzepte. 72
4.3 Darstellung des Kompetenzbereichs Indirektes Messen, der dazugehörigen Kompetenz und der Aufteilung in Konzepte. 74
4.4 Darstellung des Kompetenzbereichs Grafische Auswertung, der dazugehörigen Kompetenz und der Aufteilung in Konzepte. . . . 76
4.5 Darstellung des Kompetenzbereichs Signifikanz, der dazugehörigen Kompetenz und der Aufteilung in Konzepte. 78
4.6 Ergebnisse des Expertenratings 79

5.1 Übersicht der adressierten Kompetenzen und verwendeten Items (Schulz, 2022) für die Evaluation der Wirksamkeit der jeweiligen App. 92
5.2 Übersicht des zeitlichen Aufwands in der Pilotierung 95
5.3 Cut-off-Werte zur Interpretation von BF 98
5.4 Übersicht über die Stichproben der Studien A-D. 100
5.5 Darstellung der Itemkennwerte für die verwendeten Testinstrumente in Studie 1. 103
5.6 Ergebnisse der Studien zur Evaluation der Wirksamkeit der Lernapps 105

6.1 Messergebnisse des verwendeten Datensatzes 115
6.2 Übersicht über die Fit-Indizes zur Evaluation der Modellgüte im Rahmen der latenten Modellierung 124
6.3 Überprüfung der multivariaten Normalverteilung der Indikatoren 124
6.4 Darstellung der Fit-Indizes für das globale Messmodell 130
6.5 Übersicht zur Reliabilität der Skalen 132
6.6 Vergleich der Ausprägungen der latenten Faktoren zum Prä- und Postzeitpunkt 133
6.7 Korrelationen der Zustimmung zu den Argumentkategorien . . . 135

6.8 Darstellung der standardisierten Regressionskoeffizienten für die Datenkompetenz als Prädiktor für die Zustimmung zu den Argumentkategorien. 136
6.9 Übersicht über die Veränderung der latenten Faktoren 137
6.10 Übersicht über die verwendeten Modelle 138
6.11 Einfluss der Zustimmung zu den Argumentkategorien auf die Richtigkeit der Hypothese 140
6.12 Einfluss der Zustimmung zu den Argumentkategorien auf die Richtigkeit der Hypothese unter Berücksichtigung der Datenkompetenz 141
6.13 Einfluss der Datenkompetenz auf die Richtigkeit der gewählten fachlichen Hypothese 142
6.15 Darstellung der Zustimmung zur Argumentkategorie Messunsicherheiten für verschiedene Gruppierungen und Erhebungszeiträume 149
6.16 Einfluss der Zustimmung zur Argumentkategorie Messunsicherheiten auf den Lernerfolg 150
6.17 Gruppierung nach Eingangshypothese ($n = 237$ mit falscher Hypothese). 152
6.18 Gruppierung nach Prähypothese ($n = 168$ mit falscher Hypothese). 153

B.1 Übersicht über die ermittelten Reliabilitäten im Rahmen der Pilotierung. 232

E.1 Konvergenzparameter der verwendeten MCMC-Chains 250
E.2 Übersicht der Itemschwierigkeiten und Itemparameter für den Prä- (S11-S19) und Postzeitpunkt (S21-S29). 253
E.3 Übersicht der Itemschwierigkeiten und Itemparameter für die Items S1 bis S9 unter Verwendung des Stacked-Modells. 262
E.4 Fit-Indizes für die entsprechenden Messmodelle für die Zustimmung der Argumentkategorien 271
E.5 Übersicht der Modellvergleiche zwischen den hierarchischen Modellen im Kontext der Messinvarianz. 276
E.5 Übersicht der Modellvergleiche zwischen den hierarchischen Modellen im Kontext der Messinvarianz. 277
E.5 Übersicht der Modellvergleiche zwischen den hierarchischen Modellen im Kontext der Messinvarianz. 278
E.5 Übersicht der Modellvergleiche zwischen den hierarchischen Modellen im Kontext der Messinvarianz. 279

Abkürzungsverzeichnis

BF	Bayes-Faktor
CFI	Comparative Fit Index
DIF	Differential-item-function
ELM	Elaboration-Likelihood Model of Persuasion
Evi	Zustimmung zur Argumentkategorie Daten als Evidenz
Exp	Zustimmung zur Argumentkategorie Expertenwissen
GUM	Guide to the expression of uncertainty in measurement
HDI	Highest-densitiy-Intervalle
ICC	Item Characteristic Curve
Int	Zustimmung zur Argumentkategorie Intuition
LCSM	Latent-change-score-model
ML	Maximum-Likelihood
MLR	robustes ML-Schätzverfahren
MU	Zustimmung zur Argumentkategorie Messunsicherheiten
RMSEA	Root Mean Square Error of Approximation
ROPE	Region-of-practical-equivalence-Intervalle
SRMR	Standardized Root Mean Square
VIM	International vocabulary of metrology

Kapitel 1

Einleitung

The proton is the only stable baryon, and all of life depends on it. Therefore, it is important to understand the proton. Its radius should be a simple quantity to determine and understand, but that is not the case.

Pohl et al., 2013, S. 200

Das vorangehende Zitat adressiert einen Konflikt, der in der Teilchenphysik als Protonen-Radius-Puzzle bekannt ist. Das Proton ist ein elementarer Bestandteil unserer Welt und ein Baustein unserer Materie. Ein Verständnis dieses Teilchens und die Bestimmung seiner Eigenschaften sind daher von besonderer Relevanz für die Wissenschaft und das Verständnis unserer Welt. Trotz größter Bemühungen scheint die Größe eines Protons bis zum heutigen Zeitpunkt noch nicht vollständig verstanden zu sein. Denn Forschungsarbeiten liefern in diesem Zusammenhang unterschiedliche Werte für den Radius eines Protons. Die Unterschiede in den Messergebnissen sind aus der Theorie heraus nicht zu erklären. Die Messungen sind jedoch nicht „fehlerhaft“. Die Forscher:innen haben im Rahmen der experimentellen Bestimmung den Aufbau, die verwendeten Geräte sowie die Methode berücksichtigt, um eine Messunsicherheit für ihre Ergebnisse zu bestimmen. Die Messunsicherheit ist ein quantitativer Parameter, welcher unter anderem das Vertrauen in die eigene Messung sowie das Messergebnis widerspiegelt und ein plausibles Intervall für die Messungen angibt. Unter Berücksichtigung der Messunsicherheiten ist eine zufällige Abweichung der ermittelten unterschiedlichen Radien extrem unwahrscheinlich. Forscher:innen befinden sich daher nach wie vor im Austausch darüber, wie die Abweichung der Messwerte erklärt werden könnte. Der Konflikt ist bis zum heutigen Tag nicht eindeutig gelöst (Februar 2023).

Während das Protonen-Radius-Puzzle primär eine hohe Relevanz für die Teilchenphysik aufweist, gibt es andere wissenschaftliche Arbeiten, deren Ergebnisse eine unmittelbare Bedeutung für gesellschaftliche Herausforderungen besitzen. Zum Beispiel entwickelte die Forschergruppe um Lyu et al. (2020) komplexe Modelle auf Grundlage von Daten und Simulationen für die Berechnung

des durch die Klimakrise zu erwartenden Anstiegs des Meeresspiegels. Es konnte gezeigt werden, dass das komplexe Modell hochsensibel ist und ein adäquater Umgang mit Messunsicherheiten notwendig ist, damit die ermittelten Effekte nicht unterschätzt werden. Eine Unterschätzung des Meeresspiegelanstiegs könnte fatale Konsequenzen für die Gesellschaft haben. Die Erkenntnisse der Arbeit tragen somit dazu bei, Lösungen für die menschengemachte Klimakrise zu entwickeln, und tangieren somit viel eher die Lebensumwelt aller Menschen.

Beide vorangehenden Beispiele aus der Wissenschaft sollen exemplarisch verdeutlichen, welche Relevanz einerseits Daten und Informationen für die wissenschaftliche Erkenntnisgewinnung und gesellschaftlichen Handlungsoptionen aufweisen. Auf Grundlage von Daten und Informationen können Zusammenhänge modelliert, evaluiert und interpretiert werden. Die Untersuchung von wissenschaftlichen Fragestellungen, die Entwicklung von Modellen oder die Vorhersage von komplexen Zusammenhängen sind ohne entsprechende Datenkompetenz daher nicht vorstellbar. Andererseits zeigen die Beispiele unterschiedliche Facetten eines Umgangs mit Messunsicherheiten: Ein korrekter Umgang mit Messunsicherheiten im Rahmen einer Messung gibt Aufschluss über den Gültigkeitsbereich des Ergebnisses und kann verwendet werden, um unter anderem Schwächen in theoretischen Modellen aufzufinden. Ein inadäquater bzw. fehlender Umgang mit Messunsicherheiten hingegen kann zu schwerwiegenden Folgen führen, indem wissenschaftliche Fehlschlüsse gezogen werden. Hieraus lässt sich die Bedeutsamkeit eines Verständnisses und adäquaten Umgangs mit Messunsicherheiten für die Wissenschaft und Gesellschaft ableiten.

Auch bei einer adäquaten Auswertung von experimentellen Daten unter Berücksichtigung von Messunsicherheiten ist nicht zweifelsfrei klar, welcher Hypothese Gültigkeit zugesprochen werden kann. Von einer Gültigkeit kann dann ausgegangen werden, wenn Wissenschaftler:innen sich selbst und andere in einem argumentativen Diskurs von *Claims* überzeugen (Driver et al., 2000). Dieser Prozess, in dem Argumente entstehen, wird Argumentation genannt. Unter einer Argumentation kann daher das Zusammenbringen und Inbeziehungsetzen von *claim* und Daten bezeichnet werden (Sampson & Clark, 2008; Toulmin, 1958).

Die dargestellten Gedanken haben gemeinsam, dass sie für eine fundierte Entscheidung einen adäquaten Umgang mit Daten voraussetzen. Die Vermittlung dieses Umgangs mit Daten und Unsicherheiten und von Datenkompetenz kann als genuiner Bestandteil des Bildungsauftrags von Schulen angesehen werden. Über die Förderung von Datenkompetenz können Aspekte wie *21st Century Skills* und *Data Literacy* gefördert werden (bspw. Ananiadou & Claro, 2009; Gibson & Mourad, 2018), welche als notwendig für die Bildung von mündigen Bürger:innen angesehen werden können (KMK, 2004; NRC, 2013), die im weiteren Verlauf ihres Lebens wissenschaftliche Erkenntnisse nachvollziehen und an der fundierten und datenbasierten Entscheidungsfindung im gesellschaftlichen und politischen Kontext partizipieren müssen.

Diese Ziele sind ebenfalls im Physikunterricht der Schule verankert. Durch die Vermittlung von *Scientific Practices*, wie beispielweise das Argumentieren und Erklären (Berland & Reiser, 2009), und einer *Scientific Literacy*, unter anderem Aspekte wie das Verständnis von wissenschaftlicher Evidenz und Arbeitsweise (Gott & Duggan, 1996), sollen die Schüler:innen zum Anfertigen von evidenzbasierten Aussagen und zum Argumentieren angeregt werden (Schiepe-Tiska et al., 2012). In diesem Zusammenhang erweist sich das Experiment als eine gute Gelegenheit zum Argumentieren und datenbasierten Begründen: Einerseits können beim Experiment Hypothesen und Daten quasi „in situ“ erzeugt werden (Toulmin, 1958). Andererseits sind Messdaten stets mit einer Messunsicherheit versehen, sodass für einen adäquaten Umgang mit Daten im experimentellen Kontext auch ein Verständnis für Messunsicherheiten relevant ist.

Trotz der aufgezeigten Relevanz zeigt sich in der Schule ein anderer Zustand: Sowohl das Argumentieren als auch der adäquate Umgang mit Daten wird in der gegenwärtigen Form in der Schule vernachlässigt (bspw. Hellwig, 2012).

Vor diesem Hintergrund ist das Ziel der Arbeit, a) digitale Lernapps zur Förderung der Datenkompetenz von Schüler:innen zu entwickeln, b) die Wirksamkeit dieser Apps zu prüfen und c) zu evaluieren, inwiefern eine Förderung der Datenkompetenz die Argumentationsfähigkeit von Schüler:innen beim Experimentieren beeinflusst. Zwei empirische Studien mit $n_1 = 138$ und $n_2 = 325$ Schüler:innen wurden zur Untersuchung der dargestellten Forschungsziele durchgeführt.

Die vorliegende Arbeit trägt dazu bei, anhand einer didaktischen Struktur Lerngelegenheiten in Form von digitalen Lernapps zu entwickeln und zu evaluieren sowie aus einer physikdidaktischen Perspektive empirisch Wirkzusammenhänge zum Einfluss von Datenkompetenz auf das Argumentieren aufzuklären. Somit wird die Arbeit dem eingangs skizzierten Anspruch an eine naturwissenschaftliche Bildung, bei der das Experiment, der Umgang mit Daten und Messunsicherheiten sowie das Argumentieren stärker im Mittelpunkt stehen, gerecht.

Anmerkung zur Notation In der vorliegenden Arbeit wird Wert auf inklusive Sprache gelegt. Daher findet die Kurzschreibweise mit einem Doppelpunkt, wie beispielsweise „Schüler:innen“, Verwendung. Diese Schreibweise findet der Übersicht wegen ebenfalls in Fällen Anwendung, in denen sich aus der Kurzschreibweise heraus nicht die korrekte Form der Substantive, wie beispielsweise „den Schüler:innen“, bilden lässt.

Kapitel 2

Theoretischer Hintergrund und Stand der Forschung

2.1 Das naturwissenschaftliche Argumentieren

Naturwissenschaftlicher Unterricht hat unter anderem zum Ziel, eine naturwissenschaftliche Grundbildung bei Schüler:innen zu fördern. Diese Grundbildung, welche auch als eine *Scientific Literacy* bezeichnet werden kann, beinhaltet ebenfalls, dass Schüler:innen Ansichten und Aussagen evidenzbasiert begründen (Schiepe-Tiska et al., 2012). Dieses Ziel kann beispielsweise durch das Argumentieren gefördert werden (Osborne et al., 2013). Ferner kann das Argumentieren ebenfalls im Unterricht genutzt werden, um das Lernen über Naturwissenschaften bzw. Physik zu fördern (Manz, 2015). Dem Argumentieren kann somit ein hohe Relevanz für den naturwissenschaftlichen Unterricht zugewiesen werden und ist daher seit einiger Zeit Gegenstand fachdidaktischer Forschung. Dabei ergibt sich das Problem, dass es keine einheitliche Verwendung dieses Begriffs bzw. Prozesses gibt (Osborne & Patterson, 2011). Hinzu kommt, dass das Verständnis einer Argumentation lange mit dem Verständnis reiner Logik verknüpft war (Hintikka, 1989). Das Argumentieren wurde erst vergleichsweise spät in alltäglichen und nicht rein wissenschaftlichen Kontexten untersucht und analysiert (Kuhn, 1991; Toulmin, 1958). Dabei schlägt Toulmin (1958) unter anderem eine strukturelle Analyse von Argumenten vor, in der die funktionalen Einheiten von Argumenten differenziert und hinsichtlich ihrer Wirkung kategorisiert werden (Osborne & Patterson, 2011; Toulmin, 1958). Aus diesem Grund wird zunächst eine Definition der Begriffe Argumentation, Argument und argumentieren etabliert, bevor auf tiefergehende Aspekte und den Stand der Forschung eingegangen werden kann.

2.1.1 Definition zentraler Begriffe

In der deutschen Sprache bezeichnet die Argumentation eine Beweisführung oder das Vortragen von Argumenten (Duden, 2023). Dieses Verständnis von Argumentation ist sehr allgemein und zu unspezifisch für die vorliegende Arbeit und den naturwissenschaftlichen Kontext. Zudem sind das Verständnis eines Beweises und die Validität von Evidenz von der entsprechenden Fachrichtung abhängig (Toulmin, 1958). Für die vorliegende Arbeit wird sich daher an dem Verständnis von naturwissenschaftlicher Argumentation orientiert: Im naturwissenschaftlichen Kontext wird im Rahmen der Argumentation versucht, eine Behauptung auf der Grundlage von Evidenzen zu unterstützen, infrage zu stellen oder zu präzisieren (S. Norris et al., 2007). Die Evidenz kann dabei explizit aus Daten oder aus Schlussfolgerungen generiert werden (Osborne, 2012; Toulmin, 1958). Nachfolgend wird das Argumentieren als Prozess verstanden, in dem Argumente kreiert und gerechtfertigt werden (Osborne, 2012). Argumentation stellt das entsprechende Substantiv dar, welches die Situation beschreibt, in der das Argumentieren stattfinden kann. Ein Argument kann somit als Ergebnis des Argumentierens angesehen werden (Kuhn & Udell, 2003).

Grundsätzlich besteht ein Argument aus einer Behauptung und einer Evidenz. Dabei dient die Evidenz der Rechtfertigung der Behauptung (Osborne et al., 2004). Die Evidenz oder Begründung der Behauptung kann in empirischer und quantitativer Form vorliegen, beispielsweise in Form von Daten oder experimentellen Beobachtungen (Iordanou & Constantinou, 2015; Kuhn, 1991; Zohar & Nemet, 2002). Zusätzlich können Argumente hinsichtlich ihrer Plausibilität und Verträglichkeit mit Fakten bewertet werden (Koslowski, 1996; Manz, 2015). Daraus lässt sich der Gedanke ableiten, dass ein Argument die sinnvolle Verbindung von theoretischen Claims, die sich wissenschaftlich motivieren lassen, mit Evidenz ist. Beispielsweise greifen verschiedene Arbeiten die dargestellten Aspekte auf, um Bewertungsraster bezüglich der epistemischen Ebene und der Qualität von Argumenten zu erstellen und Argumente von Lernenden entsprechend zu klassifizieren (Kelly & Takao, 2002; Kuhn & Udell, 2003; Sandoval & Millwood, 2005). Das Element der Begründung besitzt hierbei einen besonderen Stellenwert, da es die Evidenz und die Behauptung in Beziehung zueinander setzt (Toulmin, 1958). Auf die Struktur und Analyse von Argumenten wird im Abschnitt 2.1.4 genauer eingegangen.

Neben den verschiedenen Charakterisierungen von Argumenten kann das Argumentieren ebenfalls unterschiedlich erläutert werden und unterschiedliche Zwecke erfüllen (bspw. Aydeniz & Ozdilek, 2015; D. Cohen, 1995; Duschl & Osborne, 2002; Kuhn et al., 2000; Sampson & Clark, 2008). Die verschiedenen Sichtweisen haben nach Aydeniz und Ozdilek (2015) die Aspekte „reasoned discourse“ und „justification of claims to knowledge through use of scientific

evidence“ gemeinsam. Das Argumentieren kann somit als epistemisches Instrument gesehen werden, welches einen zentralen Prozess in der wissenschaftlichen Erkenntnis darstellt (Kuhn, 1991; Osborne, 2012). Beispielsweise berichten Forscher:innen am CERN, dass drei neue exotische Teilchen entdeckt worden sind (CERN, 2022). Während die Existenz der Teilchen bereits in den 1960er Jahren vorhergesagt wurde, konnte erst durch die aktuelle Messung entsprechende Evidenz für die Existenz eines Pentaquarks, das einen Strange-Quark beinhaltet, gefunden werden. Durch diese Messung konnte der „Claim“, dass dieses Teilchen existiert, durch Evidenz, die das in der Wissenschaft anerkannte Signifikanzniveau von fünf Standardabweichungen deutlich überschreitet, begründet werden. Dieses Beispiel illustriert, wie die Zusammenführung von theoretischen Vorhersagen und datenbasierter Evidenz zu wissenschaftlicher Erkenntnis führen kann.

Neben dem Argumentieren gibt es weitere epistemische Diskursaktivitäten, wie zum Beispiel das Erklären (bspw. Ohlsson, 1996). Osborne und Patterson (2011) weisen darauf hin, dass trotz der Unterschiede zwischen den beiden Tätigkeiten in vielen Forschungsarbeiten das Argumentieren und das Erklären gleichgesetzt werden. In Abgrenzung zum Argumentieren ist beim Erklären das Explanandum bekannt, wahr und oftmals wird eine eindeutige Situation angeführt. Beispielsweise untersuchte Wagner (2018), wie Schüler:innen unter Verwendung von Modellen das Phänomen der optischen Hebung erklären. Der fachliche Zusammenhang ist eindeutig. Das Ziel der Erklärung ist, zu beschreiben, *wie* und *warum* die optische Hebung für die Beobachter:innen entsteht. Dem Argumentieren hingegen geht eine strittige Situation voraus und hat eine Überzeugungsabsicht bezüglich der eigenen Behauptung inne, die sowohl auf intra- als auch auf interpersoneller Ebene liegen kann (Ford, 2012). Die beiden diskursiven Entitäten Argumentieren und Erklären unterscheiden sich somit darin, dass eine Argumentation eine Erklärung hinsichtlich der Validität und Gültigkeit bewertet (Osborne & Patterson, 2011). Im vorigen Beispiel von Wagner (2018) könnten Schüler:innen in Diskussionen versuchen, sich gegenseitig von den konkurrierenden Erklärungen des Phänomens der optischen Hebung zu überzeugen. Somit kann eine Argumentation unter anderem auch als die Suche nach der besten Erklärung verstanden werden (Osborne & Patterson, 2011).

2.1.2 Argumentieren im naturwissenschaftlichen Unterricht

Der vorangegangene Abschnitt hat die besondere Bedeutung des Argumentierens in der Wissenschaft hervorgehoben. Neben dem Finden von Erklärungen und Modellierungen von naturwissenschaftlichen Phänomenen kann Argumentieren als zentraler Teil von *Scientific Practices* (Berland & Reiser, 2009; Jiménez-Aleixandre & Crujeiras, 2017) und *Scientific Literacy* (Schiepe-Tiska et al., 2012)

angesehen werden und besitzt daher eine große Bedeutung für Schüler:innen, um sowohl ein korrektes Bild vom Wesen der Naturwissenschaften zu zeichnen als auch ein kritisches Hinterfragen und rationales Entscheiden der Schüler:innen zu fördern (Driver et al., 2000; McComas, 2004). Darüber hinaus gibt es eine Vielzahl weiterer Aspekte, die die Bedeutung des Argumentierens für Schüler:innen hervorheben (bspw. Erduran & Jiménez-Aleixandre, 2008). Es lässt sich somit ein Desiderat von fachdidaktischen Forscher:innen ableiten, dem Argumentieren im naturwissenschaftlichen Unterricht eine größere Relevanz zukommen zu lassen (Kuhn, 1992; Newton et al., 1999; Sandoval & Millwood, 2005). Dieser Forderung kommen die deutschen Bildungsstandards des Fachs Physik in Teilen nach. Die Bildungsstandards für den Mittleren Schulabschluss adressieren das Argumentieren nur implizit, beispielsweise lautet der Standard K7 wie folgt: „Die Schülerinnen und Schüler diskutieren Arbeitsergebnisse und Sachverhalte unter physikalischen Gesichtspunkten“ (KMK, 2004, S. 12). In der Fassung für die Allgemeine Hochschulreife wird das Argumentieren öfter und explizit genannt. Beispielsweise wird der Standard K8 wie folgt formuliert: „Die Lernenden nutzen ihr Wissen über aus physikalischer Sicht gültige Argumentationsketten zur Beurteilung vorgegebener und zur Entwicklung eigener innerfachlicher Argumentationen“ (KMK, 2020, S. 17). Somit ist das Argumentieren nach bildungspolitischen Rahmen als ein Ziel des naturwissenschaftlichen Unterrichts deklariert. Ähnliche bildungspolitische Richtlinien finden sich ebenfalls in internationalen Bildungszielen (bspw. NRC, 2013; OECD, 2018).

Vor dem Hintergrund der genannten formalen und normativen Gründe sollte das Argumentieren eine zentrale Rolle im naturwissenschaftlichen Unterricht einnehmen. Empirische Studien zeigen jedoch ein anderes Bild: Erduran und Jiménez-Aleixandre (2008) zeigen zum Beispiel auf, dass der naturwissenschaftliche Unterricht weiterhin eine starke Ergebnisorientierung aufweist. Schüler:innen und Lehrer:innen sind einen lehrerzentrierten Naturwissenschaftsunterricht gewohnt, in dem Fakten von Lehrer:innen oder Schulbüchern auswendig gelernt werden (Choi et al., 2015). Im Zentrum steht oftmals der oder die Lehrer:in, die oder der den Unterricht lenkt und häufig „triadische Dialoge“ etabliert (Duschl & Osborne, 2002; Lemke, 1990). Beispielhaft für die triadischen Dialoge sind Unterrichtsmuster, in denen der oder die Lehrer:in initiiert, die Schüler:innen antworten und der oder die Lehrer:in die Antwort evaluiert. Dieses Vorgehen verdeutlicht exemplarisch die vorherrschende Lehrer:inzentrierung. Außerdem werden von Lehrer:innen oft „kochbuchartige“ Rezepte zum Problemlösen verwendet, die mehr nach Routine als kritischem Hinterfragen verlangen (Grooms et al., 2018). Gleichzeitig herrscht im Unterricht ein positivistisches Bild vor, welches dem unsicheren, diskursiven und vorläufigen Charakter der naturwissenschaftlichen Erkenntnisgewinnung entgegensteht (Driver et al., 2000; Osborne, 2012). Es wird ein Bild vermittelt, in dem „scientific claims are seen as logically and self-evidently deducible from a limited set of empirical premises“

(Osborne, 2012, S. 935). Es kann konstatiert werden, dass der naturwissenschaftliche Unterricht in der dargestellten Form nicht die Möglichkeiten bietet, das naturwissenschaftliche Argumentieren in den Unterricht zu integrieren (McNeill et al., 2016; S. P. Norris et al., 2008). Abd-El-Khalick (2008) stellt dem jedoch entgegen, dass berücksichtigt werden müsse, dass (Natur-)Wissenschaft und naturwissenschaftlicher Unterricht per se nicht unmittelbar dieselben Ziele verfolgen. Während die Ziele durchaus Unterschiede aufweisen, da in der Wissenschaft beispielsweise der Erkenntnisgewinn eine hohe Bedeutung hat und im naturwissenschaftlichen Unterricht der Bildungsauftrag bezüglich der Schüler:innen im Vordergrund steht, zeigen sich jedoch Überschneidungen in der Methodik: Die Ausbildung einer *Scientific Literacy* und von *Scientific Practices* ist Teil des Bildungsauftrags und ebenso ein essenzieller Bestandteil der (natur-)wissenschaftlichen Vorgehensweise von Forscher:innen. Das Experimentieren beispielsweise benötigt sowohl in der Wissenschaft als auch beim Einsatz in der Schule ähnliche Methoden für die adäquate Auswertung. Im Unterricht kann zwischen einer induktiven und deduktiven Vorgehensweise unterschieden werden, welche in der Wissenschaft eher weniger vorliegt. Wird das Experiment jedoch zum Erkenntnisgewinn eingesetzt, adressiert die Auswertung ähnliche Kompetenzen in beiden Bereichen. In Abschnitt 2.2.1 wird auf die Bedeutung des Experiments und die Rolle des naturwissenschaftlichen Experimentierens im Physikunterricht eingegangen und der Bezug zum Argumentieren verdeutlicht. Es ist daher nicht von der Hand zu weisen, dass die Naturwissenschaft per se und der naturwissenschaftliche Unterricht trotz der von Abd-El-Khalick (2008) genannten unterschiedlichen Ziele dennoch Gemeinsamkeiten aufweisen, zu denen, wie zuvor dargelegt, auch das Argumentieren zu zählen ist.

Verschiedene Forschungsarbeiten wie beispielsweise von Osborne (2010) zeigen daher, dass die Partizipiation von Schüler:innen in Argumentationen Teil von naturwissenschaftlichem Unterricht sei. Einen weiteren Grund zeigt Osborne (2010) in der Sichtweise auf die Schule, die sehr auf den Transfer von Wissen und Fakten statt auf den Erkenntnisgewinn konzentriert ist:

> Deep within our cultural fabric, education is still seen simplistically as a process of transmission where knowledge is presented as a set of unequivocal and uncontested facts and transferred from expert to novice. In this world-view, failure of communication is the exception and success the norm. (Osborne, 2010, S. 463)

Verschiedene fachdidaktische Forschungsarbeiten zeigen auf, dass Gelegenheiten für das Argumentieren im naturwissenschaftlichen Unterricht geschaffen werden können, wenn der Fokus des Unterrichts auf die Konstruktion von Wissen durch Analyse und Interpretation von Daten gelegt und Lernenden die Möglichkeit zur Teilnahme an diesen Tätigkeiten gegeben wird (Duschl et al., 2007; NRC, 2013; Osborne, 2012).

Dabei ist jedoch zu beachten, dass das wissenschaftliche Argumentieren drei Aspekte aufweist (Duschl, 2008): Der *kognitive* Aspekt des wissenschaftlichen Argumentierens betrachtet beispielsweise, inwiefern Begründungen und Evidenzen generiert und mit den *Claims* der Argumente verbunden werden. Beim *epistemischen* Aspekt des Argumentierens stehen das kritische Hinterfragen von Daten und Evidenzen oder die Verträglichkeit der im Rahmen der Argumentation hervorgebrachten Ideen mit bestehenden wissenschaftlichen Theorien im Mittelpunkt. Der *soziale* Aspekt des Argumentierens bezeichnet die Interaktivität und Intensität von Diskussionen in Gruppen und weist somit einen großen Einfluss auf die interpersonelle Ebene des Argumentierens auf. Um das Argumentieren in den naturwissenschaftlichen Unterricht zu integrieren, müssen Schüler:innen gleichzeitig an den kognitiven, epistemischen und sozialen Aspekten des Argumentierens teilnehmen (Grooms et al., 2018).

Zusammenfassend wird ein Wandel des Unterrichts gefordert, der einerseits *Scientific Practices* bei den Schüler:innen fördert und andererseits eine *Scientific Literacy* bei den Schüler:innen ausbildet, damit Lernende dazu befähigt werden, den gesellschaftlichen und globalen Herausforderungen der Menschheit, in Forschungsarbeiten als *Socioscientific Issues* bezeichnet, begegnen zu können (z. B. Council et al., 2008; Driver et al., 2000; Feinstein, 2011; Feinstein et al., 2013; Sadler, 2004).

2.1.3 Überblick über den Stand der Forschung zur Förderung des Argumentierens im naturwissenschaftlichen Unterricht

In dem Spannungsfeld zwischen den „guten Gründen“ für das Argumentieren im naturwissenschaftlichen Unterricht und dem aus dieser Perspektive mangelhaften Ist-Zustand hat eine Vielzahl von fachdidaktischen Arbeiten Ansätze zur Förderung des Argumentierens entwickelt und evaluiert (Abd-El-Khalick, 2008; Duschl, 2008; Duschl et al., 2007; Nussbaum & Bendixen, 2003). In den folgenden Abschnitten sollen die zentralsten Arbeiten dargestellt werden. Dabei wird der Fokus auf Randbedingungen beim naturwissenschaftlichen Argumentieren gelegt. Ausführliche Arbeiten zum Argumentieren im naturwissenschaftlichen Unterricht finden sich beispielsweise bei Erduran und Jiménez-Aleixandre (2008).

Schüler:innenebene

Eine Reihe von Arbeiten konnte aufdecken, dass Schüler:innen trotz Möglichkeiten zum wissenschaftlichen Diskurs im Unterricht die Fähigkeiten zum Argumentieren fehlen (Driver et al., 2000; Osborne, 2010; Sandoval & Millwood, 2005). In ihren Untersuchungen haben Sandoval und Millwood (2005) die schriftlichen Begründungen von Schüler:innen im Kontext der natürlichen Selektion analysiert. Dabei wurde eine Bewertungsmethode für Argumente vorgestellt,

die unter anderem die in den Argumenten herangezogene Evidenz evaluiert. Es konnte gezeigt werden, dass Schüler:innen zwar durchaus ein Bewusstsein für die Verwendung von Daten als Evidenz aufweisen, jedoch die herangezogene Evidenz oftmals nicht ausreichend war. Die Lernenden waren somit nicht in der Lage, ihre Behauptungen entsprechend mit Daten zu rechtfertigen. In einer Analyse von mehreren eigenen Arbeiten konnte Kelly (2008) zeigen, welche Auswirkungen es auf die Student:innen hat, wenn sie sich im Rahmen des Unterrichts mit dem Anfertigen von evidenzbasierten Aussagen beschäftigen. Dazu wurden mehrere Datensätze aus verschiedenen Studien untersucht. Die Lernenden sollten dabei im Fach Ozeanographie lernen, wissenschaftliche Evidenz schriftlich anzufertigen. Die Studien stammen jeweils aus dem gleichen Kurs, der über die Jahre überarbeitet wurde. Die Analyse der Ergebnisse hat gezeigt, dass durch das schriftliche Argumentieren Aspekte, wie beispielsweise *Scientific Practices* oder eine adäquate Sicht auf die Naturwissenschaften, gefördert werden können.

Die vorangehend exemplarisch dargestellten Arbeiten zeigen auf, dass durch Maßnahmen auf Ebene der Schüler:innen das Argumentieren adressiert und gefördert werden kann. Somit ergibt sich auf der Ebene der Schüler:innen eine Handlungsoption. Dabei kann grundsätzlich zwischen direkten und indirekten Ansätzen zur Förderung der Argumentationsfähigkeit bei Schüler:innen unterschieden werden.

Beispielsweise haben Bell und Linn (2000) in einer Projektarbeit die Argumente von Schüler:innen analysiert. Dabei gingen die Schüler:innen der Frage „How far does light go?“ (Bell & Linn, 2000, S. 798) nach. Im Rahmen der Arbeit wurde ein didaktischer Rahmen zur Förderung des wissenschaftlichen Argumentierens und des Verständnisses von (Natur-)Wissenschaft evaluiert. Die Arbeit hebt sich durch den Einsatz von Computermedien und dem Internet als Quellen hervor. Die Argumentation der Schüler:innen wurde durch direkte Anweisungen im Projekt initiiert. Es konnte gezeigt werden, dass die Qualität der Argumente gesteigert werden konnte, wenn die entwickelte Lernumgebung von den Schüler:innen verwendet worden ist.

Ähnliche Ergebnisse konnte Ratcliffe (1996) zeigen. Schüler:innen wurden im Rahmen ihrer Studie in Diskussionen über socioscientific issues eingebunden. Dabei wurden die Schüler:innen anhand eines didaktischen Rahmens in sechs Schritten instruiert. Obwohl der Schwerpunkt der Intervention die informationsbasierte Entscheidungsfindung der Schüler:innen war, konnte bezüglich der Argumentation ebenfalls eine Beobachtung getätigt werden. Schüler:innen, die bereits durchdachte Entscheidungsfindungsstrategien entwickelt hatten, konnten in der Diskussion ihre Entscheidungen mit validen Argumenten verstärken.

In einer großen Studie an 14 Schulen untersuchte Solomon (1992) Diskussionen im Zusammenhang mit *Socioscientific Issues*. Dabei wurde eine von den Schüler:innen geführte Diskussion initiiert. Die Ergebnisse deuten darauf hin, dass die zunehmende Erfahrung der Schüler:innen in der Teilnahme von

Experimenten zu einer höheren Qualität der Diskussion und Argumente führt.

Die dargestellten Arbeiten sollen exemplarisch zeigen, welche Effekte das direkte und indirekte Fördern der Argumentationsfähigkeit bei Schüler:innen haben kann. Sowohl direkte Instruktionen als auch eine Unterstützung der Schüler:innen durch bestimmte Aufgabenformate, Gruppenarbeiten oder Hilfestellungen kann das Argumentieren der Schüler:innen auf struktureller und qualitativer Ebene verbessern (Berland & Reiser, 2009; Iordanou & Constantinou, 2015; Jiménez-Aleixandre & Crujeiras, 2017; Patterson, 2001). Auf Grundlage der dargestellten Arbeiten wird deutlich, dass es möglich ist, das Argumentieren auf der Ebene der Schüler:innen zu adressieren und erfolgreich zu fördern. Jedoch lässt sich erkennen, dass die genannten Aspekte ebenfalls einen Anspruch an die Lehrkraft stellen. Der oder die Lehrer:in steht insofern in der Verantwortung, die genannten Aspekte, wie beispielsweise direkte Instruktionen zum Argumentieren oder das Initiieren von Diskussionen, in den Unterricht zu integrieren, damit die Schüler:innen die Argumentationskompetenz stetig verbessern können. Neben der Ebene der Schüler:innen kann daher ebenso die Ebene der Lehrer:innen adressiert werden. Diese besitzen einen großen Einfluss auf den Unterricht und haben daher eine eminent wichtige Rolle darin, Argumentationsanlässe im Unterricht zu generieren und gleichzeitig die Argumentationsfähigkeit der Schüler:innen zu fördern. Durch die Unterrichtsplanung, didaktische Entscheidungen, die Festlegung von Sozialformen oder andere Faktoren werden wesentliche Elemente des Unterrichts durch die Lehrer:innen festgelegt.

Die vorangegangenen Ausführungen zur Förderung von Argumentationen auf der Schüler:innenebene legen nahe, dass beispielsweise die Einbindung von Instruktionen und Aufgaben in den Unterricht, die das Argumentieren direkt adressieren, für die Lehrkräfte ein Instrument zur Förderung der Argumentation im Unterricht sein können.

Einflüsse der Sozialform auf das Argumentieren sind ebenfalls Gegenstand der didaktischen Forschung. Beispielsweise haben Sampson und Clark (2009) in ihrer Studie untersucht, inwiefern sich die Argumentqualität von einzelnen Schüler:innen gegenüber Gruppen unterscheidet. Es konnte festgestellt werden, dass es bezüglich der Argumentqualität keine Unterschiede gibt, jedoch der persönliche Fortschritt bezüglich der bearbeiteten Aufgaben der Schüler:innen positiv von der Gruppenarbeit profitierte.

Asterhan (2018) hat die Argumentation von Schüler:innen zu Themen der Astronomie und Wirtschaft untersucht. Dabei wurden jeweils Schüler:innen zu einer Diskussion zusammengeführt, welche unterschiedliche Standpunkte zu den Themen eingenommen hatten. Die Ergebnisse der Arbeit legen nahe, dass die Partnerarbeit alleine nicht ausreichend für einen argumentativen Austausch der Schüler:innen ist. Andere Faktoren wie beispielsweise leistungsorientierte Ziele oder hohes Selbstbewusstsein der Schüler:innen stellen sich in diesem Zusammenhang als stärkere Prädiktoren heraus. Die exemplarisch dargestellten Arbeiten

zeigen, dass die Festlegung der Sozialform beim Argumentieren im Unterricht nicht unmittelbar zu a) mehr Argumentationen zwischen den Schüler:innen und b) einer höheren Argumentqualität bei den Lernenden führt.

Zudem schlagen Berland und McNeill (2010) vor, das Ziel des Argumentierens schrittweise zu erreichen und ebenfalls Vorstufen des Argumentierens wie beispielsweise die Sinnfindung im Rahmen des Kontexts oder wissenschaftliche Artikulation der Schüler:innen zu fördern und dabei unter anderem auf vereinfachte Aufgabenstellungen zu setzen. Es kann dabei unterstützend wirken, sich nicht nur auf eine Art des wissenschaftlichen Diskurses festzulegen (McDonald & Kelly, 2012).

Berland und Reiser (2009) zeigen auf, dass die Einbindung von wissenschaftlichen Untersuchungen in den Unterricht zur Förderung von *Scientific Practices* mit Schwierigkeiten verbunden sein kann. Eine erfolgreiche Einbindung einer solchen Untersuchung kann sich jedoch positiv auf das genannte Ziel auswirken: Beispielsweise konnten H.-T. Chen et al. (2016) zeigen, dass durch eine argumtationsgestützte wissenschaftliche Untersuchung in der Grundschule eine signifikante Steigerung der Argumentqualität als auch beim Lernen von Wissenschaft bei den Schüler:innen entstehen kann. Neben der Förderung der Argumentationsfähigkeit durch einzelne Bestandteile des Unterrichts oder der gesamten Unterrichtsgestaltung schlagen Forscher:innen die Verwendung von Curricula vor, die zur Förderung der Argumentationsfähigkeit der Schüler:innen verwendet werden können (bspw. Berland & McNeill, 2010; Lehesvuori et al., 2013). Beispielsweise präsentieren H.-T. Chen et al. (2016) ein zwölfwöchiges Programm, in dem die Schüler:innen wöchentlich zum Argumentieren angeregt werden. Die Arbeit von Finn et al. (2006) stellt mehrere Lernumgebungen zur Verfügung, in der die Schüler:innen der Frage „Where have all the creatures gone?“ nachgehen. Die einzelnen Lernumgebungen haben unter anderem das Ziel, das evidenzbasierte Argumentieren der Schüler:innen zu fördern. Die Curricula beinhalten vorgefertigtes Material und Instruktionen für die Integration in den Unterricht durch die Lehrkraft. Der Ansatz, das naturwissenschaftliche Argumentieren im Unterricht durch den Einsatz von speziellen Unterrichtsmaterialien oder die Verwendung von vorbereiteten Instruktionen zu fördern, wird von verschiedenen Autor:innen auch unabhängig von Curricula vorgeschlagen (bspw. Manz & Renga, 2016; Osborne et al., 2004; Sampson et al., 2017; Shakespeare, 2004).

Lehrer:innenebene

Während die genannten Ansätze erneut die Schüler:innen im Fokus haben und Lehrer:innen Möglichkeiten in Form von Handlungsanweisungen, Instruktionen, Curricula sowie Materialien zur Förderung der Argumentationsfähigkeit durch die Etablierung einer „epistemischen Kultur“ (Manz, 2015) im Unterricht geben, gibt es ebenso Forschungsergebnisse, die die Lehrperson in den Fokus stellen. Einerseits zeigen von Aufschnaiter et al. (2008) auf, dass die Lehrer:innen stärker

die Lernvoraussetzungen der Schüler:innen insbesondere bezüglich des Argumentierens bei der Vorbereitung von Argumentationen berücksichtigen müssen. In ihrer Arbeit konnte gezeigt werden, dass Schüler:innen eher an Argumentationen teilnehmen, wenn die Abstraktionsschwelle sowie der Inhalt nah an den eigenen Fähigkeiten liegen. An diesem Umstand ändert auch die Bereitstellung von zusätzlichen Informationen in der Aufgabenstellung nur wenig, da die Schüler:innen die weiteren Informationen nicht verwendet haben.

Andererseits konnten verschiedene Arbeiten zeigen, dass sowohl die Förderung der Argumentationsfähigkeit bei Schüler:innen als auch die Vermittlung der Argumentationsfähigkeit durch das PCK der Lehrkraft begrenzt wird (Clarke et al., 2016; Wang & Buck, 2016). Dazu ist auch das wissenschaftliche Argumentieren zu zählen. Duschl und Osborne (2002) äußern sich dazu wie folgt:

> Before we can ask teachers to engage their students in argumentation and use the information they acquire from the process to plan subsequent lessons or evaluate students learning, it is essential to provide some theoretical guidance to answer such questions. (Duschl & Osborne, 2002, S. 62)

Es ist daher davon auszugehen, dass die erfolgreiche Etablierung und Einbindung des Argumentierens in den Unterricht ebenfalls von professionellen Handlungskompetenzen und dem didaktischen Fachwissen abhängig sein kann (Bauer et al., 2017). Auch wenn der Zusammenhang zwischen persönlichen epistomologischen Überzeugungen der Lehrer:innen und ihrer Argumentationsfähigkeit nicht geklärt ist (Liu & Roehrig, 2017), besteht weiterhin die Notwendigkeit, Lehrer:innen in Argumentationsmethoden zu fördern, um die Argumentation von Lernenden zu unterstützen (Duschl & Osborne, 2002; Ozdem Yilmaz et al., 2011). Außerdem konnten Christodoulou und Osborne (2014) sowie Clarke et al. (2016) zeigen, dass qualitative Unterschiede in der Sichtweise auf das Argumentieren zwischen „frischen“ und erfahrenen Lehrkräften existieren. Insbesondere Clarke et al. (2016) zeigen dabei Bereiche auf, in denen die Lehrkräfte Unterstützung brauchen. Manz und Renga (2016) haben Unterrichtsepisoden untersucht und dabei anhand eines Kodierschemas das epistemische Level der Episode bestimmt. Ziel der Untersuchung war es, ein Verständnis für die Evidenzkonstruktion im naturwissenschaftlichen Unterricht zu entwickeln. Auf Grundlage ihrer Beobachtungen konnten Manz und Renga (2016) didaktische Implikationen zur Förderung von Argumentationen und Evidenzkonstruktion, wie beispielsweise die Aktivität der Schüler:innen zu lenken, gezielte Fragen oder Meinungsverschiedenheiten zu fördern, ableiten. Die vorangehend dargelegten Forschungsergebnisse zeigen, dass Lehhrkräfte vor allem durch Fortbildungen oder eine überarbeitete Ausbildung in diesem Bereich unterstützt werden können.

Im Rahmen von verschiedenen Forschungsarbeiten wurden sowohl intensive mehrwöchige (bspw. Choi et al., 2015; Fishman et al., 2017; Zaccarelli et al., 2018) als auch begleitende Fortbildungen (bspw. Hilton et al., 2016; Zohar, 2007)

empirisch untersucht. Die Wirksamkeit der Fortbildungen konnte in beiden Fällen nachgewiesen werden. Dabei wurden den Lehrer:innen sowohl verschiedene Strategien als auch Materialen zur Verfügung gestellt, anhand derer die diskursiven Unterrichtsphasen gestaltet wurden. Eine exemplarische Übersicht zu Fortbildungen in diesem Kontext ist bei Pehmer et al. (2015) einsehbar.

Durch Fortbildungen alleine ist es jedoch nicht möglich, das didaktische Fachwissen der Lehrkräfte zu adressieren und zu fördern. Es besteht Handlungsbedarf in der Ausbildung der Lehrer:innen. Zu diesem Aspekt gibt es verschiedene Arbeiten, welche das Argumentieren von Referendar:innen untersucht haben. In diesem Zusammenhang benennt Osborne (2012) Fähigkeiten und Themen, die notwendig für Lehrkräfte sind, um das Argumentieren in den Unterricht einzubinden. Weitere Forschungsarbeiten deuten darauf hin, dass insbesondere das fachdidaktische Wissen bezüglich des Argumentierens adressiert werden muss: Einerseits ist damit das explizite Wissen über das Argumentieren und andererseits die Wahrnehmung und Adaption von Argumenten durch die Lehrkräfte gemeint (McNeill et al., 2016; Wang & Buck, 2016). Dabei konnten verschiedene Handlungsbarrieren festgestellt und Lösungsansätze auf Grundlage verschiedener Forschungsarbeiten benannt werden (Wang & Buck, 2016, S. 602).

Die dargestellten Schwächen konnten bereits bei Referendar:innen beobachtet werden (Acar, 2016). Beispielsweise konnten Cetin et al. (2014) zeigen, dass es bei Referendar:innen ebenfalls einen Zusammenhang zwischen dem Fachwissen und dem Argumentieren im Rahmen von socioscientific issues gibt. Damit zeigt sich ein Indiz, dass die bereits von Zohar und Nemet (2002) aufgezeigten Zusammenhänge durch die Lehramtsbildung nur gering beeinflusst werden. Eine Untersuchung von Yardimci Capkinoglu et al. (2022), welche durch *Peer Review* das Argumentieren von Referendaren fördern wollte, findet Anhaltspunkte dafür, dass die Argumentqualität zunimmt, wenn die Argumentation mit einem „Peer Review“ verbunden wird.

Zusammenfassend lässt sich sagen, dass es viele Möglichkeiten gibt, um das Argumentieren auf der Ebene der Lehrer:innen, Schüler:innen und der Lehrer:innenausbildung zu fördern. Die empirische Wirksamkeit der Argumentationsförderungsansätze ist nicht eindeutig. So gibt es umfangreiche Studien, welche am Ende nur eine geringe Wirkung aufweisen (Osborne et al., 2013). Solche Studien fügen sich gut in das Bild ein, das beispielsweise von Asterhan und Schwarz (2016) gezeichnet wird.

In der vorliegenden Arbeit wird ein Ansatz verwendet, welcher sich aus der Arbeit von Ludwig (2017) ableitet. Ludwig (2017) hat in seiner Arbeit den Einfluss personaler und situationaler Faktoren auf das Argumentieren beim Experimentieren untersucht.

Dazu haben die Schüler:innen in realen und virtuellen Experimenten den funktionalen Zusammenhang zwischen der Periodendauer T eines Fadenpendels und der Pendelmasse m untersucht. Die dazugehörige Hypothese der Lernenden

wurde vor und nach dem Experiment erhoben. Das Experiment war den Schüler:innen unbekannt und daher war der ermittelte Datensatz die einzige Evidenz, die herangezogen werden konnte, um den Zusammenhang zwischen den physikalischen Größen zu bestimmen. Es konnte unter anderem gezeigt werden, dass das Fachwissen der Schüler:innen eine positive Wirkung auf eine wissenschaftlich adäquate Art der Rechtfertigung besitzt und intuitive Rechtfertigungen eher hemmt (vgl. Abschnitt 2.1.4). Außerdem zeigte sich im untersuchten Kontext, dass eine evidenzbasierte Begründung mit einer höheren Chance zu einem Lernerfolg führt. In der Studie von Ludwig (2017) wurde theoretisches Wissen der Mechanik erhoben. Verschiedene Forschungsarbeiten legen jedoch nahe, dass für das wissenschaftliche Argumentieren auf Grundlage von experimentellen Daten Datenkompetenz relevant ist und einen positiven Einfluss auf datenbasierte Begründungen aufweist (bspw. Gott & Duggan, 2007; Kanari & Millar, 2004). Pols (2019) konnte aufzeigen, dass Schüler:innen ohne entsprechende Datenkompetenz Behauptungen nicht belegen und keine datenbasierten Schlussfolgerungen ziehen konnten. Wiederum zeigen weitere Forschungsarbeiten, dass das Adressieren und Fördern von Kompetenzen im Umgang mit Daten und Unsicherheiten das Argumentieren und die Argumentationsqualität fördern kann (Donner, 2019; Kok, 2022).

Vor diesem Hintergruwnd wird daher für den weiteren Verlauf der Arbeit angenommen, dass auf Grundlage der Ergebnisse von Ludwig (2017) das Argumentieren beim Experimentieren durch die Datenkompetenz adressiert und gefördert werden kann. Eine theoretische Darstellung zur Bedeutung der Datenkompetenz im naturwissenschaftlichen Unterricht kann Abschnitt 2.3 entnommen werden. Der gewählte Ansatz adressiert hauptsächlich die Schüler:innen und wird in Form von kurzen Interventionseinheiten umgesetzt, welche durch gezielte Förderung von Kompetenzen zum Umgang mit Daten und Unsicherheiten die Argumentationsfähigkeit der Schüler:innen adressieren und fördern sollen.

Bei der Gestaltung der Intervention wird sich an Forschungsarbeiten orientiert, welche den Einsatz von Lernumgebungen adressiert haben (bspw. Bell & Linn, 2000; Iordanou & Constantinou, 2015). Vor dem Hintergrund, dass die Praktikabilität von vorgefertigen curricularen Einheiten sowie langen Interventionen sehr gering ist, werden kurze und modulare Lerneinheiten konzipiert, welche kontextfrei und daher gut in den Unterricht einzubringen sind. Obwohl die Intervention und die Apps methodisch auf der Schüler:innenebene operieren, können Lehrer:innen ebenfalls von den Apps profitieren. Durch die Kontextfreiheit, Modularität und Kürze ist eine Integration in den Unterricht einfach möglich (Kardas & Ludwig, 2021).

2.1.4 Analyse von Argumenten

Im vorangehenden Abschnitt wurde eine Übersicht zur Förderung der Argumentationsfähigkeit im Unterricht gegeben. In den dargestellten Forschungsarbeiten wurde die Wirksamkeit der Untersuchungen durch die Evaluation der entsprechenden Argumente bewertet. Die vorliegende Arbeit hat ebenfalls zum Ziel, das Argumentieren der Schüler:innen zu fördern. Die Evaluation von Argumenten nimmt daher eine zentrale Rolle bei der Bewertung der Wirksamkeit der Intervention ein. Vor diesem Hintergrund erfolgt eine kurze Übersicht über Verfahren zur Evaluation von Argumenten. Abschließend wird dargestellt, wie die Wirksamkeit bezüglich der Förderung der Argumentationsfähigkeit in der vorliegenden Arbeit bewertet wird.

Sampson und Clark (2008) stellen dar, dass Argumente grundsätzlich auf einer strukturellen, inhaltlichen und „nature of justification“ (Art der Rechtfertigung) Ebene untersucht werden können. Verschiedene Rahmen, welche von Forscher:innen verwendet werden, lassen sich stets einem oder mehreren Aspekten zuordnen. Als struktureller Ausgangspunkt wird oftmals das Schema von Toulmin (1958) verwendet, welches die Bestandteile eines Arguments charakterisiert. Der Schwerpunkt der inhaltlichen Analyse liegt in der Naturwissenschaftsdidaktik darin, ob das Argument fachlich korrekt ist. Bei einer Beurteilung der Art der Begründung wird untersucht, welche Informationen verwendet werden, um die Behauptung innerhalb des Arguments zu stützen.

Strukturelle Analyse von Argumenten

Die Evaluation auf Grundlage der strukturellen Ebene findet in vielen Arbeiten Anwendung. In dem von McNeill und Krajcik (2011) vorgeschlagenen Evaluationsvorgehen werden unter anderem strukturelle Elemente der Argumente ausgewertet. Zudem kann anhand der Bestandteile des Arguments eine Aussage über die Komplexität und Persuasion von Argumentationen bewertet werden. In ihrer Arbeit untersuchen McNeill und Krajcik (2011) die Elemente „claim, evidence, reasoning“ und „rebuttal“. *Claim* bezeichnet in diesem Fall die Aussage oder Folgerung, die sich auf die Problemstellung bezieht. Unter *Evidence* werden wissenschaftliche Daten verstanden, welche den *Claim* unterstützen. Die Verbindung von *Claim* und *Evidence* kann als *Reasoning* angesehen werden. Es geht dabei um die Rechtfertigung des *Claim* auf Grundlage der *Evidence*. *Rebuttal* umfasst einerseits das Beschreiben von weiteren Erklärungen zu dem Problem und andererseits das Anfertigen von Gegenargumenten bzw. Evidenz, um vorgelegte Argumente bzw. Erklärungen zu entkräften. Jedes Element kann in dem vorgeschlagenen Schema auf einer Skala von null bis zwei bewertet werden. Auf diese Weise kann ebenfalls eine Steigerung der Argumentqualität ermittelt werden (bspw. Christodoulou & Osborne, 2014). Einen weiteren strukturellen

Rahmen für das Argumentieren stellt das Toulmin-Schema dar, welches innerhalb eines Arguments die sechs Elemente „datum, warrant, backing, qualifier, conclusion“ und „rebuttal“ identifiziert (Toulmin, 1958). Das Toulmin-Schema findet in vielen didaktischen Arbeiten zum Argumentieren Anwendung und wird ebenfalls verwendet, um komplexere Rahmungen für die Evaluation von Argumenten zu konstruieren (bspw. Bell & Linn, 2000; Berland & McNeill, 2010; Weinberger & Fischer, 2006). Es stellt die Grundlage für das strukturelle Schema von McNeill und Krajcik (2011) dar. Es wird jedoch kritisiert, dass die reine Betrachtung der strukturellen Ebene nur bedingt mit der Argumentqualität zusammenhängt (Kuhn & Lerman, 2021).

Inhaltliche Analyse von Argumenten

Eine inhaltliche Analyse von Argumenten erfolgt hinsichtlich ihrer fachlichen Richtigkeit im Rahmen des betrachteten Kontextes. Aus dieser Perspektive heraus kann der Fortschritt des Argumentierens mit dem aktuellen Lernfortschritt verknüpft werden (Berland & McNeill, 2010). Vor dem Hintergrund der Einbettung von Argumentationsanlässen in die Problemlösung stellt dieses Charakteristikum einen wichtigen Indikator für den Erfolg und das Argumentieren dar (Weinberger & Fischer, 2006).

Analyse der Art der Begründung

Abschließend bleibt nach dem Schema von Sampson und Clark (2008) die Analyse der Begründung eines Arguments. Im Vordergrund steht hierbei, wie die Behauptungen durch die Schüler:innen gerechtfertigt werden (Sampson & Clark, 2008, S. 451). Diese Perspektive weist insbesondere in den Naturwissenschaften eine große Bedeutung auf:

> Argumentation, with its emphasis on justification of claims and on the coordination among claims and evidence, may support the development of epistemic criteria and more generally the enculturation in the practices of the scientific community. (Erduran & Jiménez-Aleixandre, 2008, S. 10)

Die Analyse der Art der Begründung ermöglicht somit nicht nur auf struktureller und inhaltlicher Ebene, sondern auch auf epistemischer und diskursiver Ebene Argumente zu bewerten und qualitativ einzuordnen. Die Art der Begründung kann daher zur Analyse der Argumentqualität verwendet werden (bspw. Ludwig et al., 2019; McNeill & Krajcik, 2011; Weinberger & Fischer, 2006). An dieser Stelle soll der Evaluationsrahmen von Ludwig et al. (2019) hervorgehoben werden, in dem der Einfluss von personalen und situationalen Faktoren auf das Argumentieren beim Experimentieren untersucht wurde. Dieses Instrument erlaubt es, die Art der Begründung von Schüler:innen in die Argumentkategorien abzubilden. Der Evaluationsrahmen von Ludwig (2017) umfasst qualitativ die zehn Argumentkategorien *Intuition, Ignoranz, Experimentierkompetenz (zen-*

tral), Experimentierkompetenz (peripher), Eignung des Experiments, heranziehen einer (falschen) physikalischen Theorie, Expertenwissen, Messunsicherheiten (implizit), Messunsicherheiten (explizit) und *Daten als Evidennz.* Vor dem Hintergrund des Experimentierens und damit einhergehend der Bedeutsamkeit von Messdaten und Evidenzen sowie der hohen didaktischen Relevanz aus wissenschaftlicher und schulischer Perspektive wurden vier Kategorien ausgewählt, welche zusätzlich über Likert-skalierte Items abgebildet werden können (Ludwig, 2017, S. 41 f.). Dabei wurde ebenfalls berücksichtigt, inwiefern eine Quantizifierung der Kategorie sinnvoll ist (bspw. Argumentkategorie Ignoranz). Aus diesen Überlegungen heraus hat Ludwig (2017) die Argumentkategorien Intuition, Expertenwissen, Messunsicherheiten und Daten als Evidenz ausgewählt, die im weiteren Verlauf der Arbeit verwendet werden, um die Art der Begründung der Proband:innen quantitativ zu erfassen. Dabei handelt es sich um eine Selbstauskunft der Schüler:innen, vorgegebene Aussagen zu den Argumentkategorien wie bspw. „Ich habe die Messungenauigkeiten beim Experimentieren in hohem Maße berücksichtigt“ auf einer fünfstufigen Likert-Skala von „trifft gar nicht“ bis „trifft völlig zu“ zu bewerten. Zu jeder der vier Argumentkategorien müssen die Schüler:innen fünf Aussagen bewerten. Für eine ausführlichere Beschreibung des Instruments wird auf Abschnitt 6.1 und die Arbeit von Ludwig (2017) verwiesen.

Vor dem Hintergrund der administrativ bedingten normativen Gründe für das Argumentieren und der in den vorangehenden Abschnitten dargelegten Bedeutung der Argumentation als Instrument zur Förderung und Etablierung von *Scientific Practices* und *Scientific Literacy* im Unterricht wird in der vorliegenden Arbeit die Analyse der Art der Rechtfertigung zur Evaluation der Schüler:innenargumente verwendet, da so die epistemische Sicht auf das Argumentieren nachvollzogen werden kann:

„Analyzing the nature of justifications allows us to focus on the „epistemic dimension“ in argument (Weinberger & Fischer, 2006, S. 74) and, consequently, learn about how students construct knowledge in argumentation and why they support or reject a certain claim (Ludwig et al., 2019, S. 821).“

Aufgrund des quantitativen Charakters, der Erfassung von rationalen sowie nicht-rationalen Rechtfertigungen und des experimentellen Kontexts wird dabei das Instrument von Ludwig et al. (2019) verwendet. Die Argumentkategorien Messunsicherheiten und Daten als Evidenz erhalten eine hohe Zustimmung, wenn die Entscheidung der Lernenden evidenzbasiert auf Grundlage der Daten und Unsicherheiten getroffen wird. Die Argumentkategorien Intuition und Expertenwissen erhalten eine hohe Zustimmung, wenn die Begründung gefühlsbasiert, auf Grundlage von Heuristiken oder den Verweis auf externe Autoritäten erfolgt. Eine ausführliche Darstellung und Unterscheidung der Argumentkategorien erfolgt in Abschnitt 6.1. Durch die Verwendung des Instruments von Ludwig et al. (2019) ist es daher im Rahmen der vorliegenden Arbeit möglich, ebenfalls den Einfluss von nicht naturwissenschaftlich-adäquaten Rechtfertigungen auf

das Argumentieren und den Lernerfolg zu untersuchen.

Für eine detaillierte Übersicht über die dargestellten Ebenen zu Möglichkeiten der Argumentanalyse sei auf die Arbeit von Sampson und Clark (2008) verwiesen.

2.1.5 Intrapersonelle Ebene des Argumentierens

In der bisherigen Darstellung des Argumentierens kann der Prozess sowohl als individuelle als auch kollektive Tätigkeit verstanden werden. Diesbezüglich äußern sich Driver et al. (2000) wie folgt:

> [...] arguing is a human practice that is situated in specific social settings. From this situated perspective, argument can be seen to take place as an individual activity, through thinking and writing, or as a social activity taking place within a group a negotiated social act within a specific community. (S. 290 f.)

Die vorliegende Sichtweise verdeutlicht, dass Argumentieren auf der inter- und der intraindividuellen Ebene betrachtet werden kann. Beide Aspekte des Argumentierens sind Gegenstand der empirischen Bildungsforschung (Ford, 2012). Die zuvor dargelegten Forschungsarbeiten zielen dabei auf beide Aspekte ab. Arbeiten, welche Peer- oder Gruppendiskussionen als Forschungsgegenstand aufweisen, untersuchen den Schwerpunkt der interindividuellen Argumentation (bspw. Asterhan, 2018; Bell & Linn, 2000; Choi et al., 2015; Christodoulou & Osborne, 2014; Clark et al., 2007; Duschl & Osborne, 2002; Iordanou & Constantinou, 2015). In diesem Fall hat das Argumentieren das Ziel, die Person gegenüber vom eigenen Standpunkt zu überzeugen (Osborne et al., 2004; Sandoval & Millwood, 2005). Hingegen betrachten Arbeiten, die Argumente analysieren, welche in einer selbstständigen Tätigkeit von Schüler:innen, zum Beispiel individuelles Experimentieren oder schriftliches Notieren von Argumenten, entstanden sind, die intraindividuelle Ebene des Argumentierens (bspw. Aydeniz & Ozdilek, 2015; Bilican & Aydeniz, 2016; Iordanou & Constantinou, 2015; Ludwig, 2017). Nach Rocci (2005) lassen sich die beiden Aspekte des Argumentierens durch folgenden Gedanken miteinander verbinden:

> Argumentative processes are shown to operate both at the interpersonal level, and at the intra-personal level in soliloquy. In fact, all persuasive processes contain, in order to succeed, an element of soliloquial argumentation: to persuade through argumentation means to induce somebody else to let him/herself be convinced by the argument. (S. 97)

Somit weisen die Ebenen des Argumentierens neben ähnlichen Zielen eine weitere starke Verbindung zueinander auf. Im Rahmen der vorliegenden Arbeit werden die Schüler:innen selbstständig Hypothesen aufstellen, experimentelle Daten auswerten und ihre Hypothesen auf Gültigkeit evaluieren. Dies ist ein individueller Prozess, der nach der dargestellten Sichtweise als intraindividuelle Argumentation aufgefasst werden kann. Zudem fertigen die Schüler:innen schriftliche und offene Begründungen an, in denen sie ihre Behauptungen überzeugend darlegen wollen. Es ist daher legitim, davon auszugehen, dass die vorliegende Arbeit den Einfluss der in Abschnitten 5.1.2 und 6.1 dargelegten Einflussfaktoren auf das Argumentieren untersucht.

2.2 Argumentieren beim Experimentieren

2.2.1 Warum ist das Experiment als Argumentationsanlass geeignet?

In Abschnitt 2.1.2 wurde aufgezeigt, dass das Argumentieren auf vielfältige Weise in den Unterricht eingebunden werden kann. Die dargelegten Forschungsarbeiten und -ergebnisse haben dabei oftmals verschiedene Kontexte wie beispielsweise *Socioscientific Issues* oder physikalische Experimente verwendet. In der vorliegenden Arbeit wird das Argumentieren im Kontext eines Experiments zum Fadenpendel untersucht. Der nachfolgende Abschnitt soll die Bedeutung des Argumentierens im Rahmen des Experimentierens aufzeigen sowie die Wahl des Kontextes begründen.

Einerseits kommt im Rahmen der bildungspolitischen Vorgaben für den naturwissenschaftlichen Unterricht im Bereich der Erkenntnisgewinnung dem Experimentieren eine große Rolle zu (vgl. KMK, 2004; Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016; NRC, 2013). So wird beispielsweise von den Schüler:innen der Sekundarstufe I für den mittleren Schulabschluss erwartet, einfache Experimente durchführen und die Experimente bzw. gegebenenfalls die Daten auswerten zu können (KMK, 2004, S. 11). In der Sekundarstufe II wird das Experimentieren sowohl bei der Sachkompetenz als auch bei der Erkenntnisgewinnungskompetenz genannt. Im Bereich der Sachkompetenz werden neben der Durchführung des Experiments (S4) ebenfalls Auswerteverfahren sowie die Anwendung dieser auf Messergebnisse (S6) genannt. Im Bereich der Erkenntnisgewinnungskompetenz wird das Experiment beispielsweise im Standard E5, „Die Lernenden planen geeignete Experimente und Auswertungen zur Untersuchung einer physikalischen Fragestellung“, genannt (KMK, 2020,

S. 14 f.). Andererseits weisen die naturwissenschaftlich-didaktischen Modelle des Experimentierens zudem noch Aspekte wie die Aufbereitung von Daten, die Hypothesenbildung sowie die Diskussion der Ergebnisse auf (Hammann, 2004; Nawrath et al., 2011). Im Rahmen des Experimentierens können Lernende auf Grundlage naturwissenschaftlicher Evidenz aus selbstständig erhobenen Messdaten begründen. Der Prozess des Experimentierens kann daher ebenfalls aus der Sicht der in Abschnitten 2.1.1 und 2.1.2 dargestellten Merkmale als Argumentieren angesehen werden (Gott & Duggan, 2007; Kanari & Millar, 2004; Lunetta et al., 2007; Osborne & Patterson, 2011). Zusätzlich zum Argumentieren können durch das Experiment weitere Ziele des naturwissenschaftlichen Unterrichts adressiert werden (Bennett, 2003; Hofstein, 2017). Hofstein (2017) sagt diesbezüglich:

> Inquirytype laboratories are central to learning science, since students are involved in the process of conceiving problems and scientific questions, formulating hypotheses, designing experiments, gathering and analyzing data, and drawing conclusions about scientific problems or phenomena. (S. 359)

Die Bedeutung des Experimentierens wird von weiteren Autor:innen bestätigt: Dem praktischen Arbeiten kommt eine zentrale Rolle beim Unterrichten eines Verständnisses von wissenschaftlicher Evidenzen zu (Gott & Duggan, 1996). Durch das Experimentieren kann zudem eine einzigartige Konstellation von Lernen und Bewerten transportiert werden, die eine zentrale und unverkennbare Rolle in der naturwissenschaftlichen Bildung aufweisen (Hofstein, 2017; Hofstein & Lunetta, 2004; Zhang, 2016). Zudem kann das *Scientific Reasoning* beim Experimentieren adressiert und gefördert werden (Kind, 2013; Kind & Osborne, 2017). Vor diesem Hintergrund kann dem Experimentieren sowie dem Argumentieren ein besonderer Stellenwert bei der Förderung von epistemischen und diskursiven Arbeitsweisen sowie von *Scientific Practice* und *Scientific Reasoning* konstatiert werden.

Dabei zeigen jedoch verschiedene Forschungsarbeiten auf, dass die Rolle des praktischen Arbeitens bzw. Experimentierens im Unterricht undefiniert und unbestimmt ist (Gott & Duggan, 1996; Xu & Talanquer, 2013). Historisch bedingt werden Experimente oft mit einem konfirmatorischen Aspekt verwendet und können im lehrerzentrierten Unterricht nur Informationen transportieren (Hofstein, 2017; Zhang, 2016). Außerdem wird kritisiert, dass das Experimentieren nicht nur „hands on“ beinhalten sollte, sondern auch „minds on“ (Abrahams & Reiss, 2012). Somit wird der von Hofstein (2017) adressierte inquirybased wissenschaftliche Diskurs nur periphär tangiert. Diesbezüglich präzisiert Millar (2004):

> Practical work of a more open-ended, investigative kind can develop students' tacit knowledge of scientific enquiry. It is, however, hard to

> argue that this kind of knowledge is needed for 'scientific literacy', though it is clearly of value to students who wish to follow more advanced courses in science. (S. 20)

Somit soll das explorative Vorgehen beim Experimentieren nicht entwertet werden. Im Rahmen der Prüfung von Hypothesen und Modellen kommt dem inquiry-based Ansatz beim Experimentieren eine größere Rolle zu. Daher wird im Rahmen dieser Arbeit ein solches Vorgehen näher betrachtet.

Der unbestimmten Rolle des Experimentierens im Unterricht stehen Aussagen von Hofstein und Lunetta (1982, 2004) gegenüber, die mit zwei Untersuchungen im Abstand von 22 Jahren festgestellt haben, dass die Herausforderung, das Experimentieren in den Unterricht einzubinden, weiterhin existiert, jedoch aufgrund der neuen wissenschaftlichen Erkenntnisse und rapiden technologischen Entwicklung Lehrer:innen weiterhin befähigt werden müssen, entsprechende Aspekte in ihren Unterricht einzubinden (Hofstein & Lunetta, 1982, 2004). Ferner haben Gott und Duggan (1995) aufgezeigt, wie praktisches Arbeiten grundsätzlich in den Unterricht eingebunden werden kann, da es eine Schlüsselrolle beim Unterrichten von Evidenzen aufweist (Gott & Duggan, 1996; Gott & Duggan, 1995).

Ludwig (2017) hat in seiner Forschungsarbeit das Argumentieren beim Experimentieren untersucht. Im Kontext des Fadenpendels haben die Schüler:innen experimentell den Zusammenhang zwischen der Periodendauer und der Pendelmasse eines Fadenpendels untersucht, wobei der Zusammenhang aus der Theorie durch

$$T = 2\pi\sqrt{\frac{l}{g}}$$

mit T der Periodendauer, l der Pendellänge und g der Erdbeschleunigung definiert ist. Durch den Einsatz seines Instruments konnte er die Art der Begründung und deren Auswirkung auf den Lernerfolg in Form der fachlich korrekten Hypothese erfassen. Es konnte unter anderem gezeigt werden, dass Lernende mit einem höheren Fachwissen ihre Hypothesen eher evidenzbasiert und weniger intuitiv begründen. Weiterhin konnte gezeigt werden, dass die Verwendung von evidenzbasierten Begründungskategorien einen positiven Effekt und die intuitiven Begründungskategorien sowie der Verweis auf externe Kompetenz einen negativen Effekt auf die Richtigkeit der Hypothese aufweisen. Somit kann über das Fachwissen sowohl die Art der Begründung als auch der Lernerfolg der Schüler:innen adressiert werden. Vor diesem Hintergrund ist das Ziel der geplanten Interventionen, das für das Experimentieren benötigte Fachwissen zu adressieren. Der Schwerpunkt liegt in diesem Rahmen auf dem Auswerten der experimentell ermittelten Daten, da diese als wichtiger Teil des Experimentierens und Argumentierens aufgefasst werden können (Gott & Duggan, 2007; Kind, 2013; Klahr & Dunbar, 1988; Millar, 2004; Nawrath et al., 2011). Die Inhalte der Intervention und die Darlegung der für die adäquate Auswertung

von experimentellen Daten benötigten Kompetenzen erfolgt in den Abschnitten 2.3 und 4.1.1.

2.2.2 Nicht-hypothesenkonforme Daten als Ausgangspunkt

Argumentationsanlässe können sich im Unterricht zufällig ergeben oder durch die Lehrkraft bewusst herbeigerufen werden. Grundsätzlich kann eine Argumentation somit auf verschiedene Weisen initiiert werden. Allgemein charakterisiert Kopperschmidt (1989) zwei Voraussetzungen für das Stattfinden von Argumentationen: Als notwendige Bedingung bedarf es einer strittigen Situation. Zudem wird als hinreichende Bedingung formuliert, dass die teilnehmenden Personen die Bereitschaft und die Fähigkeit zum Argumentieren aufweisen (Kopperschmidt, 1989).

Die in Abschnitt 2.1.2 dargestellten Forschungsarbeiten adressieren auf verschiedenen Wegen das Argumentieren im naturwissenschaftlichen Unterricht. Die zuvor dargestellte besondere Eignung des Experiments wird zwar in einigen Forschungsarbeiten berücksichtigt (bspw. Gott & Duggan, 1996; Kanari & Millar, 2004; Kind et al., 2011; Ludwig, 2017; Lunetta et al., 2007), jedoch findet das Experiment trotz der herausgearbeiteten Eignung zur Adressierung des Argumentierens wenig Anklang (Ludwig, 2017). Im Rahmen der vorliegenden Arbeit wird die strittige Situation im Kontext des Experimentierens herbeigeführt. So formulieren Hammann et al. (2008, S. 34), dass „die Analyse der Daten vor dem Hintergrund der gebildeten Hypothesen“ als eine zentrale Tätigkeit beim Experimentieren angesehen werden kann. In diesem Fall kann daraufhin eine strittige Situation initiiert werden, wenn „nicht die richtigen Kriterien oder Methoden angewendet werden, es Widersprüche gibt oder kognitive Konflikte entstehen, wenn also der Wahrheitsanspruch von Aussagen in Zweifel gezogen werden kann“ (Gromadecki, 2008, S. 92 f.).

In der vorliegenden Arbeit wird beabsichtigt, durch einen kognitiven Konflikt die intrapersonale Argumentation der Lernenden zu initiieren. Der kognitive Konflikt soll dadurch induziert werden, dass die Schüler:innen ein Experiment bearbeiten, zu dem es starke fachlich defizitäre Präkonzepte seitens der Schüler:innen gibt (vgl. Abschnitt 6.1). In diesem Fall sind die experimentellen Daten bei korrekter Auswertung nicht mit der Hypothese der Schüler:innen konform. Dabei ist die Konfrontation der Lernenden mit den „nicht-hypothesenkonformen“ Daten das zentrale Element, um den kognitiven Konflikt und damit die intraindividuelle Argumentation herbeizuführen (Ludwig, 2017, S. 12).

Im Rahmen des Experimentierens wird dieser Argumentationsanlass in verschiedenen Forschungsarbeiten verwendet (Berland & Lee, 2010; Ludwig, 2017; Ludwig et al., 2019). Dabei kann der kognitive Konflikt nicht nur im experimentellen Kontext als Argumentationsanlass verwendet werden. Chinn

und Brewer (1993) zeigen auf, wie Lernende mit „anomalous data“[1] umgehen und welche Implikationen dieser Umgang mit dem resultierenden kognitiven Konflikt auf die gesellschaftliche Rolle des Individuums aufweist. Insbesondere vor dem Hintergrund der steigenden Informationsverfügbarkeit und „post-truth conditions“ weist dieser Aspekt des Umgangs mit nicht-hypothesenkonformen Daten eine wachsende Bedeutung auf (Barzilai & Chinn, 2020). Dabei weist Ludwig (2017, S. 12) jedoch darauf hin, dass die entsprechenden Klassifizierungen vom Umgang mit nicht-hypothesenkonformen Daten mit Vorsicht zu betrachten sind, da entsprechende Arbeiten (bspw. Chinn & Brewer, 1993, 2001) nicht das schulische Experimentieren betrachten und den Fokus mehr auf den Konzeptwechsel als auf das Argumentieren legen. Ein durch nicht-hypothesenkonforme Daten induzierter kognitiver Konflikt bei den Schüler:innen kann dazu führen, dass weitere kognitiv-psychologische Phänomene berücksichtigt werden müssen. In diesem Zusammenhang können ein *Confirmation Bias* und die Persistenz von Überzeugungen den Konzeptwechsel der Schüler:innen beeinflussen (Chinn & Brewer, 1993; Ludwig, 2017). Im Rahmen der vorliegenden Arbeit und des den Schüler:innen vorgelegten experimentellen Kontext können durch nicht-hypothesenkonforme Daten *Confirmation Bias* oder Persistenz von Überzeugungen auf Ebene der Schüler:innen eine praktische Relevanz aufweisen. Nach Nickerson (1998) bezeichnet die Begrifflichkeit *Confirmation Bias* „[. . .] the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand“. Aufgrund des in der vorliegenden Arbeiten gewählten Kontextes und der damit einhergehenden fachlich defizitären Präkonzepte ist davon auszugehen, dass Schüler:innen bei der Auswertung von einem *Confirmation Bias* geleitet werden können. Ferner konnte festgestellt werden, dass der *Confirmation Bias* nicht nur als einfacher Fehler angesehen werden kann, sondern sich als Teil des Argumentationsmusters der Schüler:innen manifestieren kann (Jones & Sugden, 2001). Darüber hinaus zeigt sich bei verankerten Konzepten von Schüler:innen eine hohe Persistenz, die dem Konzeptwechsel gegenübersteht (Jones & Sugden, 2001; Kang et al., 2010). Neben diesem dargelegten Einfluss des *Confirmation Bias* konnte in mehreren Studien außerdem gezeigt werden, dass die Schüler:innen unerwartete oder unklare Daten für ihre Hypothesen und Argumentationen nicht berücksichtigen (Chinn & Brewer, 1993; Kanari & Millar, 2004; Ludwig et al., 2019). Für eine detaillierte Darstellung der genannten kognitionspsychologischen Einflussfaktoren auf den kognitiven Konflikt wird auf Ludwig (2017, S. 12 f.) verwiesen.

Die vorangehende Darstellung soll verdeutlichen, welche Wirkung ein *Confirmation Bias* und die Persistenz von Präkonzepten auf die durch einen kognitiven Konflikt induzierte intrapersonelle Argumentation haben können. Unter Berücksichtigung dieser Einflussfaktoren soll nachfolgend die empirische Befundlage zum Argumentieren beim Experimentieren untersucht werden.

[1] Übersetzung nach Ludwig (2017) mit *nicht-hypothesenkonformen Daten*

2.2.3 Modellierung von Argumentationsprozessen beim Experimentieren durch das Elaboration-Likelihood Model of Persuasion

Darüber hinaus konnte Ludwig (2017) zeigen, dass für die Betrachtung der Persuasion von Lernenden beim Argumentieren beim Experimentieren das Elaboration-Likelihood Model of Persuasion (ELM) herangezogen werden kann (Ludwig, 2017, S. 30 ff.).

Das ELM ist ein Modell aus der Kommunikationspsychologie, welches eine Perspektive für die Verarbeitung von persuasiven Mitteilungen bietet (bspw. Greenwald, 1968; Petty & Cacioppo, 1986). Im Modell wird zwischen der zentralen und peripheren Route unterschieden, über die persuasive Informationen verarbeitet werden können. Eine Verarbeitung von persuasiven Informationen über die zentrale Route kann mit einer Prüfung der Informationen nach rationalen Kriterien einhergehen. In diesem Fall setzt sich die Person intensiv mit den Inhalten auseinander und denkt sorgfältig darüber nach. Eine Verarbeitung von Informationen über diese Route kann zu nachhaltigen Einstellungsänderungen führen (P. K. Murphy & Mason, 2006). Eine Bewertung von persuasiven Informationen über die periphäre Route erfolgt eher weniger rational-logisch, sondern kann eher von Einflussfaktoren wie intuitiven Heuristiken, Hinweisreizen, Glaubwürdigkeit der Quelle usw. beeinflusst werden. Eine Einstellungsveränderung, die über diese Route hervorgerufen wurde, ist im Vergleich zur durch die zentrale Route hervorgerufenen Einstellungsveränderung weniger zeitlich stabil. Aus der Perspektive des naturwissenschaftlichen Unterrichts ist die Verarbeitung der persuasiven Informationen über die zentrale Route daher zu bevorzugen.

Im Rahmen der vorliegenden Arbeit kann das Experiment zum Fadenpendel mit nicht-hypothesenkonformen Daten als persuasive Informationen, die die Schüler:innen verarbeiten sollen, und das Aufstellen der Hypothese als Einstellungsänderung aufgefasst werden (Ludwig, 2017, S. 45). Rückschlüsse auf eine Verarbeitung der Informationen über die zentrale und periphäre Route können anhand der verwendeten Argumentkategorien gezogen werden: Eine höhere Zustimmung zu den Argumentkategorien Messunsicherheiten und Daten als Evidenz deutet auf eine eher rationale Auseinandersetzung mit den Daten hin und kann somit einer Verarbeitung der persuasiven Informationen über die zentrale Route zugeordnet werden. Eine höhere Zustimmung zu den Argumentkategorien Expertenwissen und Intuition spricht für eine eher nicht rationale Verarbeitung der experimentellen Daten. Die Schüler:innen berücksichtigen bei ihrer Entscheidung eher eine Form von Intuition, Heuristiken oder Eigenschaften der Daten bzw. des Experiments. Die Verwendung dieser Argumentkategorien kann somit eher der periphären Route des ELM zugeordnet werden. Im Sinne des ELM können die Argumentkategorien nach Ludwig (2017) bezüglich ihres kognitiven Charakters in eher rationale und eher nicht rationale Kategorien

getrennt werden.

2.2.4 Empirische Befundlage zum Argumentieren beim Experimentieren

Trotz der dargelegten fachdidaktischen Perspektive, die die Eignung des Experiments als Argumentationsanlass motiviert, zeigt sich eine Diskrepanz in der fachdidaktischen Forschung zwischen Theorie und Praxis: Es existieren nur wenige Arbeiten, die explizit das Argumentieren beim Experimentieren untersuchen (Hofstein, 2017; Ludwig, 2017).

In der Untersuchung von Ludwig (2017) haben Schüler:innen experimentell den Zusammenhang zwischen der Periodenlänge und der Pendelmasse eines Fadenpendels untersucht. Es wurde untersucht, wie die Schüler:innen auf Grundlage selbstständig erhobener Messdaten fachlich physikalische Hypothesen aufstellen und ihre Entscheidung begründen. Für die Erfassung der Argumentation der Schüler:innen wurde das in Abschnitt 2.1.4 vorgestellte Instrument von Ludwig (2017) verwendet. Unter anderem konnte dabei gezeigt werden, dass Schüler:innen, die über ein höheres Fachwissen in der Mechanik verfügen, eher weniger auf Heuristiken oder Gefühlen basierend, sondern eher mehr datenbasiert argumentieren. Darüber hinaus konnte gezeigt werden, dass der Lernerfolg der Schüler:innen höher ausfällt, wenn mehr datenbasiert und weniger intuitiv argumentiert wird. Im gleichen experimentellen Kontext konnten Kanari und Millar (2004) zeigen, dass die Schüler:innen im Fall von fehlender Kovariation zwischen unabhängiger und abhängiger Variable Probleme mit datenbasierten Begründungen aufweisen.

In diesem Zusammenhang zeigt die Arbeit von Pols (2019) spannende Ergebnisse. Der Schwerpunkt der Untersuchung lag in der selbstständig Auswertung von experimentellen Daten. Die Schüler:innen führten Experimente zur Pendelschwingung von Spiderman und der gleichförmigen Bewegung einer Murmel auf der Ebene durch. Die Autoren stellen fest, dass trotz der in den Bildungsplänen vorgegebenen Kompetenzen zur Auswertung von empirischen Daten Schüler:innen Schwierigkeiten mit der Auswertung ebensolcher Daten aufweisen, und resümieren: „Based upon these outcomes, it can not be justified that students carry out experiments in which they analyse empirical data and draw conclusions out of these without the teacher's help“(Pols, 2019, S. 7). Diese Ergebnisse stehen im Zeichen von Arbeiten, die unterstützende Maßnahmen beim Experimentieren vorschlagen (bspw. Kim & Song, 2006; Rickey & Stacy, 2000).

Das Auswerten von experimentellen Daten steht ebenfalls im Fokus der Arbeit von Kok (2022). In einer Interventionsstudie wurde untersucht, wie sich die Förderung von Kompetenzen zu einem adäquaten Umgang mit Daten und Unsicherheiten auf den Vergleich von experimentell ermittelten Messdaten

auswirkt. Es konnte unter anderem gezeigt werden, dass die Schüler:innen nach die Daten nach der Intervention öfter korrekt ausgewertet haben.

Die dargestellten Studien untersuchen verschiedene Situationen: Während ein Teil der dargelegten Arbeiten (Kanari & Millar, 2004; Kind et al., 2011; Ludwig, 2017; Pols, 2019) das Argumentieren beim gesamten Experimentierprozess analysiert, fokussiert sich der andere Teil auf die Auswertung von experimentell ermittelten Messdaten (Kind et al., 2011; Kok, 2022). Die Übertragbarkeit der Ergebnisse der beiden Studien ist daher mit Vorsicht zu betrachten (Ludwig, 2017). Nichtsdestotrotz lässt sich erkennen, dass fehlende Kompetenzen im Umgang mit Daten und Unsicherheiten die Argumentation von Schüler:innen hemmen bzw. verhindern (Kanari & Millar, 2004; Kind et al., 2011). Des Weiteren konnten Ludwig et al. (2019) zeigen, dass Schüler:innen beim Auswerten von experimentellen Messdaten eine Sensibilität für Messunsicherheiten aufweisen. Da jedoch die Kompetenzen für einen adäquaten Umgang mit Messunsicherheiten bei den Lernenden nicht vorhanden waren, behielten die Lernenden ihre falschen Hypothesen bei. Darüber hinaus konnte gezeigt werden, dass fehlendes Wissen um den Messprozess und Messunsicherheiten dazu führen kann, dass Schüler:innen Zusammenhänge von nicht miteinander kovariierenden physikalischen Größen wie beispielsweise der Periodendauer und Pendelmasse beim Fadenpendel nicht auf Grundlage von Messdaten erkennen und folgern können (Kanari & Millar, 2004).

Es lässt sich somit konstatieren, dass ein adäquater Umgang mit Daten und Unsicherheiten beim Argumentieren beim Experimentieren von hoher Relevanz ist. Der anhand exemplarischer Arbeiten aufgezeigte Sachverhalt kann durch die Aussagen von Kind et al. (2011) zusammengefasst werden:

> Where students had to carry out an experiment, argumentation was often short, as reliance on their data was paramount. Measurements were given credence by frequency and regularity of collection, while possibilities for error were ignored. These data point to changes to existing practices being required in order to achieve authentic scientific inquiry in school laboratory work. (S. 2)

Zwischenfazit

Die dargelegte empirische Befundlage zum Argumentieren beim Experimentieren lässt unter anderem zwei Beobachtungen zu: Einerseits zeigt sich, dass Fachwissen daten- und evidenzbasiertes Argumentieren fördern kann. Während in der Arbeit von Ludwig (2017) das Fachwissen im Bereich der Mechanik ermittelt wurde, liegt auf Grundlage der präsentierten Ergebnisse die Vermutung nahe, dass Fachwissen im Bereich des Experimentierens eine höhere Relevanz für das Argumentieren beim Experimentieren aufweisen kann. Andererseits kann ein fehlendes Verständnis eines adäquaten Umgangs mit Daten und Unsicherheiten den Argumentationsprozess beim Experimentieren hemmen. Vor diesem Hintergrund

kann der adäquate Umgang mit Daten und Unsicherheiten beim Experimentieren als ein wesentliches Element zur Förderung des Argumentierens beim Experimentieren identifiziert werden. Daher wird in Abschnitt 2.3 ein Überblick über diesen Themenbereich gegeben, wobei sowohl fachwissenschaftliche Elemente als auch Ergebnisse fachdidaktischer Forschung herausgearbeitet werden.

2.3 Umgang mit Daten und Unsicherheiten im Physikunterricht

Der adäquate Umgang mit Daten und Unsicherheiten im naturwissenschaftlichen Unterricht weist nicht nur auf Grundlage der zuvor aufgezeigten Aspekte eine große Bedeutung auf, sondern ist mittlerweile auch aus normativen Gründen heraus in den Unterricht zu integrieren (vgl. KMK, 2004; NRC, 2013). Nachfolgend wird zunächst die fachwissenschaftliche Perspektive auf das Auswerten der Daten unter Berücksichtigung der Messunsicherheiten dargelegt. Im Rahmen der Arbeit wird wie in Abschnitt 2.2.1 das Experiment als Argumentationsanlass kontextualisiert. Aus diesem Grund wird sich der nachfolgende Abschnitt verstärkt mit dem Umgang mit experimentellen Daten beschäftigen. Der dargestellte Bereich entspricht nur einem kleinen Bereich der allgemeinen Datenkompetenz oder dem Verständnis von *Data Literacy* und erhebt daher keinen Anspruch auf Vollständigkeit. Der Umgang mit experimentellen Daten, Messunsicherheiten und Messgeräten stellt einen Teil der Metrologie dar, welche nachfolgend als Ankerpunkt der Arbeit dienen soll.

2.3.1 Grundlagen der Metrologie

Die Metrologie kann als „science of measurement and its application“ (JCGM, 2012, S. 16) definiert werden und stellt somit die fachwissenschaftliche Grundlage für die adäquate Erfassung, Darstellung und Auswertung von experimentellen Daten dar. Sprachlich gestaltet sich dieser Bereich im Deutschen nicht eindeutig: Historisch sind Begriffe wie Messfehler oder auch „Fehlerrechnung“ gewachsen, die im heutigen Sinne anders interpretiert und trotzdem weiterhin in entsprechenden Forschungsarbeiten verwendet werden (Heinicke et al., 2010; Klinger & Wörlen, 1991; Nagel et al., 2021). Nicht immer sind Forscher:innen bereit, sich an die entsprechenden engen sprachlichen Vorgaben der Metrologie zu halten (Nagel, 2021). Zudem gibt es Diskrepanzen zwischen den Begrifflichkeiten im

englisch- und deutschsprachigen Raum. So kann zwar die Messunsicherheit als *Measurement Uncertainty* bezeichnet werden, jedoch wird Messabweichung als *Measurement Error* bezeichnet, welches im Konflikt mit der Übersetzung von error (übersetzt mit „Fehler“) steht (JCGM, 2008). Ebenso werden Konzepte falsch aufgefasst. Beispielsweise werden Unsicherheiten verallgemeinert, ohne auf die Kategorisierung Typ A oder Typ B Unsicherheit und die damit einhergehende Aussagekraft der Unsicherheiten bezüglich der Präzision und Richtigkeit einzugehen (JCGM, 2008; Nagel, 2021). Aus diesem Grund wird für die nachfolgende Arbeit zunächst eine Definition der Begriffe durchgeführt. Während der Guide to the expression of uncertainty in measurement (GUM) (JCGM, 2008) den adäquaten Umgang mit Messunsicherheiten vollumfänglich adressiert, stellt das International vocabulary of metrology (VIM) eine Sammlung von Definitionen aller wesentlicher Kenngrößen dar. Der GUM kann als internationaler Vorschlag des *Bureau International des Poids et Mesure* verstanden werden, einheitliche Verfahren, Konzepte und Bezeichnungen im Bereich des Messens zu etablieren. Die Inhalte des GUM wurden vollständig in die DIN-Norm 1319 aufgenommen. Aufgeteilt in vier Teile stellen sie somit auch in Deutschland die normative Referenz für den Umgang mit Messunsicherheiten dar (*DIN 1319-1*, 1995; *DIN 1319-2*, 2005; *DIN 1319-3*, 1996; *DIN 1319-4*, 1999). Für die Definition der Begriffe Messunsicherheit und Messabweichung wird auf die Definitionen innerhalb des VIM zurückgegriffen. Nachfolgend wird unter Messunsicherheit ein „non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used“ (JCGM, 2012, S. 25) verstanden. Messabweichung hingegen bezeichnet „measured quantity value minus a reference quantity value“(JCGM, 2012, S. 22). Dieses Verständnis der Begriffe Messunsicherheit und -abweichung wird in Abschnitt 2.3.5 aufgegriffen und vor dem Hintergrund der fachdidaktischen Forschung und der Präkonzepte der Schüler:innen kontextualisiert.

Obwohl die Vorgaben des GUM international als normative Referenz gelten (ISO/TMBG, 2008) und Teil des Standardkatalogs 17.020 sind (ISO, 2022), ist die Verwendung des GUM nicht unumstritten. Kritisiert wird die *simplifizierte* Behandlung von zufälligen systematischen Fehlern im Rahmen des GUM oder die fehlende genaue Klassifikation der logischen Zusammenhänge zwischen der Messunsicherheit und der Genauigkeit einer Messung (Grabe, 2001; Krouwer, 2003; Ye et al., 2016). Verschiedene Arbeiten, die versuchen, diesen genannten Zusammenhang zu systematisieren und neue *Elementarfehlermodelle* zu generieren (bspw. Weiji, 1980), werden jedoch aufgrund der nicht theoriebasierten Annahmen ebenfalls kritisiert (Schmidt, 2003).

Die dargestellten Kritiken entstammen aus dem Fachbereich der Geodäsie. Es handelt sich dabei um sehr fachspezifische Aspekte, welche im Rahmen anderer Fachbereiche nicht derart negativ formuliert worden sind. So finden sich viele Forschungsarbeiten sowohl in der Physikdidaktik (Allie et al., 2003;

Buffler et al., 2008; Hellwig, 2012; Schulz, 2022) als auch in der Fachliteratur zum Umgang mit Messdaten (Kirkup & Frenkel, 2006; Lira, 2016; Möhrke & Runge, 2020), die den GUM integrieren und dessen Anwendung in der Praxis demonstrieren. Nachfolgend wird daher trotz der ausgewiesenen Kritikpunkte auf den GUM als fachwissenschaftliche Referenz im Umgang mit experimentellen Daten und Messunsicherheiten zurückgegriffen.

2.3.2 Daten und Messunsicherheiten in bildungspolitischen Vorgaben

Während das GUM und VIM als fachwissenschaftliche Nachschlagewerke verwendet werden können, sind sie im Zuge der Kompetenzorientierung des Physikunterrichts und damit der zu unterrichtenden Inhalte schwer als strukturelle Basis für die Adressierung von Daten und Messunsicherheiten einzubinden. Nach Durchführung einer Sach- und Strukturanalyse im Bereich des adäquaten Umgangs mit Messunsicherheiten entwickelte Hellwig (2012) ein Sachstrukturmodell, welches Inhalte des GUM strukturiert und thematisch sortiert. Dabei werden Konzepte formuliert, die für den Umgang mit Daten nötig sind, und vier unterschiedlichen Kompetenzbereichen zugeordnet. Diese Systematisierung ermöglicht es, Kompetenzen bzw. Konzepte in diesem Bereich zu formulieren, und wird daher als ein Stützelement für die Entwicklung der Lernumgebungen (s. Abschnitt 4.1.1) verwendet und stellt neben dem GUM die fachwissenschaftliche Richtigkeit der Struktur dar. Dabei findet sich ebenfalls ein reduziertes Sachstrukturmodell in der Arbeit von Hellwig (2012), welches sich inhaltlich auf die für die Sekundarstufe I relevanten inhaltlichen Konzepte der Metrologie beschränkt.

Neben der fachwissenschaftlichen Korrektheit der adressierten Aspekte ist für die Entwicklung der Apps und der dazugehörigen Rahmenbedingungen ebenfalls von Bedeutung, dass die Inhalte relevant für Lernende in der Schule sind. Der Umgang mit Daten und Messunsicherheiten, insbesondere in Bezug auf die Analyse von empirischen Daten, ist international als Bildungsziel formuliert (bspw. NRC, 2013), mit dem Ziel eine *Scientific Literacy* auszubilden (OECD, 2015). Vor diesem Hintergrund hat Pols (2019) diverse Dokumente analysiert und auf Gemeinsamkeiten untersucht. Aus der hohen Überschneidung der unterschiedlichen Arbeiten formuliert er „10 data analysis skills studentes aged 15 should process“ (Pols, 2019, S. 3), welche im Folgenden mit den Inhalten ausgewählter Bildungspläne Deutschlands für das Fach Physik verglichen werden.

Dabei gibt es durchaus Unterschiede in den Plänen. Während im berliner Rahmenlehrplan (LISUM, 2015) auf den adäquaten Umgang mit Daten und Messunsicherheiten in den Kompetenzen und Standards an mehreren Stellen Bezug genommen wird, beispielsweise auf implizite Weise in dem Kompetenzbereich „Vergleichen und Ordnen“ oder „Hypothesenbildung“ (LISUM, 2015, S. 19 f.) oder explizit in Formulierungen der Art „Die Schülerinnen und Schüler

können grobe, zufällige und systematische Messfehler unterscheiden“ (LISUM, 2015, S. 21), jedoch im Bereich der Themen und inhaltlichen Felder nicht konkret auf die Fähigkeiten mit Messunsicherheiten eingegangen wird, zeigt sich unter anderem im Bildungsplan des Landes Baden-Württemberg eine thematische Aufarbeitung bezüglich Messunsicherheiten sowohl in den prozessbezogenen Kompetenzen, unter anderem im Bereich Erkenntnisgewinnung unter dem Aspekt „zielgerichtet experimentieren“ (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016, S. 10), als auch in den inhaltsbezogenen Kompetenzen, z. B. im Gebiet des Elektromagnetismus in der Kompetenzstufe 3 (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016, S. 27). In den Bildungsstandards der Sekundarstufe II für das Fach Physik werden Messunsicherheiten konkret adressiert. Der Standard E7 fordert beispielsweise, dass Messunsicherheiten bei der Erkenntnisgewinnung berücksichtigt werden und die Konsequenzen von Messunsicherheiten bezüglich der Ergebnisse analysiert werden müssen (KMK, 2020, S. 15). Trotz der strukturellen Unterschiede überschneiden sich die Bildungspläne in den adressierten Kompetenzen zum Umgang mit Daten und Messunsicherheiten und lassen sich mit den Formulierungen von Pols (2019) vereinen.

Die zuvor dargestellten Aspekte werden in Abschnitt 4.1.1 aufgegriffen und zu einer didaktischen Struktur vereint. Die Notwendigkeit dieser neuen Struktur ergibt sich anhand der in Abschnitt 2.1.2 formulierten Form der Intervention bzw. Apps und der in Abschnitt 3 geforderten Modularität als Entwicklungsziel. Nachfolgend werden die Inhalte der didaktischen Struktur aus der fachdidaktischen Forschung zum Umgang mit Daten und Messunsicherheiten herausgearbeitet. Es kann an dieser Stelle konstatiert werden, dass neben der Bedeutung der Daten und Messunsicherheiten im Kontext des Argumentierens (vgl. Abschnitt 2.2.2) die Einbindung dieser Kompetenzen in den Unterricht ebenso normativ motiviert werden kann und somit eine eminente Bedeutsamkeit für den naturwissenschaftlichen Unterricht und im Speziellen den Physikunterricht aufweist.

2.3.3 Was bedeutet adäquater Umgang mit experimentellen Daten und Messunsicherheiten?

Neben den in Abschnitt 2.3.2 dargestellten normativen Gründen spricht auch eine Reihe von anderen Faktoren für die Implementierung und Adressierung von Datenkompetenzen im Physikunterricht. Einerseits ist das physikalische Experiment ein wichtiger Bestandteil der wissenschaftlichen Erkenntnisgewinnung und nimmt daher ebenso einen großen Stellenwert im Physikunterricht ein (Solomon, 1980). Neben der experimentell-praktischen Kompetenz sind das Auswerten und Interpretieren der ermittelten Messwerte von großer Bedeutung (Maiseyenka et al., 2013; P. Murphy & Gott, 1984). Bei der Messung muss stets berück-

sichtigt werden, dass der „wahre“ Wert nicht ermittelt werden kann und ein Messergebnis die Zusammenfassung aller Informationen bezüglich der Messung darstellt (JCGM, 2008). Auf diese Weise gibt das Messergebnis Eigenschaften des Messvorgangs wieder und gibt an, wie groß das Vertrauen in die Messung ist (JCGM, 2008). Ein adäquater Umgang mit Daten und Unsicherheiten insbesondere vor dem Hintergrund des Experimentierens ist somit unabdingbar und für den Erkenntnisgewinn in der Physik notwendig (Buffler et al., 2009). Andererseits können sowohl mit dem Argumentieren (vgl. Abschnitt 2.1) als auch mit dem Experimentieren und dem Auswerten experimenteller Daten Charakteristika naturwissenschaftlichen Lernens wie *Scientific Practices, Scientific Litaracy, Data Literacy* und *Nature of Science* adressiert werden, welche als Ziel von naturwissenschaftlichem Unterricht angesehen werden können (Berland & Reiser, 2009; Jiménez-Aleixandre & Crujeiras, 2017; Millar & Osborne, 1998; OECD, 2018; Schiepe-Tiska et al., 2012). Einerseits kann den Schüler:innen auf diese Weise ein korrektes Bild über das Wesen der Naturwissenschaften vermittelt werden. Andererseits nimmt die Relevanz von Aspekten wie beispielsweise *Data Literacy* im Hinblick auf die Gesellschaft, beispielsweise unter dem Aspekt *Critical Thinking*, zu. *Critical Thinking* kann als „process by which people make decisions about what to trust and what to do“ (Heim et al., 2022, S. 1) verstanden werden. Die Bedeutung dieser Fähigkeit nimmt in Zeiten einer hohen Datenflut in wirtschaftlichen und politischen Belangen (Gould, 2017) und der verstärkten Verbreitung von Fehlinformationen und Verleugnung wissenschaftlicher Fakten (Barzilai & Chinn, 2020) für die Lernenden stetig zu. Diesbezüglich kommt dem adäquaten Umgang mit Daten und Unsicherheiten eine essenzielle Rolle zu, da der Umgang als Teil von *Data Literacy*, *Information Literacy* und *Statistical Literacy* einen wichtigen Beitrag zum *Critical Thinking* leistet (Schield, 2004). Schield (2004) zeigt in seiner Arbeit dabei, wie die drei genannten Aspekte zusammenhängen:

> Data literacy is needed to access, manipulate and summarize the data. But statistical literacy is needed to guide in that process while information literacy sets the overall context for evaluating the sources of data and the appropriate manipulations. (S. 8)

Die dargelegten Aspekte verbindet, dass sie die Förderung der Datenkompetenz ebenfalls adressieren. Nachfolgend wird daher deklariert, welches Wissen bzw. welche Kompetenzen die Schüler:innen erlangen bzw. ausbilden müssen.

Aus fachwissenschaftlicher Perspektive lässt sich die Frage nach dem adäquaten Umgang mit Daten und Messunsicherheiten mit dem Verweis auf den GUM einfach beantworten. Alle relevanten Aspekte der Messtechnik und Metrologie sind dort entsprechend dargestellt. Der GUM ist jedoch recht umfangreich und weist mittlerweile sechs Teile auf (JCGM, 2020). Zudem übersteigen die Inhalte des GUM die für die Schüler:innen benötigten Kompetenzen auf fach-

licher Ebene sowohl bezüglich der mathematischen Voraussetzungen als auch der angewendeten Verfahren (vgl. Abschnitt 2.3.2). Es wird daher auf die theoretischen Ausarbeitungen von Fachdidaktiker:innen zurückgegriffen, in denen entsprechende Kompetenzen formuliert werden.

Duggan und Gott (2002) sowie Millar und Osborne (1998) sind der Frage nachgegangen, welche Ziele der naturwissenschaftliche Unterricht der *Zukunft* aufweisen sollte. Duggan und Gott (2002) heben dabei die Bedeutung von Evidenz und Unsicherheiten hervor:

> The ability to sift through evidence, recognize when it was „good“ evidence or recognize where there was no evidence were crucial in action groups and in informed personal decision making. Unterstanding concepts such as risk and uncertainty were also crucial. (S. 674)

Beide genannten Arbeiten betonen die Rolle einer *Scientific Literacy*, fokussieren sich jedoch auf unterschiedliche Ebenen des Verständnisses: Duggan und Gott (2002) nennen das *Procedural Knowledge* als wichtige Komponente der naturwissenschaftlichen Bildung, wohingegen Millar und Osborne (1998) das *Conceptual Knowledge* als wichtiges Element identifizieren. Nach Hiebert und Lefevre (1986) kann unter *Conceptual Knowledge* „a network in which the linking relationships are as prominent as the discrete pieces of information“ verstanden werden (Hiebert & Lefevre, 1986, S. 3 f.). Als *Procedural Knowledge* kann das Wissen über Algorithmen, Regeln und Strategien zum Problemlösen angesehen werden (Hiebert & Lefevre, 1986, S. 7 ff.). Bezüglich der von Duggan und Gott (2002) genannten Evidenz ist es somit notwendig, dass Schüler:innen einerseits das Wissen darüber haben, was Evidenz ist und wie sie generiert werden kann, und andererseits die Evidenz zur Lösung von Problemen anwenden können.

Vor dem Hintergrund des experimentellen Kontextes kann die Evidenz auf Grundlage von Beobachtungen oder Messdaten gewonnen werden. Ein Ziel eines adäquaten Umgangs mit Daten und Messunsicherheiten ist somit, die Schüler:innen dazu zu befähigen, im Rahmen des Experimentierens aus den Messdaten Evidenz generieren und anwenden zu können (Gott & Duggan, 2003; Millar et al., 1994). Welche Kompetenzen für diese Evidenzgewinnung notwendig sind, ist beispielsweise Pols (2019) nachgegangen. In seiner Arbeit hat er verschiedene internationalen Bildungspläne auf Gemeinsamkeiten untersucht und daraus zehn Datenkompetenzen generiert (s. Tabelle 2.1). Diese Datenkompetenzen erheben keinen Anspruch auf Vollständigkeit, zeigen jedoch wesentliche Elemente, um aus Daten Evidenz gewinnen zu können. Es fällt auf, dass diese Kompetenzen einen Schwerpunkt auf die Auswertung von Daten aufweisen und nur periphär auf die Bedeutung der Messunsicherheit eingehen. Da Unsicherheiten ein Teil der Messung und somit unvermeidlich sind (Hellwig, 2012), wird daher nachfolgend die Bedeutung der Daten *und* Messunsicherheiten beim Experimentieren dargestellt.

Tab. 2.1: Zehn Datenanalysekompetenzen, die sich aus internationalen Bildungsplänen synthetisieren lassen. Übernommen nach Pols (2019).

1	Visualise data graphically
2	Use theoretical models to infer wether the origin constitutes a measurement
3	Draw a line of best fit through a data set and articulate that choice
4	Distinguish between a linear and non-linear relation
5	Qualitatively describe a dataset or trend
6	Describe qualitative similarities and differences between datasets
7	Draw a conclusion which is supported by the dataset
8	Indicate restrictions concerning the data-analysis and conclusions
9	Estimate the calue of a point within the range of available data (interpolation)
10	Estimate the value of a point outside the range of available data (extrapolation)

2.3.4 Daten und Unsicherheiten beim Experimentieren

In Abschnitt 2.2.1 wurde sowohl die Bedeutung des naturwissenschaftlichen Experiments als auch dessen Eignung für das Argumentieren hervorgehoben. Im weiteren Verlauf wird daher der Schwerpunkt darauf gelegt, welche Relevanz ein adäquates Verständnis von Daten und Unsicherheiten beim Experimentieren aufweist.

Eine wichtige Grundlage für die vorliegende Untersuchung ist die Arbeit von Hellwig (2012). Im Rahmen ihrer Dissertation hat Hellwig (2012) ein Sachstrukturmodell zum Umgang mit Daten und Messunsicherheiten entwickelt. In der Arbeit wird betont, dass Schüler:innen in der Lage sein sollten, „quantitative Experimente durchzuführen, auszuwerten und auf dieser Grundlage eine gegebene wissenschaftliche Fragestellung beantworten zu können“ und ein Verständnis dafür zu entwickeln, „wie in den Naturwissenschaften mithilfe von experimenteller Forschung neues Wissen generiert wird“ (Hellwig, 2012, S. 5). Das Strukturmodell enthält die vier Dimensionen *Grundsätzliche Existenz von Messunsicherheiten, Einfluss von Messunsicherheiten auf das Messwesen, Erfassung von Messunsicherheiten* und *Aussagekraft von Messunsicherheiten*, welche jeweils mehrere Konzepte beinhalten, die in ihrer Gesamtheit ein umfassendes Verständnis von Messunsicherheiten darstellen. Neben der Erfassung von Messunsicherheiten und der Auswertung von Daten unter Berücksichtigung von Messunsicherheiten zeigt das Modell, dass ein adäquates Verständnis von Messunsicherheiten bereits in der Planung des Experiments beginnt (Ziel der Messung). Messunsicherheiten sind daher nicht nur bei der direkten Erfassung von Messgrößen relevant, sondern auch in der Vorbereitung, Planung und Auswertung eines Experiments zu berücksichtigen.

Gott und Duggan (1995) untersuchen in ihrem Buch das forschende Lernen

im naturwissenschaftlichen Unterricht, in dem das Experimentieren eine wichtige Rolle einnimmt. Unter anderem stellen die Autoren vor, welche Einflussfaktoren die Evidenz im Rahmen des Experiments beeinflussen. Als Faktoren werden *Design, Measurement* und *Data Handling* genannt (Gott & Duggan, 1995, S. 110). Eine explizite Nennung von Messunsicherheiten gibt es in der Arbeit von Gott und Duggan (1995) nicht, jedoch adressieren die verschiedene Faktoren ein adäquates Verständnis von Daten und Messunsicherheiten implizit: Unter *Design* verstehen die Autoren Aspekte wie die relevanten Variablen für die Untersuchung zu bestimmen und adäquate Variablenkontrollstrategien anzuwenden. Beide Aspekte sind vor dem Messen zu berücksichtigen und reduzieren die Chance für Messfehler während des Messprozesses. Der Aspekt *Measurement* stellt den umfangreichsten Bereich dar und berücksichtigt Aspekte wie die richtige Skala und eine sinnvolle Genauigkeit zu verwenden oder auch Messungen gegebenenfalls zu wiederholen. Diese Aspekte beinhalten ein adäquates Verständnis von Messunsicherheiten und adressieren indirekt bereits beispielsweise Konzepte aus dem Sachstrukturmodell von Hellwig (2012). Der Aspekt *Data Handling* berücksichtigt unter anderem, ob Daten korrekt interpretiert oder Grafiken angemessen angefertigt wurden. In beiden Aspekten ist ein Umgang mit Unsicherheiten eminent. Zusammenfassend stellt die Arbeit von Gott und Duggan (1995) übersichtlich dar, welche Einflussfaktoren für die Generierung von Evidenzen aus Messdaten beim Experimentieren notwendig sind. Gleichzeitig wird deutlich, dass der Messprozess beim Experimentieren wichtig ist, viele der zuvor genannten Faktoren von Unsicherheiten beeinflusst werden und ein adäquates Verständnis von Daten und Unsicherheiten förderlich für Evidenzen ist. Die dargestellten Einflussfaktoren können verschiedenen Konzepten des Sachstrukturmodells zugeordnet werden und verdeutlichen somit, welche Relevanz der explizite und implizite Umgang mit Messunsicherheiten für das Experimentieren aufweist.

Millar und Osborne (1998) sind der Frage nachgegangen, was einen erfolgreichen naturwissenschaftlichen Unterricht ausmacht und welche Anforderungen an den naturwissenschaftlichen Unterricht gestellt werden. Dabei wurde als zentrales Ziel eine *Scientific Literacy* von Schüler:innen deklariert (Millar & Osborne, 1998, S. 2001). Unter anderem formulieren die Autoren dabei das Ziel, dass die Lernenden durch „practical work“ (Millar & Osborne, 1998, S. 2001), zu dem auch das Experimentieren zu zählen ist, selbstständig lernen sollen. Beispielsweise sollen Schüler:innen in der Lage sein, einfache Untersuchungen zur Untersuchung der Beziehung zweier Variablen zu planen. Neben der Planung der Experimente wird zusätzlich explizit auf Messunsicherheiten eingegangen:

> In the early years of secondary school, pupils should learn, through their own practical work and in other ways:
>
> - that no observation or measurement can ever be sure of matching exactly the ‘true’ value - that there is always some uncer-

tainty;

- that repeating measurements and taking an average is a good method for reducing the effect of random error; [...].
(Millar & Osborne, 1998, S. 2001)

Beide Aspekte beziehen sich auf die direkte Messung von Messwerten. Die genannten Konzepte lassen sich, wie bereits die Konzepte von Duggan und Gott (2002), im Sachstrukturmodell von Hellwig (2012) verorten: Die Dimension *Ursachen der Messunsicherheit* beinhaltet Konzepte, welche die unterschiedlichen Quellen von Messunsicherheiten darstellen und verdeutlichen, dass eine Messung immer unsicherheitsbehaftet ist. In der Dimension Erfassung von Messunsicherheiten finden sich Konzepte, welche unterschiedliche Verfahren zur Unsicherheitsermittlung in Abhängigkeit des Messverfahrens vermitteln. Die „Ermittlungsmethode A“ (Hellwig, 2012, S. 67) beinhaltet dabei den von Millar und Osborne (1998) genannten Aspekt der Mehrfachmessung von Messwerten.

Die zuvor genannte Wichtigkeit eines adäquaten Verständnisses von Daten und Unsicherheiten kann an den Messdaten beobachtet werden, die für diese Arbeit verwendet und von Schüler:innen ausgewertet worden sind: Betrachtet man beispielsweise die Messdaten aus Tabelle 6.1, zeigt sich, dass die Messwerte derselben Messreihe streuen. Eine reine Betrachtung des Mittelwerts würde zu dem Schluss führen, dass mit wachsender Pendelmasse eine Zunahme der Periodendauer zu beobachten ist. Diese Schlussfolgerung ist jedoch fachlich falsch, da die genannten physikalischen Größen im Kontext des mathematischen Pendels unabhängig voneinander sind. Werden die Messreihen unter Berücksichtigung der Messunsicherheit miteinander verglichen, zeigt sich, dass die Ergebnisse sich überschneiden und somit nicht unterscheidbar voneinander sind. Daraus ergibt sich die fachlich korrekte Schlussfolgerung, dass die Erhöhung der Pendelmasse keinen Einfluss auf die Periodendauer nimmt.

Die exemplarisch dargelegten Arbeiten sollen aufzeigen, dass die Relevanz des adäquaten Umgangs mit Daten und Unsicherheiten bezüglich des Experimentierens bereits seit einiger Zeit Gegenstand der fachdidaktischen Forschung ist. In Abschnitt 2.3.5 wird dieser Aspekt vor dem Hintergrund des Experiments und der Präkonzepte von Schüler:innen genauer beleuchtet. Neben der fachdidaktischen Perspektive gibt es ebenso fachwissenschaftliche Literatur, welche einen adäquaten Umgang mit Daten und Unsicherheiten beim Experimentieren und praktischen Arbeiten adressiert und für weiterführende Analysen herangezogen werden kann (bspw. Krystek, 2020; Möhrke & Runge, 2020; Taylor et al., 1998).

2.3.5 Fachdidaktische Arbeiten zu Messdaten im Physikunterricht

Nachfolgend werden ausgewählte fachdidaktische Arbeiten dargestellt, welche das Verständnis von Messunsicherheiten beim Experimentieren und den adäquaten Umgang mit Daten und Unsicherheiten bei Student:innen bzw. Schüler:innen untersucht haben.

Die Relevanz eines adäquaten Umgangs mit Daten und Unsicherheiten beschränkt sich nicht nur auf den Schul- bzw. Universitätskontext: Duggan und Gott (2002) haben untersucht, welche Relevanz das prozedurale Verständnis von Evidenz für Beschäftigte der Industrie und der Öffentlichkeit, die alltäglich mit Wissenschaft interagieren, besitzt. Es konnte dabei gezeigt werden, dass ein ausgeprägtes Verständnis von Evidenzen verstärkt dazu führt, dass alltägliche Probleme effektiv gelöst werden können. In einer Untersuchung konnte gezeigt werden, dass beispielsweise Pflegepersonal im Krankenhaus bei der Betreuung von Patienten ebenfalls auf Konzepte von Evidenz zurückgreift (Aikenhead, 2005). Eine explizite Berücksichtigung von Messunsicherheiten oder -abweichungen findet im letzten Beispiel jedoch nicht statt. Neben einer *Scientific Literacy* als allgemeines Bildungsziel für Schüler:innen zeigt sich, dass das prozedurale Wissen über Evidenz der Lernenden ebenfalls für verschiedene Berufe und alltägliche Probleme gebraucht werden kann. Duggan und Gott (2002) fordern daher unter anderem, dass grundsätzliche Konzepte von Evidenz und deren Anwendung in der Schule gefördert werden müssen (Duggan & Gott, 2002, S. 674).

Wie in Abschnitt 2.3.4 dargelegt wurde, kann ein adäquater Umgang mit Daten und Unsicherheiten beim Experimentieren gefördert werden. Gleichzeitig zeigen Arbeiten wie beispielsweise von Kung (2005) auf, dass ein Verständnis der Messung und der Unsicherheiten notwendig ist, um Experimente erfolgreich zu beenden. Kung (2005) weist darauf hin, dass unter Wissenschaftler:innen verschiedene Methoden zur Abschätzung der Unsicherheiten existieren. Die Vermittlung von Konzepten sowie eines Verständnisses des Messprozesses und der adäquaten Auswertung soll daher im Fokus stehen und nicht die Berechnungsvorschrift. Kung (2005) stellt verschiedene Experimente vor und beschreibt, wie in diesem Zusammenhang die Konzepte vermittelt werden können. Nach einem konzeptbasierten Kurs konnte gezeigt werden, dass der Anteil der Schüler:innen, die zum Vergleich von Messreihen neben dem Mittelwert die Messunsicherheit verwenden, verdoppelt werden konnte.

Einen ähnlichen Ansatz untersuchten Buffler et al. (2009). Entwickelt wurde ein Kurs, welcher zum Ziel hatte, bei den Studierenden Konzepte zum Messen und Messunsicherheiten zu vermitteln, welche den for Standardization (2004)-Normen entsprechen. Hauptaugenmerk lag hierbei darauf, dass die Student:innen ein probabilistisches Verständnis für das Messen entwickeln. Es konnte gezeigt werden, dass das Verständnis der Studierenden bezüglich des Messens und der

Messunsicherheiten in verschiedenen Messsituationen verbessert werden konnte (Pillay et al., 2008). Kung (2005) merkt jedoch an, dass nur mit einem Kurs oder einer Sequenz nicht ein vollständiges Verständnis von Messunsicherheiten gefördert werden kann, sondern dass „[...] reaching such a stage would require a whole series of laboratory curriculum reforms“ (Kung, 2005, S. 772). Erschwerend kommt hinzu, dass trotz der dargestellten Relevanz eines adäquaten Umgangs mit Daten und Unsicherheiten das Thema in der Schule im naturwissenschaftlichen Unterricht oftmals vernachlässigt wird (Hellwig, 2012). Zusätzlich sind Lehrer:innen und Forscher:innen mit verschiedenen Schwierigkeiten beim Lehren konfrontiert, welche nachfolgend kurz dargestellt werden.

Schüler:innenvorstellungen zum Messen und zu Messunsicherheiten

Bereits in der Primarstufe ist es möglich, Schüler:innen für Messunsicherheiten zu sensibilisieren. Beispielsweise konnten Munier et al. (2013) in ihrer Untersuchung feststellen, dass Schüler:innen verschiedene Unsicherheitsquellen beim Messen benennen können. Außerdem war es nach Bearbeitung entsprechender Aufgaben möglich, den Schüler:innen Eigenschaften von Konfidenzintervallen näherzubringen.

Ähnliche Ergebnisse werden von Masnick und Klahr (2003) berichtet. Untersucht wurden ebenfalls Grundschüler:innen und ihr Verständnis von den Ursachen und Konsequenzen von Messfehlern. Dabei wurde eine neue Klassifizierung für Messfehler erstellt und die vier Kategorien „Design Error, Measurement Error, Execution Error“ und „Interpretation Error“ sind gebildet worden. Damit konnten Fehlerquellen verschiedenen Phasen des Experimentierens zugeordnet werden. Die Untersuchung hat unter anderem gezeigt, dass die Schüler:innen der Primarstufe Messfehlerursachen benennen und erkennen konnten. Die Autoren konkludieren daher, dass „these results tell us that second- and fourth-grade children understand many things about error“ (Masnick & Klahr, 2003, S. 96). Gleichzeitig weisen die Autoren darauf hin, dass die Kinder weiterhin viel über die Messfehler lernen müssten. Das Verständnis von Kindern in diesem Bereich sollte daher weiter von Forscher:innen untersucht werden, da fachdidaktische Forschung bezüglich des Messprozessverständnisses oftmals ältere Proband:innen zum Ziel hat (Masnick & Klahr, 2003).

Obwohl Schüler:innen der Grundschule durchaus über ein qualitatives Verständnis von Unsicherheiten beim Messen verfügen, können bei älteren Schüler:innen fachlich defizitäre Präkonzepte in diesem Zusammenhang identifiziert werden, die im Folgenden dargestellt werden. Beispielsweise haben Séré und Journeaux (1995) unter anderem gezeigt, dass für Lernende Messen die Suche nach dem *richtigen* Wert bedeutet. Abweichungen von diesem Wert werden als *Fehler* angesehen. Entgegen den Überzeugungen, die von Masnick und Klahr (2003) bei Grundschüler:innen festgestellt werden konnten, lehnen Schüler:innen eine Streuung von Messwerten ab (Séré & Journeaux, 1995). Ein Verständnis von Messwiederholungen fehlt den Schüler:innen, da diese nur vorgenommen werden,

um den ersten Messwert zu bestätigten bzw. bei zwei unterschiedlichen Werten zwischen dem ersten und zweiten Messwert durch erneute Messung entscheiden zu können (Séré & Journeaux, 1995). Diese Erkenntnis deckt sich mit anderen fachdidaktischen Arbeiten (Kanari & Millar, 2004; Lubben & Millar, 1996). Lubben und Millar (1996) konnten zeigen, dass Schüler:innen kein Verständnis einer reliablen Messung haben. Zwar wurde eine Messwiederholung nicht verneint, jedoch konnte gezeigt werden, dass mit steigendem Alter und steigender Erfahrung der Proband:innen eine Veränderung des Messprozessverständnisses beobachtet werden konnte. Während die jüngeren Lernenden wie bei Séré und Journeaux (1995) eine Messwiederholung eher als Bestätigung für bestehende Messwerte sahen, konnte bei älteren Schüler:innen gezeigt werden, dass die Streuung der Messwerte bei der Datenauswertung berücksichtigt wird. Dabei wurde von den Lernenden der Mittelwert verwendet, welcher für Schüler:innen weniger intuitiv ist als andere zentrale Lagemaße (Séré et al., 1993). Kanari und Millar (2004) haben das Messen der Schüler:innen sowie die Datenauswertung im Rahmen von zwei Experimenten untersucht. Dabei mussten die Schüler:innen entweder einen Gegenstand mittels Kraftmesser über eine Ebene ziehen oder beim Fadenpendel mögliche Einflussfaktoren der Periodendauer untersuchen, wobei verschiedene Versuchsanweisungen hinsichtlich der Kovariation beim Messprozess vorgegeben worden sind. Es konnte unter anderem gezeigt werden, dass die Schüler:innen unterschiedlich häufig bei den beiden Experimenten gemessen haben. Dennoch deckt sich das gezeigte Verständnis der Lernenden bezüglich des Messprozesses mit den vorherigen dargelegten Arbeiten. Die vorangehenden Arbeiten legen somit nahe, dass Schüler:innen eine Messwiederholung nicht unmittelbar mit der Reliabilität einer Messung in Einklang bringen können.

Lubben et al. (2001) haben das prozedurale Datenverständnis von Studienanfänger:innen untersucht. Dabei wurden neben dem Verständnis einer Messwiederholung und der Streuung von Messwerten ebenfalls Aspekte wie der Messprozess an sich sowie die Datenauswertung und der Datenvergleich untersucht. Es konnte gezeigt werden, dass die Begründungen der Studienanfänger:innen in zwei Typen kategorisiert werden können: „point and set reasoning“ (Buffler et al., 2001; Lubben et al., 2001). Das *Point Reasoning* zeichnet sich durch das Konzept aus, dass jede Messung der richtige Wert sein könnte und einzelne Messwerte Messergebnisse darstellen. Mehrfache Messungen stellen Ausnahmen dar und erfüllen dabei eher die Ziele, welche im vorigen Absatz dargestellt wurden (Lubben et al., 2001, S. 312). Beim *Set Reasoning* wird angenommen, dass jede Messung nur eine Approximation des wahren Werts darstellt. Damit ergeben sich Konsequenzen für die Messung:

> As a consequence, a number of measurements are required to form a distribution that clusters around some particular value. The best information regarding the true value is obtained by combining the measurements using theoretical constructs in order to describe the

data collectively. (Lubben et al., 2001, S. 313)

Eine Auswertung der Messung erfolgt, indem ein vollständiges Messergebnis bestehend aus dem Mittelwert und der Standardabweichung aufgestellt wird. Beide Aspekte weisen eine hohe Relevanz auf, wenn es um den Datenvergleich geht (Lubben et al., 2001, S. 312). Das vorgestellte Instrument zur Klassifizierung von datenbasierten Begründungen von Schüler:innen findet in vielen fachdidaktischen Arbeiten Verwendung bzw. wird zu komplexeren Modellen in der Datenanalyse erweitert (bspw. Evangelinos et al., 2002; Kung & Linder, 2012; Masnick & Morris, 2022; Pols et al., 2021). Dabei ist es möglich, dass Mischargumentationen entstehen, die auf der Skala zwischen *Point Reasoning* und *Set Reasoning* zu verorten sind (Buffler et al., 2001; Volkwyn et al., 2004). Kung (2005) weist darauf hin, dass Schüler:innen Routinen zur Auswertung von Daten verwenden, beispielsweise Mittelwerte und Standardabweichungen, aber die zugrunde liegenden Konzepte nicht verstanden haben und somit nicht in der Lage sind, datenbasierte Argumente zu bilden. Diese Vorstellung der Lernenden kann als Beispiel für die Mischargumentation genommen werden, welche Volkwyn et al. (2004) in ihrer Arbeit darstellen. Die Mischargumentation zeichnen sich dadurch aus, dass den Lernenden bewusst ist, dass Messsunsicherheiten existieren und der *exakte* Wert daher experimentell nicht ermittelbar ist. Die Schüler:innen sind jedoch weiterhin der Überzeugung, dass das Ziel der Messung die Bestimmung des *richtigen* Werts ist. Die Abgrenzung zur Set-Argumentation besteht darin, dass die Lernenden, welche die Mischargumentation verwenden, den probabilistischen Charakter der Messung und des Messergebnisses nicht internalisiert haben und ein deterministisches Modell weiterhin die dominate Vorstellung der Schüler:innen ist (Lubben et al., 2001). Zusammenfassend lässt sich erkennen, dass die Vorstellungen von Lernenden im Spektrum des Point- und Set-Paradigmas abgebildet werden können und Schüler:innen ein intuitives und propädeutisches Verständnis für Messunsicherheiten und Streuung besitzen (Schulz, 2022), jedoch trotz der bildungspolitischen Vorgaben in diesem Bereich Schüler:innen weiterhin Defizite aufweisen können (Pols et al., 2021).

Ein fachlich defizitäres Verständnis von Daten und Unsicherheiten kann sich negativ auf andere Aspekte des Messprozesses sowie der Entscheidungsfindung von Schüler:innen auswirken: Beispielsweise haben Kok et al. (2019) untersucht, welchen Einfluss die Anzahl an Nachkommastellen auf die Schlussfolgerungen von Schüler:innen hat. Dabei wurde die Hypothese der Lernenden zweimal erhoben: nach Betrachtung des Experiments und nach Bereitstellung eines entsprechenden Datensatzes. Es konnte unter anderem gezeigt werden, dass mit zunehmender Anzahl an Nachkommastellen der Anteil an Lernenden, die von einer falschen zu einer korrekten Hypothese wechseln, abnimmt und der Anteil an Lernenden, die von einer fachlich richtigen zu einer falschen Hypothese wechseln, zunimmt. Kok et al. (2019) führen dieses Verhalten der Lernenden auf ein fehlendes Verständnis von Messunsicherheiten und Streuung zurück. Ein

anderes Beispiel zeigt sich in der Arbeit von Ludwig et al. (2019). In dieser Arbeit wurde die Hypothese der Schüler:innen vor und nach dem Experimentieren erhoben. Es konnte unter anderem gezeigt werden, dass Schüler:innen sich der beim Experimentieren zu berücksichtigenden Unsicherheiten bewusst sind. Eine Untersuchung des Einfluss dieses *Bewusstseins* für Messunsicherheiten zeigt einen negativen Einfluss auf die Richtigkeit der fachlichen Hypothese. Ludwig et al. (2019) führen dies darauf zurück, dass Schüler:innen sich konservativ verhalten und ihre falschen Hypothesen beibehalten, obwohl die unsicherheitsbehafteten Daten erkannt worden sind. Eine mögliche Erklärung von Ludwig et al. (2019) ist, dass Lernende aufgrund des fehlenden Wissens und Verständnisses von Daten und Unsicherheiten nicht in der Lage sind, die Daten adäquat auszuwerten. Sie folgern daraus: „This is strong evidence for the claim that we do need to foster explicitly students' abilities in evaluating data and uncertainties in school science labs“ (Ludwig et al., 2019, S. 4).

Die exemplarischen Arbeiten von Kok et al. (2019) und Ludwig et al. (2019) weisen einen ähnlichen Rahmen auf, unterscheiden sich jedoch in der Art der Datenbereitstellung. Während Kok et al. (2019) die Daten den Schüler:innen bereitstellten, wurden die Daten bei Ludwig et al. (2019) von den Schüler:innen selbst erhoben. Den Einfluss dieser unterschiedlichen Datenbereitstellung auf die Evaluation von Daten im naturwissenschaftlichen Unterricht hat Pfeiler (2019) untersucht. Die Evaluation der Daten wurde in diesem Fall als Glaubwürdigkeit der Daten erfasst. Zunächst entwickelte Pfeiler (2019) ein Codiersystem, anhand dessen die Glaubwürdigkeit der Daten unter Berücksichtigung der von für Schüler:innen relevanten Einflussfaktoren evaluiert wird. Das System umfasst die übergeordneten Kategorien *Eigenschaften des Experiments, Eigenschaften von Autoren, Eigenschaften der Daten* und *Prüfen/Abgleichen*. In einer quantitativen Untersuchung, in der die Proband:innen zunächst selbst experimentiert haben und anschließend je nach Gruppenzugehörigkeit mit den eigenen Daten, denen eines bzw. einer Mitschüler:in oder einer Lehrperson konfrontiert wurden, wurden Unterschiede zwischen First- und Second-Hand-Data untersucht. Zwischen den drei Gruppen konnten keine Unterschiede in der Glaubwürdigkeit für Schüler:innen festgestellt werden. Ebenso unterschied sich das Verhalten der Schüler:innen bezüglich ihrer Hypothesen zum Experimentieren in den Gruppen nicht. Es zeigte sich, dass für die Schüler:innen andere Faktoren, wie beispielsweise die Streuung der Datensätze oder die Passung zur aufgestellten Eingangshypothese, einen höheren Einfluss auf die Evaluation von Daten besitzen. Dabei konnte gezeigt werden, dass die Streuung für Schüler:innen für und gegen die Glaubwürdigkeit von Daten spricht.

2.4 Digitale Medien im Physikunterricht

Für die Entwicklung der Intervention wird auf digitale Lernaufgaben in Form von Lernapps zurückgegriffen. Grundsätzlich können digitale Medien im Physikunterricht in die Formate Hardware, Software und Online-Formate gegliedert werden (Kircher et al., 2020, S. 457 f.) und es kann davon ausgegangen werden, dass digitale Medien mittlerweile Einzug in den Klassenraum gehalten haben (Fleischer et al., 2020; Lorenz et al., 2017).

Die Wirksamkeit des Einsatzes digitaler Medien hinsichtlich des Lernens von Schüler:innen ist dabei nicht per se gegeben (Clark et al., 2009; Haßler et al., 2016; Hillmayr et al., 2020; Schnotz, 2008). Der Einsatz digitaler Medien ist ein komplexes Feld (Clark et al., 2009) und das Potenzial digitaler Medien im naturwissenschaftlichen Unterricht scheint noch nicht ausgeschöpft zu sein (Hanekamp, 2014). Der Einsatz digitaler Medien im Unterricht beruht oftmals auf dem Ersatz analoger Lehrmittel (Kramer et al., 2019).

Verschiedene Forschungsarbeiten konnten zeigen, dass der Einsatz digitaler Medien nur erfolgen sollte, wenn er didaktisch wertvoll ist, die Vielfalt der Medien genutzt wird oder selbst entwickelte Inhalte verwendet werden (Bano et al., 2018; Hanekamp, 2014; Hillmayr et al., 2020). Es lässt sich konstatieren, dass der Einfluss digitaler Medien von weiteren Einflussfaktoren nur schwer zu isolieren ist (Hamilton et al., 2016; Humes, 2017).

Zur Kategorisierung digitaler Medien wird in der vorliegenden Arbeit das SAMR-Modell von Puentedura (2006) verwendet. Die Kategorisierung erfolgt anhand der Funktionalität und der Aufgabenformate, die durch den Einsatz der digitalen Medien in den Unterricht implementiert werden. Dient ein digitales Medium dazu, ein analoges Medium zu ersetzen, so würde dieser Einsatz als Substitution (S) im SAMR-Modell kategorisiert werden. Ein Einsatz digitaler Medien im Unterricht, der die Funktionalität analoger Medien verstärkt, entspricht im SAMR-Modell dem Aspekt Augmentation (A). Werden durch digitale Medien Aufgabenformate erstellt, welche in analoger Form nur mit sehr hohem Aufwand formuliert werden können, da sie eine hohe Interaktion erfordern, oder nicht realisiert werden können, werden die Bereiche Modification (M) und Redefinition (R) adressiert.

Das SAMR-Modell wird weniger als grundlegendes Instrument zur Gestaltung der Apps verwendet, da unter anderem die Kontextfreiheit des Modells kritisch betrachtet wird (Hamilton et al., 2016), sondern dient zur Klassifikation der innerhalb der Lernapps verwendeten digitalen Elemente hinsichtlich ihrer Funktionalität. Ziel der Entwicklung ist, die verwendeten digitalen Aufgaben und Elemente in den Bereichen Modification und Redefinition zu verankern, da

einem Einsatz digitaler Medien unter Verwendung dieser Aufgabenformate im naturwissenschaftlichen Unterricht eine positive Wirkung auf den Lernerfolg nachgewiesen werden konnte (Clark et al., 2009; Hamilton et al., 2016; Hillmayr et al., 2020; Kramer et al., 2019; Romrell et al., 2014).

Weiterhin können digitale Lernapps genutzt werden, um aktives Lernen und eine Reflexion während des Lernprozesses zu ermöglichen (Moreno & Mayer, 2005). Gleichzeitig können durch digitale Lernapps beispielsweise Simulationen, Interaktionen und antwortbasierte Rückmeldungen verwendet werden, sodass eine autonome, individuelle und selbstregulierte Erarbeitung der Lerninhalte ermöglicht werden, die sich positiv auf die Motivation und den Lernerfolg auswirken können (Clark et al., 2009; Hillmayr et al., 2020; Pfost et al., 2020; Rutten et al., 2012). Die Umsetzung der genannten Eigenschaften digitaler Lernaufgaben wird in Abschnitt 4.1.3 motiviert und konkretisiert.

2.5 Zusammenfassung und Ableitung des Forschungsvorhabens

In den vorangehenden Abschnitten wurde die Bedeutung des naturwissenschaftlichen Argumentierens beim Lernen von Naturwissenschaften aufgezeigt. Anhand der exemplarischen Darstellung von fachdidaktischen Arbeiten zur Förderung der Argumentationsfähigkeit im naturwissenschaftlichen Unterricht sollten verschiedene Ansätze gegenübergestellt werden. Insbesondere wurden dabei das naturwissenschaftliche Experiment als Argumentationsanlass und die Relevanz eines adäquaten Verständnisses von Daten und Unsicherheiten für die Naturwissenschaften und den naturwissenschaftlichen Unterricht hervorgehoben. Dabei wurde auf ein adäquates Verständnis der Datenkompetenz eingegangen und der Bezug zum naturwissenschaftlichen Experiment hergestellt. Die dargestellten Aspekte des Argumentierens, des Experimentierens und eines adäquaten Verständnisses von Daten und Messunsicherheiten adressieren wichtige Bestandteile einer naturwissenschaftlichen Bildung. Die genannten Aspekte können verwendet werden, um bei Lernenden eine *Scientific Literacy* und *Scientific Practices* zu fördern. Die vorliegende Arbeit stellt in diesem Sinne eine Fortsetzung der Arbeit von Ludwig (2017) dar, der das Fachwissen der Schüler:innen über einen Mechaniktest ermittelt hat und unter anderem eine positive Wirkung des Fachwissens auf das rationale Argumentieren beim Experimentieren zeigen konnte (vgl. Abschnitt 2.2.1). Das naturwissenschaftliche Experiment kann als guter Anlass für das Argumentieren betrachtet werden: Behauptungen und Hypothesen müssen mit Beobachtungen und Daten in Einklang gebracht werden.

Dieser Prozess kann nach Toulmin (1958) als Argumentieren aufgefasst werden, setzt jedoch voraus, dass Schüler:innen experimentelle Daten korrekt auswerten können (Kanari & Millar, 2004). Vor diesem Hintergrund ist es naheliegend anzunehmen, dass ein adäquater Umgang mit Daten und Unsicherheiten relevantes Fachwissen für das Argumentieren beim Experimentieren darstellt. Daher ist es Ziel dieser Arbeit, zu untersuchen, wie eine Förderung der Datenkompetenz auf das Argumentieren beim Experimentieren auswirkt. Dazu werden im weiteren Verlauf der Arbeit drei Studien durchgeführt: Zunächst werden in einem ersten Schritt zehn digitale Lernapps entwickelt, die anschließend in einer zweiten Studie exemplarisch auf ihre Wirkung im Hinblick auf die Steigerung von Datenkompetenz untersucht werden (vgl. Kapitel 5.2). Im letzten Schritt wird dann der Zusammenhang zwischen der Datenkompetenz und dem naturwissenschaftlichen Argumentieren im Kontext des Experimentierens untersucht. Anschließend werden Forschungsfragen aufgestellt, welche die genannten Aspekte empirisch untersuchen sollen.

Kapitel 3

Entwicklungsziele und Forschungsfragen

Entwicklungsziele

Im Rahmen der vorliegenden Arbeit soll der Einfluss von Datenkompetenz auf das Argumentieren beim Experimentieren untersucht werden. Die Konzepte für den adäquaten Umgang mit Daten und Unsicherheiten werden über kurze Interventionen in Form von digitalen Lern-Apps in den Unterricht integriert. Der Ansatz beinhaltet daher eine indirekte Förderung der Argumentationsfähigkeit und agiert auf der Schüler:innenebene (vgl. Abschnitt 2.1.3). Es ergibt sich das folgende zentrale Entwicklungsziel für die Lern-Apps:

> Die Lernapps sollen Konzepte und Kompetenzen im Umgang mit Daten und Unsicherheiten vermitteln.

Anhand der zuvor dargelegten Grundzüge des Projekts werden die Lernapps unter Berücksichtigung der folgenden Kriterien entwickelt:

- modularer Einsatz
- maximale Bearbeitungsdauer von 45 Minuten
- Verwendung ab Klassenstufe 8
- Verwendung von Datensimulationen
- selbstregulierte Bearbeitung

Forschungsfragen und -hypothesen

Ein Ziel der Arbeit ist neben der Entwicklung von digitalen Lernapps zur Förderung eines adäquaten Umgangs mit Daten und Unsicherheiten (s. Abschnitte 3 und 4.1) die Förderung der Argumentationsfähigkeit von Schüler:innen beim Experimentieren. Für dieses Ziel ist es essenziell, die Wirksamkeit der Apps im Hinblick auf die Förderung eines adäquaten Umgangs mit Daten und Unsicherheiten zu evaluieren. Daraus resultiert Forschungsfrage 1:

1. Inwieweit fördert das selbstständige Bearbeiten der Lernapps die Ausprägung der jeweiligen Kompetenzen im Umgang mit Daten und Messunsicherheiten bei Schüler:innen?

In Abschnitt 2.3.5 wurden Ergebnisse verschiedener fachdidaktischer Arbeiten berichtet, die die Messdaten im Physikunterricht und im Physikstudium untersucht haben. Dabei konnte gezeigt werden, dass Interventionen, die Schüler:innen Konzepte eines adäquaten Umgangs mit Daten und Messunsicherheiten vermitteln, eine positive Wirkung haben können (bspw. Buffler et al., 2009; Kok, 2022; Kung, 2005). Bezüglich Forschungsfrage 1 wird daher angenommen, dass die durch die jeweiligen Apps adressierten Konzepte in den vier Kompetenzbereichen direktes und indirektes Messen, grafische Auswertung sowie Signifikanz gefördert werden können.

1. Die Kompetenzen, die durch die Apps in dem Kompetenzbereich ...
   a) ... Direktes Messen adressiert werden, werden bei den Schüler:innen durch die Bearbeitung der App gefördert.
   b) ... Indirektes Messen adressiert werden, werden bei den Schüler:innen durch die Bearbeitung der App gefördert.
   c) ... Grafische Auswertung adressiert werden, werden bei den Schüler:innen durch die Bearbeitung der App gefördert.
   d) ... Signifikanz adressiert werden, werden bei den Schüler:innen durch die Bearbeitung der App gefördert.

Um zu untersuchen, inwiefern die Argumentationsfähigkeit der Schüler:innen beim Experimentieren gefördert werden kann, wird auf das Instrument zur Erhebung der Art der Begründung von Ludwig (2017) zurückgegriffen. Die Betrachtung und Auswertung der Art der Begründung ermöglicht Rückschlüsse darauf, ob Schüler:innen beim Argumentieren auf Grundlage der experimentellen Messwerte Evidenzen formuliert haben. Forschungsfrage 2 untersucht dahingehend, welchen Einfluss die durch die Lernapps adressierten Konzepte und Kompetenzen im Umgang mit Daten und Unsicherheiten auf die Art der Begründung der Schüler:innen besitzen.

2. Welchen Einfluss haben die nach den Instruktionen erworbenen Kompetenzen im Umgang mit Daten und Messunsicherheiten auf die Ausprägung der Zustimmung zu den Argumentkategorien?

Wie in Abschnitt 3 motiviert und in Abschnitt 2.3 dargelegt, ist ein adäquates Verständnis von Daten und Unsicherheiten notwendig, um auf Grundlage von experimentellen Daten evidenzbasiert argumentieren zu können. Vor diesem Hintergrund ist davon auszugehen, dass die Zustimmung zu den rationalen und

evidenzbasierten Argumentkategorien Daten als Evidenz und Messunsicherheiten eine positive Wirkung durch die Datenkompetenz erfährt. Entsprechend wird für die Zustimmung der Argumentkategorien Intuition und Expertenwissen eine Abnahme der Zustimmung bei zunehmender Datenkompetenz erwartet, da diese Argumentkategorien gefühlsbasiert sind, auf Heuristiken basieren oder auf externe Autorität verweisen. Zusammengefasst ergeben sich die folgenden Forschungshypothesen:

2. Eine höhere Datenkompetenz der Schüler:innen wirkt sich ...
   a) ... negativ auf die Zustimmung zur Argumentkategorie Intuition aus.
   b) ... negativ auf die Zustimmung zur Argumentkategorie Expertenwissen aus.
   c) ... positiv auf die Zustimmung zur Argumentkategorie Messunsicherheiten aus.
   d) ... positiv auf die Zustimmung zur Argumentkategorie Daten als Evidenz aus.

Im Rahmen der Arbeit wird untersucht, wie sich unterschiedliche Arten von Begründungen auf das Argumentieren bzw. den Lernerfolg auswirken. Dabei wird zwischen den Begründungskategorien Intuition, Expertenwissen, Daten als Evidenz und Messunsicherheiten unterschieden. Die Wirkung der Zustimmung zu den Argumentkategorien auf die Richtigkeit einer Hypothese wird mit Forschungsfrage 3 untersucht.

3. Welchen Einfluss hat die Stärke der Verwendung der vier Argumentkategorien auf die Richtigkeit der Hypothese nach dem Auswerten?

Den Schüler:innen wurde im Rahmen der Studie ein unbekanntes Experiment gezeigt. Die fachlich richtige Hypothese lässt sich daher nur auf Grundlage der bereitgestellten Messdaten ableiten. Die Verbindung der eigenen Hypothese mit den Daten stellt in diesem Fall das Argumentieren dar. Dieser Prozess kann durch Datenkompetenz und Fachwissen begünstigt werden (Kanari & Millar, 2004; Ludwig, 2017). Für Forschungsfrage 3 wird daher davon ausgegangen, dass eine höhere Zustimmung zu den eher daten- und evidenzbasierten Argumentkategorien Daten als Evidenz und Messunsicherheiten einen positiven Effekt auf die Wahl der physikalisch korrekten Hypothese aufweist. Entsprechend wird vermutet, dass eine höhere Zustimmung zu den Kategorien Intuition und Expertenwissen mit einer fachlich falschen Hypothese einhergeht.

3. Eine höhere Zustimmung zur Argumentkategorie ...
   a) ... Intuition hat einen negativen Einfluss auf die Richtigkeit der Hypothese nach dem Experiment.

b) ... Expertenwissen hat einen negativen Einfluss auf die Richtigkeit der Hypothese nach dem Experiment.

c) ... Messunsicherheiten hat einen positiven Einfluss auf die Richtigkeit der Hypothese nach dem Experiment.

d) ... Daten als Evidenz hat einen positiven Einfluss auf die Richtigkeit der Hypothese nach dem Experiment.

Die vorliegende Arbeit untersucht darüber hinaus, in welchem Zusammenhang Datenkompetenz und die Richtigkeit einer Hypothese stehen. Es wird daher in Forschungsfrage 4 untersucht, welchen Einfluss Kompetenzen im Umgang mit Daten und Unsicherheiten auf das Aufstellen von physikalisch richtigen Hypothesen auf Grundlage gegebener physikalischer Kontexte und Daten haben.

4. Welchen Einfluss haben die Kompetenzen der Schüler:innen im Umgang mit Daten und Messunsicherheiten darauf, auf Grundlage gegebener physikalischer Kontexte und Daten physikalisch richtige Hypothesen aufzustellen?

Aus den zuvor dargelegten Arbeiten wird für Forschungsfrage 4 davon ausgegangen, dass Datenkompetenz einen positiven Einfluss auf die Wahl der korrekten wissenschaftlichen Hypothese aufweist. Es ergibt sich daher die folgende Forschungshypothese bezüglich Forschungsfrage 4.

4. Eine höhere Datenkompetenz der Schüler:innen resultiert in einer größeren Chance für die Richtigkeit der Hypothese.

Die dargestellten Forschungshypothesen werden in den Abschnitten 5.2 und 6.2 als statistische Hypothesen im Rahmen der praktischen Relevanz (Studie 2) bzw. Signifikanztestung (Studie 3) verwendet. Die Forschungshypothesen stellen gerichtete Alternativhypothesen dar (Döring & Bortz, 2016, S. 659). Die dazugehörigen Nullhypothesen werden nicht genannt. Für die Nullhypothesen werden die exklusiven gegenteiligen Effekte angenommen.

Kapitel 4

Entwicklung der Lernapps zum Umgang mit Daten und Messunsicherheiten

4.1 Entwicklung

Der nachfolgende Abschnitt soll die Entwicklung und Evaluation der Apps genauer darlegen. Dabei wird zunächst auf Grundlage der in den Abschnitten 2.3.2 und 2.3.5 dargelegten bildungspolitischen Vorgaben und fachdidaktischen Erkenntnisse eine didaktische Struktur entwickelt. Anschließend werden Eigenschaften der Apps dargestellt, welche in allen Modulen funktional implementiert wurden, und die konkrete Umsetzung digitaler Tools der Apps wird prototypisch anhand der App A1 gezeigt.

4.1.1 Didaktische Strukturierung

Ein Ziel der vorliegenden Arbeit ist die Entwicklung von digitalen Lernapps zur Förderung von Datenkompetenz im Umgang mit Daten und Unsicherheiten beim Experimentieren. Zur sinnvollen Abgrenzung der Inhalte innerhalb der Apps und zur Lernzieldefinition wurden die zu vermittelnden Kompetenzen im Rahmen einer didaktischen Strukturierung den einzelnen Apps zugeordnet. Bei der Erstellung der Strukturierung wurde auf das Sachstrukturmodell von Hellwig (2012) und entsprechende Fachliteratur zurückgegriffen (JCGM, 2008, 2012).

Somit basiert die didaktische Strukturierung auf zwei Fundamenten: Die fachliche Richtigkeit der entwickelten Strukturierung wird durch Fachliteratur wie JCGM (2008, 2012) sichergestellt. Darüber hinaus sollen die Apps sich auf die schulrelevanten Inhalte fokussieren. Aus diesem Grund stellen die im Bildungsplan des Landes Baden-Württemberg (Ministerium für Kultus, Jugend

Tab. 4.1: Übersicht der didaktischen Struktur. Dargestellt sind die Kompetenzbereiche A-D sowie die dazugehörigen Konzepte, jeweils nummeriert von 1 bis 3.

A Direktes Messen	**B Indirektes Messen**
A1 Streuung von Messwerten	B1 Addition absoluter Unsicherheiten
A2 Unsicherheitsbestimmung	B2 Addition relativer Unsicherheiten
A3 Messabweichung	B3 Zusammengesetzte Unsicherheit nach Gauß*
C Grafische Auswertung	**D Signifikanz**
C1 Unsicherheitsflächen	D1 Vergleich von Messergebnissen
C2 Ausgleichsgeraden	D2 Boxplots
C3 Drei-Geraden-Methode	D3 Hypothesentests*

* nicht Teil der Sekundarstufe I. Die Entwicklung einer App zu diesem Konzept ist nicht Gegenstand der vorliegenden Arbeit.

und Sport Baden-Württemberg, 2016) für die Sekundarstufe I und in fachdidaktischen Arbeiten wie Hellwig (2012) und Schulz (2022) formulierten Kompetenzen und Konzepte das zweite Fundament der didaktischen Strukturierung dar.

Hellwig (2012) berücksichtigt Richtlinien und Standards von Standardwerken (Hellwig, 2012, S. 54) und strukturiert in ihrer Arbeit die Inhalte und Konzepte der Metrologie (JCGM, 2008, 2012). Daraus wurden vier zentrale Dimensionen zum Verständnis und Umgang mit Messunsicherheiten herausgearbeitet (Hellwig, 2012, S. 55). Das fachliche Modell von Hellwig (2012) adressiert alle Aspekte des Themenkomplexes Messunsicherheiten und kategorisiert diese entsprechend einer vorangehenden Sachstrukturanalyse ebendieses Themenkomplexes (Hellwig, 2012, S. 53). Aus den einzelnen Aspekten und Unteraspekten des Modells (Hellwig, 2012, S. 55) wurde die didaktische Struktur erstellt, um so die Modularität der Lernapps (vgl. Abschnitt 3) und die Reduktion auf die für die Sekundarstufe I relevanten Kompetenzen im Umgang mit Daten und Messunsicherheiten gewährleisten zu können (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016).

Für die didaktische Strukturierung konnten so vier zentrale Kompetenzbereiche identifiziert werden. Zwei Kompetenzbereiche adressieren direktes und indirektes Messen und zwei Kompetenzbereiche die Auswertung von Messwerten und -ergebnissen. Eine Übersicht über die Konzepte kann Tabelle 4.1 entnommen werden. Die Grundfertigkeiten werden durch die Kompetenzbereiche *Direktes Messen* und *Indirektes Messen* abgedeckt. Der erste Kompetenzbereich beinhaltet die Bestimmung von Messunsicherheiten, die Dokumentation von

Messergebnissen sowie die Unterscheidung zwischen einer Messunsicherheit und -abweichung. Methoden und Strategien, um die Fortpflanzung von Unsicherheiten in mathematischen Operationen zu berücksichtigen, sind dem Kompetenzbereich Indirektes Messen zugeordnet.

Die Auswertung von Messwerten und -ergebnissen unter Berücksichtigung von Messunsicherheiten wird durch zwei Kompetenzbereiche abgedeckt. Einerseits fokussiert sich der Kompetenzbereich *Grafische Auswertung* auf die Darstellung von Unsicherheitsflächen im Koordinatensystem und die Eigenschaften einer Ausgleichsgeraden. Andererseits werden im Kompetenzbereich *Signifikanz* verschiedene Methoden zur Ermittlung der metrologischen Verträglichkeit von Messergebnissen untereinander oder mit Referenzwerten adressiert.

Die Modularität der Kompetenzbereiche wird gewährleistet, indem jedem Kompetenzbereich drei weitere Konzepte zugeordnet werden, die je von einer App adressiert werden (s. Tabellen 4.2-4.5). Nachfolgend werden die entsprechenden Konzepte der Kompetenzbereiche Direktes Messen, Indirektes Messen, Grafische Regression und Signifikanz aufgeschlüsselt. Zudem wird dargestellt, aus welchen fachwissenschaftlichen Aspekten und Konzepten des Sachstrukturmodells sich die Konzepte der Apps ergeben. Die Arbeit von Hellwig (2012) wird an dieser Stelle als Bezugspunkt verwendet, da für die Evaluation der Apps hinsichtlich der Wirksamkeit im Abschnitt 5.1.2 Items und Kompetenzen verwendet werden, die Schulz (2022) auf Grundlage des Sachstrukturmodells entwickelt hat. Dadurch können eine inhaltliche und methodische Konsistenz der Arbeit gewährleistet werden.

Als Bezeichnung werden die Kompetenzbereiche mit den Buchstaben A-D bezeichnet. Das entsprechende Konzept wird mit einer Ziffer codiert. Beispielsweise adressiert die App A2 den Kompetenzbereich Direktes Messen und zeigt den Schüler:innen die Bestimmung von Typ A und B Unsicherheiten (vgl. Tabelle 4.2).

Direktes Messen

Der Kompetenzbereich A (Direktes Messen) soll Schüler:innen ermöglichen, beim direkten Messen während des Experimentierens Messwerte und -ergebnisse adäquat unter Berücksichtigung der Messunsicherheiten zu notieren. Eine kurze Zusammenfassung des Kompetenzbereichs kann Tabelle 4.2 entnommen werden. Die Stufe A1 – Streuung von Messwerten – hat die Kenntnis über die grundlegende Existenz von Messunsicherheiten als Lernziel. Es werden qualitative und quantitative Aspekte zur Streuung von Messwerten adressiert. Als eine mögliche Abschätzung für die Unsicherheit wird der Maximalfehler eingeführt. In diesem Fall wird der Betrag der maximalen Abweichung zwischen Messwerten und arithmetischem Mittel als Unsicherheit verwendet. Darauf aufbauend wird das vollständige Messergebnis als Mittelwert $\pm$ Unsicherheit konstruiert.

Nachfolgend wird der Unterschied zwischen der didaktischen Strukturierung der vorliegenden Arbeit und dem Sachstrukturmodell von Hellwig (2012) verdeut-

Tab. 4.2: Darstellung des Kompetenzbereichs Direktes Messen, der dazugehörigen Kompetenz und der Aufteilung in Konzepte.

A Direktes Messen Schüler:innen können Messwerte adäquat dokumentieren und Messunsicherheiten mit unterschiedlichen Erfassungsmethoden angeben sowie Messunsicherheiten und -abweichungen unterscheiden.	
A1 Streuung von Messwerten	Jede Messung ist mit einer Unsicherheit versehen. Zur Angabe eines Messergebnisses gehört stets die Messunsicherheit.
A2 Unsicherheitsbestimmung	Unsicherheiten können prinzipiell auf zwei Arten bestimmt werden: Die Unsicherheit muss recherchiert werden oder sie kann mit statistischen Mitteln abgeschätzt werden.
A3 Messabweichung	Definition und Nennung der Eigenschaften der Messabweichung.

licht, der sich insbesondere in der Implementation mehrerer Aspekte des Sachstrukturmodells in ein Konzept der didaktischen Strukturierung ausdrückt. Das Konzept A1 adressiert mehrere Konzepte des Sachstrukturmodells von Hellwig (2012). Einerseits wird die Dimension *Ursachen der Messunsicherheit* adressiert, da der thematische Schwerpunkt der App die Definition und Eigenschaft von Messunsicherheiten sind. Es wird nur ein Teil des Unterpunkts *Unterscheidung zwischen Messunsicherheit und Messabweichung* adressiert. Außerdem beinhaltet das Konzept A1 Konzepte der Dimension *Einfluss von Messunsicherheiten auf das Messwesen*. Dem Unterpunkt *Ziel der Messung* und *Unkenntnis des wahren Werts* kommt eine große Bedeutung zu, da den Schüler:innen vermittelt wird, dass der wahre Wert nicht bestimmt werden kann und jede Messung mit einer Unsicherheit versehen ist. Außerdem werden die Dimension *Erfassung von Messunsicherheiten* und das Konzept *Erfassung einer Unsicherheitskomponente bei direkter Messung* adressiert. Die Schüler:innen lernen den zuvor definierten Maximalfehler als Abschätzung kennen und stellen damit das Messergebnis als reduzierte rechteckige Wahrscheinlichkeitsdichtefunktion dar, wobei ein Vertrauensbereich von 100 % berücksichtigt wird (Hellwig, 2012, S. 125). Abschließend werden Aspekte der Dimension *Messergebnis als Zusammenfassung sämtlicher Informationen* in das inhaltliche Konzept A1 integriert. Den Schüler:innen soll so vermittelt werden, dass das vollständige Messergebnis sich aus einem Mittelwert und einer Unsicherheit zusammensetzt.

Das Konzept A2 – Unsicherheitsbestimmung – adressiert die Typ A und Typ B Unsicherheiten. Es wird dargestellt, wie die Methoden verwendet werden und wann sie Anwendung finden. Die Methoden werden, didaktisch reduziert, als „Nachschlage-“ und „Wiederholmethode“ bezeichnet (Hellwig, 2012, S. 125).

Das Konzept A2 adressiert zwei Dimensionen des Sachstrukturmodells.

Durch die Bestimmung von Unsicherheiten und Formulierung von Messergebnissen wird das inhaltliche Konzept der Dimension *Aufstellen einer Wahrscheinlichkeitsdichtefunktion* aufgegriffen. Speziell wird das Konzept *Erfassung einer Unsicherheitskomponente bei direkter Messung* adressiert. Zudem beinhaltet das Konzept A2 Inhalte aus der Dimension *Ursachen der Messunsicherheit*. Den Schüler:innen werden unterschiedliche Ermittlungsmethoden, deren Anwendungsfälle und Einflussgrößen vermittelt. Damit werden die inhaltlichen Konzepte *Endlichkeit von Darstellungen, „Faktor Mensch“* und *Einflussgrößen* adressiert.

Im Konzept A3 – Messabweichung – wird die Messabweichung eingeführt und den Schüler:innen wird im Kontext von Experimentiersituationen, Grafiken und Versuchsprotokollen die Unterscheidung von Messunsicherheit und -abweichung gezeigt.

Dieses Konzept setzt sich aus Unteraspekten der Dimensionen *Unterscheidung zwischen Messunsicherheit und Messabweichung* und *Rückschlüsse auf die Messung* zusammen. Da die Lernenden in dem Konzept A3 die Unterschiede zwischen Messunsicherheit und -abweichung lernen, wird die erste Dimension direkt adressiert. Die zweite Dimension wird nur partiell in dem Konzept *Verlässlichkeit der Messung und ihres Ergebnisses* im Unterpunkt *Rückschlüsse auf die Messung* adressiert. Den Schüler:innen wird vermittelt, dass sowohl die Messunsicherheit als auch die Messabweichung die Unsicherheitsbilanz beeinflussen können. Anhand der Unsicherheitsbilanz kann untersucht werden, wie stark sich die Einflussgrößen auf diese auswirken.

Indirektes Messen

Beim Kompetenzbereich B (Indirekte Messen) sollen Schüler:innen lernen, wie Unsicherheiten sich fortpflanzen oder sich aus mehreren Bestandteilen zusammensetzen können. Systematisch wird dabei von elementaren Rechenoperationen bis hin zu beliebigen funktionalen Zusammenhängen variiert. Dabei wird sich an dem didaktisch reduzierten Vorgehen von Hellwig (2012) orientiert, anhand desen die Unsicherheiten als Summe absoluter Unsicherheiten bei Anwendung von Strichrechnung und als Summe relativer Unsicherheiten bei Anwendung von Punktrechnung sinnvoll abgeschätzt werden können(Hellwig, 2012, S. 133). Eine Übersicht des Kompetenzbereichs kann Tabelle 4.3 entnommen werden. Kompetenzbereich B gliedert sich ebenfalls in drei Konzepte, wobei das dritte Konzept inhaltlich eher für die Sekundarstufe II relevant ist (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016).

Die Konzepte B1 und B2 sollen den Schüler:innen die Fortpflanzung von Unsicherheiten bei elementaren Rechenoperationen vermitteln. Das Konzept B1 – Addition absoluter Unsicherheiten – adressiert die Addition und Subtraktion von unsicherheitsbehafteten Messwerten. Den Schüler:innen wird vermittelt, dass die Unsicherheit einer Summe bzw. einer Differenz von unsicherheitsbehafteten Messwerten aus der Summe der einzelnen Unsicherheiten besteht. Damit wird der bereits in Konzept A1 skizzierte Ansatz, 100 %-ige Vertrauensbereich zu

Tab. 4.3: Darstellung des Kompetenzbereichs Indirektes Messen, der dazugehörigen Kompetenz und der Aufteilung in Konzepte.

B Indirektes Messen	
Schüler:innen können Ergebnisse aus unsicherheitsbehafteten Messwerten bestimmen und die dazugehörigen Unsicherheiten bei Verwendung der Grundrechenarten (Addition, Subtraktion, Multiplikation und Division) angeben bzw. sinnvoll abschätzen. Auf einem erhöhten Niveau können die Schüler:innen die Fortpflanzung der Messunsicherheiten für jede Rechenart unter Verwendung der differentiellen Formel von Gauß bestimmen.	
B1 Addition absoluter Unsicherheiten	Verrechnung von Messwerten mit Konstanten und zusammengesetzten Unsicherheiten in physikalischen Gesetzmäßigkeiten (Strichrechnung).
B2 Addition relativer Unsicherheiten	Verrechnung von Messwerten mit Konstanten und zusammengesetzten Unsicherheiten in physikalischen Gesetzmäßigkeiten (Punktrechnung).
B3 Zusammengesetzte Unsicherheit nach Gauß	Weiterführende Rechnungen über mehrere Stufen, in denen Messwerte auftreten, allgemeiner Ansatz zur Unsicherheitsfortpflanzung.

nutzen, konsequent fortgeführt. Im Rahmen des Sachstrukturmodells lässt sich das Konzept in die Dimension *Fortpflanzung der Messunsicherheiten* einordnen. Es handelt sich dabei um das Konzept *Zusammensetzung der Messunsicherheit aus mehreren Komponenten* und dem Aspekt *schrittweise Bestimmung der Gesamtunsicherheit*. Das Konzept B1 adressiert dabei Summen bzw. Differenzen von gemessenen Größen und aus Messwerten sowie *exakten* Zahlen.

Das Konzept B2 – Addition relativer Unsicherheiten – beinhaltet die Fortpflanzung von Unsicherheiten bei Produkten und Quotienten. Dabei werden die relativen Unsicherheiten der entsprechenden Messwerte addiert, um die relative Unsicherheit des Ergebnisses zu bestimmen. Bei dieser Abschätzung der Unsicherheit handelt es sich ebenfalls um eine didaktische Reduktion, welche nicht immer den exakten 100 %-igen Vertrauensbereich abdeckt. Ein Beispiel hierzu findet sich in den Ausführungen zur Gestaltung der App B2 (s. Anhang A).

Den Abschluss des Kompetenzbereichs B bildet das Konzept B3 – Zusammengesetzte Unsicherheit nach Gauß – welches die Fortpflanzung von Unsicherheiten auf einen beliebigen funktionalen Zusammenhang über eine differentielle Formel herleitet. Dieses Konzept kann in dieselbe Dimension *Fortpflanzung der Messunsicherheiten* (Hellwig, 2012) eingeordnet werden. Aufgrund der Komplexität der Formel sowie der Voraussetzungen in Form von Differentialrechnung wird dieses Konzept nicht für die Sekundarstufe I aufbereitet.

Grafische Auswertung

Für die Auswertung von Messwerten und -ergebnissen werden zwei Kompetenzbereiche für die didaktische Strukturierung aufbereitet. Der Kompetenzbereich

C (Grafische Auswertung) ermöglicht den Lernenden, zwei physikalische Größen, die unsicherheitsbehaftet ermittelt und grafisch dargestellt werden, auf mögliche lineare Zusammenhänge zu untersuchen. Dabei werden unter anderem Unsicherheitsflächen im Koordinatensystem und Eigenschaften von Ausgleichsgeraden eingeführt. Der Kompetenzbereich gliedert sich in drei Konzepte, welche wiederum von drei Apps adressiert werden (vgl. Tabelle 4.4).

Das Konzept C1 – Unsicherheitsflächen – adressiert den Aspekt, dass grafisch aufgetragene Messwertepaare nicht nur in einer Dimension mit einer Unsicherheit behaftet sein können (vgl. App A1), sondern ebenfalls in der jeweils anderen. So entstehen bei der grafischen Auftragung der Messwertepaare Unsicherheitsflächen. So wird konzeptuell dargestellt, dass sowohl die abhängige als auch die unabhängige Messgröße Unsicherheiten aufweisen können. Im Sachstrukturmodell kann dieses Konzept mehreren Dimensionen zugeordnet werden. Einerseits wird die Dimension *Dokumentation von Messergebnissen* des Sachstrukturmodells (Hellwig, 2012) adressiert. Das Konzept C1 vermittelt den Lernenden, wie Unsicherheitsflächen grafisch dargestellt werden können. Dieser Aspekt kann als Teil des Konzepts *Ergebnis der Messung* und des Punkts *Dokumentation von Messergebnissen* angesehen werden. Die Darstellung der unsicherheitsbehafteten Messwerte stellt ein Messergebnis des Experiments dar, welches in den nachfolgenden Konzepten C2 und C3 noch mit einer Ausgleichsgeraden ergänzt werden kann. Andererseits wird die Dimension *Aufstellen einer Unsicherheitsbilanz* adressiert. Die Unsicherheitsbilanz, die im Rahmen des Konzepts aufgestellt wird, entspricht der Unsicherheitsfläche, welche alle im Rahmen der direkten und indirekten Messung aufgetretenen Messunsicherheiten bilanziert. Somit wird das Konzept *Zusammensetzung der Messunsicherheit aus mehreren Komponenten* durch das Konzept C1 abgedeckt.

Nach Darstellung der Unsicherheitsflächen im Koordinatensystem adressiert die App C2 – Ausgleichsgerade – die Eigenschaften einer Ausgleichsgeraden. Den Schüler:innen wird vermittelt, dass eine Ausgleichsgerade nur dann gültig ist, wenn sie alle Unsicherheitsflächen des dargestellten Datensatzes schneidet. Der Umgang mit Ausreißern wird an dieser Stelle explizit nicht aufgenommen. Im Rahmen des Sachstrukturmodells lässt sich das Konzept C2 zwei Dimensionen zuordnen. Die Dimension *Grafische Regression einer linearen Regression* und das Konzept *Grafische Durchführung einer linearen Regression* zeigen die Inhalte des Konzepts C2 auf. Eine gültige Ausgleichsgerade muss alle Unsicherheitsflächen schneiden. Ebenso wird durch das Konzept C2 die Dimension *Genauigkeit des Schätzwerts* adressiert. Einerseits kann anhand der eventuellen Existenz einer gültigen Ausgleichsgeraden die Gültigkeit eines theoretischen linearen Zusammenhangs und ebenso die Schätzung der Parameter evaluiert werden. Dieses Konzept kann in der zuvor genannten Dimension unter dem Konzept *Verlässlichkeit der Messung* unter dem Teilaspekt *Genauigkeit des Schätzwerts* verortet werden.

Tab. 4.4: Darstellung des Kompetenzbereichs Grafische Auswertung, der dazugehörigen Kompetenz und der Aufteilung in Konzepte.

C Grafische Auswertung Schüler:innen können zwei Größen mit Messunsicherheiten auf Grundlage von grafischen Regressionen auf lineare Zusammenhänge untersuchen. Die sinnvollste Ausgleichsgerade kann über die Drei-Geraden-Methode ermittelt werden.	
C1 Unsicherheitsflächen	Darstellung von Messwerten mit Unsicherheiten im Koordinatensystem. Einführung von Unsicherheitsbalken in x- und y-Richtung.
C2 Ausgleichsgeraden	Einführung der Grundlagen für die grafische lineare Regression. Schüler:innen lernen die Eigenschaften von Ausgleichsgeraden kennen.
C3 Drei-Geraden-Methode	Eine Ausgleichsgerade ist in den seltensten Fällen eindeutig. Eine sinnvolle Ausgleichsgerade kann über die Drei-Geraden-Methode ermittelt werden.

Das letzte Konzept der grafischen Auswertung C3 – Drei-Geraden-Methode – behandelt die Uneindeutigkeit von Ausgleichsgeraden. Während eine *beste* Ausgleichsgerade mathematisch durch die Minimierung der Residuen erreicht werden kann (z. B. Field et al., 2014, S. 248), ist dieses Konzept bzw. Verfahren in der Schule und insbesondere in der Sekundarstufe I aufgrund fehlender mathematischer Kenntnisse der Schüler:innen nur schwer umsetzbar. Es wird daher auf den Vorschlag einer sinnvollsten Ausgleichsgeraden von Hellwig (2012, S. 80) eingegangen. Zunächst werden eine möglichst steile und eine möglichst flache Ausgleichsgerade eingezeichnet. Die Winkelhalbierende dieser Ausgleichsgeraden kann als sinnvollste Ausgleichsgerade motiviert werden. Dieses Konzept findet sich im Sachstrukturmodell im Aspekt *Grafische Durchführung einer Regression* in der Dimension *Grafische Regression einer linearen Regression* wieder.

Signifikanz

Neben der grafischen Auswertung von Messwerten und -ergebnissen ist in der Schule die rechnerische Auswertung relevant. Dieser Aspekt wird in der didaktischen Strukturierung durch den Kompetenzbereich D (Signifikanz) adressiert. Die Schüler:innen lernen das Konzept der metrologischen Verträglichkeit kennen. Dabei wird auf die besondere Aussagekraft von Messunsicherheiten eingegangen. Eine Übersicht des Kompetenzbereichs und der dazugehörigen Konzepte kann Tabelle 4.5 entnommen werden.

Das Konzept D1 – Vergleich von Messergebnissen – beinhaltet die Unterscheidbarkeit von verschiedenen Messergebnissen. Die metrologische Verträglichkeit wird anhand der grafischen Überlappung der Unsicherheitsintervalle eingeführt und didaktisch reduziert. Sofern eine Überschneidung im Rahmen

der jeweiligen 100 %-igen Vertrauensbereiche vorliegt, sind die Messergebnisse nicht voneinander unterscheidbar. Die Verwendung dieses didaktisch reduzierten Verständnisses von metrologischer Verträglichkeit ist in der Sekundarstufe I legitim, da a) weiterhin eine Anschlussfähigkeit für weiterführende Vorgehen durch eine andere Art der Unsicherheitsabschätzung (vgl. App A1) motiviert werden kann und b) für Hypothesentests das gewissermaßen willkürlich gewählte Signifikanzniveau adressiert und zu 5 % geändert werden kann.

Das Konzept D1 entspricht einem Spezialfall der allgemeinen Darstellung aus dem Sachstrukturmodell. In der Dimension *Aussagekraft von Messunsicherheiten* beinhaltet das Konzept *Vergleich eines Messergebnisses mit einem Referenzwert* den Aspekt *Metrologische Verträglichkeit mit anderen Messergebnissen*, welches das Konzept D1 darstellt. Das Konzept D1 ist jedoch derart gestaltet, dass der Vergleich mit einem Literaturwert als Referenzwert ebenso motiviert werden kann.

Das Konzept D2 – Boxplots – veranschaulicht die Streuung von Messwerten anhand eines Boxplots. Die Betrachtung der Quantile ermöglicht, den Grad der Überlappung zu beurteilen und damit einen differenzierten Vergleich von Messergebnissen vorzunehmen. Dieses Konzept kann mehreren Dimensionen des Sachstrukturmodells zugewiesen werden. Erstens wird die Dimension *Vergleich eines Messergebnisses mit einem Referenzwert* adressiert. Dieser Aspekt des Konzepts C2 deckt sich mit dem Konzept C1. Der Boxplot ist eine alternative Darstellung der Messwerte. Eine solche Darstellung kann daraufhin ebenfalls verwendet werden, um die Messwerte unterschiedlicher Gruppen bzw. Experimentator:innen miteinander zu vergleichen. Es handelt sich dabei um den Aspekt *Verträglichkeit mit anderen Messergebnissen* im Konzept *Vergleich von Messwerten*. Ebenso adressiert das Konzept C2 die Dimension *Vergleich innerhalb einer Messreihe mit einem Referenzwert*. Unter Verwendung des Boxplots kann die Spannweite der Messung bestimmt werden und somit können auch Ausreißer und Anomalien thematisch aufgearbeitet werden. Abschließend wird die Dimension *Verlässlichkeit der Messung* adressiert. Anhand des Boxplots kann mit Schüler:innen über den Grad des Vertrauens in die Daten gesprochen werden. Eine Erklärung und Interpretation des Quantilbegriffs können für ein besseres Verständnis der Daten verwendet werden. Dies entspricht dem Konzept *Verlässlichkeit der Messung* und speziell dem Teilkonzept *Grad des Vertrauens*.

Auf der höchsten Ebene beim Vergleich von Messergebnissen setzt das Konzept D3 – Hypothesentests – an. Bei diesem Konzept handelt es sich wie beim Konzept B2 um vertiefende Inhalte. Den Schüler:innen sollen die Grundlagen eines Signifikanztests vermittelt werden. Dazu werden der Signifikanzbegriff, das Signifikanzniveau sowie die Fehler erster und zweiter Art eingeführt. Im Rahmen des Sachstrukturmodells kann das Konzept D3 nicht einsortiert werden. Die Dimension *Vergleich von Meswerten* beinhaltet in dem Konzept *Vergleich eines Messergebnisses mit einem Referenzwert* den Aspekt *Verträglichkeit mit anderen*

Tab. 4.5: Darstellung des Kompetenzbereichs Signifikanz, der dazugehörigen Kompetenz und der Aufteilung in Konzepte.

D Signifikanz	
Schüler:innen können Messergebnisse miteinander vergleichen und die Güte der Überlagerung bewerten. Auf einem erhöhten Niveau können die Schüler:innen die Grundlagen eines Hypothesentests nachvollziehen.	
D1 Vergleich von Messergebnissen	Unterscheidbarkeit wird als Überschneidung von Unsicherheitsintervallen eingeführt. Insbesondere über den grafischen Zugang und der dazugehörigen Überlappung der Intervalle wird die Verträglichkeit von Messergebnissen verdeutlicht.
D2 Boxplots	Einführung von Boxplots als weitere Darstellung eines Messergebnisses. Untersuchung der Quartile als propädeutischer Aspekt für die Einführung von Signifikanz.
D3 Hypothesentests	Einführung des Signifikanzbegriffs. Bedeutung des Signifikanzniveaus sowie Fehler erster und zweiter Art.

Messergebnissen keine Signifikanztestung, sondern als höchste Stufe das *normalized error ratio*, welches als Maß der Überlappung gedeutet werden kann (Hellwig, 2012, S. 77 f.). Die Überprüfung der metrologischen Verträglichkeit anhand von Signifikanztestung findet jedoch auch in Schulen Anklang (bspw. S. Neumann, 2021). Neben der alltägliche Relevanz von Signifikanz in der Gesellschaft wird das Konzept der Signifikanz im Fach Mathematik in der Sekundarstufe unterrichtet (KMK, 2012, S. 18 ff.), sodass das Konzept D3 für diese Lernenden eine sinnvolle Erweiterung der metrologischen Verträglichkeit darstellt.

4.1.2 Validität der Passung zwischen didaktischer Struktur und Sachstrukturmodell

Da die Entwicklung der didaktischen Struktur und damit einhergehend der Apps sowie die Evaluation der Wirksamkeit der Apps sich auf die Arbeiten von Hellwig (2012) und Schulz (2022) stützt, wurde die Reliabilität der Zuordnung von Aspekten des Sachstrukturmodells (Hellwig, 2012) in die Konzepte der didaktischen Struktur anhand eines Expertenratings überprüft. Jeder der Kompetenzbereiche A-D wurde von einem bzw. einer Expertin fachdidaktischer Forschung im Themenfeld Messunsicherheiten evaluiert, indem die Konzepte der didaktischen Strukturierung den Konzepten und Aspekten des Sachstrukturmodells zugewiesen werden sollten. Dabei wurden nur Konzepte und Aspekte evaluiert, die in der Sekundarstufe I Relevanz aufweisen. Die Expert:innen

Tab. 4.6: Übereinstimmung zwischen den Expert:innen und dem Autor bezüglich der Verträglichkeit der Konzepte der didaktischen Strukturen und des Sachstrukturmodells nach Hellwig (2012).

Konzept	A1	A2	A3	B1	B2	C1	C2	C3	D1	D2
Cohens κ	0,76	0,76	0,72	0,81	0,81	0,88	1	1	0,77	0,92

ordneten Aspekte des Sachstrukturmodells den Konzepten der didaktischen Strukturierung zu und konnten bei jeder Zuordnung angeben, ob das Konzept vollständig, teilweise oder nicht von Aspekten des Sachstrukturmodells adressiert wird. Für die Evaluation des Expertenratings wurde daher ein gewichtetes Cohens Kappa verwendet. Die Übereinstimmung der Expert:innen hinsichtlich der Beurteilung kann Tabelle 4.6 entnommen werden.

Bis auf das Konzept A3 lässt sich mit einem gewichteten Cohens Kappa von $\kappa \geq 0{,}75$ eine substanzielle bis perfekte Übereinstimmung zwischen den Expert:innen und dem Autor erkennen. Die im Vergleich mit den restlichen Konzepten geringere Zustimmung des Konzepts A3 lässt sich auf unterschiedliche Gewichtungen der Konzepte zurückführen. Das Konzept A3 wurde zwar auf dieselben Aspekte des Sachstrukturmodells (Hellwig, 2012) zurückgeführt, jedoch wurden die Konzepte unterschiedlich hinsichtlich der Bewertung „trifft vollständig zu“ und „trifft teilweise zu“ zugeordnet. In den restlichen Konzepten gab es neben der Übereinstimmung der adressierten Konzepte des Sachstrukturmodells ebenfalls stärkere Übereinstimmungen bezüglich der Gewichtung der Konzepte.

Abschließend lässt sich sagen, dass durch die Expertenstudie die inhaltliche Validität bestätigt werden kann. Damit können den Konzepten der didaktischen Strukturierung (vgl. Kapitel 4.1.1) und den dazugehörigen Apps eindeutig Konzepte und Aspekte des Sachstrukturmodells sowie auf deren Grundlagen formulierte Kompetenzen zugeordnet werden (Hellwig, 2012; Schulz, 2022). Auf diese Weise kann von einer guten Passung des Testinstruments von Schulz (2022) zur Ermittlung der in den Apps vermittelten Kompetenzen ausgegangen werden.

4.1.3 Methodische und didaktische Charakteristika der Lernapps

Ein Entwicklungsziel der zu entwickelnden Lernapps ist eine Form von Modularität. Darunter ist einerseits die inhaltliche Modularität zu verstehen, welche durch die didaktische Struktur gegeben ist (s. Kapitel 4.1.1). Andererseits sollen die Lernapps so gestaltet werden, dass die Lehrkraft keine weiteren fachdidaktischen Kenntnisse für die Integration ebendieser in den naturwissenschaftlichen Unterricht benötigt. Die Lernsequenzen sollen außerdem aus Gründen der Prak-

tikabilität eine Dauer von 20 bis 25 Minuten aufweisen. Dies soll die Verwendung der Lernapps durch Lehrkräfte sowie die Integration der Lernapps in den Unterricht begünstigen.

Die Zielgruppe der Intervention stellen Lernende der achten und neunten Klasse dar. Einerseits sind diese Lernenden mit dem Unterrichtsfach Physik vertraut. Andererseits kennen die Schüler:innen die Thematik Daten aus dem Mathematikunterricht (KMK, 2003) und konnten im Physikunterricht bereits Erfahrungen im Experimentieren sammeln. Ferner adressieren die Inhalte und Konzepte der Lernapps (s. Abschnitt 4.1.1) die Bildungsstandards der Mittleren Reife (KMK, 2004). Daher bietet sich die zuvor genannte Zielgruppe an, da unter Verwendung der Apps entsprechende Kompetenzen adressiert werden können.

Die Entwicklung von digitalen Apps wurde dabei grundlegend aus zwei Beweggründen heraus beschlossen. Die Verbreitung und Verwendung der Lernapps kann in digitaler Form wesentlich einfacher koordiniert und realisiert werden als bei analogen Lehrmaterialien. Die Apps wurden als webbasierte Shiny Applikationen als OER entwickelt, somit können beide zuvor genannten Aspekte durch die Bereitstellung eines zentralen Links adressiert werden (https://www.physikalische-schulexperimente.de/physo/Apps_zum_Umgang_mit_Daten_und_Messunsicherheiten). Außerdem kann die Aufbereitung der Inhalte, Konzepte und Aufgaben auf interaktive Elemente und Vorzüge digitaler Medien zurückgreifen (vgl. Abschnitt 2.4).

Darstellung der Inhalte und Konzepte

Bei der didaktischen Aufbereitung der Inhalte und Konzepte innerhalb der Lernapps werden unterschiedliche interaktive Gestaltungsmöglichkeiten von Shiny (Chang et al., 2021; R Core Team, 2021) verwendet, die nachfolgend kurz exemplarisch dargelegt werden. Im Rahmen der Lernapps C2 und C3 müssen die Lernenden Ausgleichsgeraden zu gegebenen Datensätzen anfertigen. Die Ausgleichsgeraden der Schüler:innen können beliebig oft gezeichnet und zudem vom System auf Korrektheit überprüft werden. Die Schüler:innen können sich dabei vollständig auf die Eigenschaften einer korrekten Ausgleichsgeraden konzentrieren und durch Anpassen von Kennwerten eines linearen Zusammenhangs die Ausgleichsgerade zum gegebenen Datensatz ermitteln.

Ein weiterer Vorteil der digitalen Umsetzung ist, dass Grafiken verändert und Hervorhebungen implementiert werden können. So können die Schüler:innen beispielsweise in der App D2, in der die prozentuale Überschneidung von Messergebnissen auf Grundlage von Boxplots visualisiert und ermittelt wird, durch Variation der Daten die Abbildung dahingehend verändern, dass der Grad der Überschneidung geändert werden kann. Durch eine zusätzliche farbliche Hervorhebung der Überschneidung wird den Schüler:innen somit deutlich die Bedeutung der Quartile beim Vergleich von Messergebnissen in Form von Boxplots vermittelt. Zuvor können die Schüler:innen bei der Erstellung des Boxplots Bestandteile in ihrem individuellen Tempo erstellen und gegebenenfalls zusätzliche

Elemente anzeigen lassen, die relevante mathematische Inhalte wiederholen. Auf diese Weise können die Lernenden selbstreguliert den Boxplot als Darstellung der Messergebnisse verstehen.

Bei der Bestimmung von Unsicherheiten kann grundsätzlich in Typ A und Typ B Unsicherheiten unterschieden werden. Im Rahmen der App (A2) werden die beiden Konzepte erarbeitet, indem die Schüler:innen zwei virtuelle Versuche durchführen. Die Schüler:innen können somit gleichzeitig die Eigenschaften der Unsicherheitsarten wie die Abhängigkeit verwendeter Geräte verstehen und diese anschließend bestimmen. Das virtuelle Messen bietet die Möglichkeit, das Experiment auf die für die Vermittlung des Konzepts relevanten Elemente zu reduzieren und die benötigte Zeit sowie eventuelle Vorbereitungen zu reduzieren.

Ein weiterer zentraler Bestandteil der Lernapps soll die Verwendung von Simulationen sein. In der Lernapp A1 wid eine Simulation verwendet, um den Schüler:innen die zufällige Streuung von Messwerten zu visualisieren und somit erfahrbar zu machen. Darauf aufbauend kann die Sinnhaftigkeit eines vollständiges Messergebnisses aus einem Maß der zentralen Tendenz sowie einer exemplarischen Typ A Unsicherheit vermittelt werden (A1). Die Simulation kann durch die Schüler:innen beliebig oft wiederholt und variiert werden.

Feedback bezüglich der Schüler:innenantworten

Im Rahmen der Bearbeitung der Lernapps werden die Lernenden an verschiedenen Stellen mit Eingaben oder Aufgaben konfrontiert. Die Eingaben der Schüler:innen können unmittelbar auf Korrektheit überprüft werden. Die Lernenden erhalten daraufhin eine entsprechende Rückmeldung, die basierend auf der Antwort der Lernenden erfolgt. Im Rahmen der Aufgabenkonstruktion wurden bei den Distraktoren Präkonzepte der Lernenden im Umgang mit Daten und Unsicherheiten (vgl. Anhang A) berücksichtigt. Auf diese Weise können den Lernenden nach Eingabe der Antwort Hinweise gegeben werden, die die Schüler:innen hinsichtlich eines Konzeptwechsels von einem fachlich defizitären Präkonzept zu einem fachlich korrekten Konzept unterstützen sollen. Ferner wird auf diese Weise sichergestellt, dass die Schüler:innen in Erarbeitungsphasen neuer Konzepte (s. Anhang A) interaktiv mitwirken können, ohne jedoch fachlich falsche Inhalte zu vermitteln.

Neben dem Feedback an Schüler:innen im Anschluss an getätigte Eingaben in den Lernapps werden den Lernenden sowohl im Rahmen der Erarbeitung von Prototypen als auch bei der Anwendung des Konzepts auf neue Situationen unterstützende Elemente angeboten. Einerseits können fachliche Inhalte wie Maße der zentralen Tendenz oder Streumaße, die zur Erarbeitung des von der App adressierten metrologischen Konzepts benötigt werden, auf Wunsch von Lernenden ein- und ausgeblendet werden. Andererseits können Lernende Unterstützung beim Lösen der Aufgaben erhalten, falls die Aufgabe nicht vollständig verstanden wurde oder ein Lösungsansatz benötigt wird.

Zusammenfassung

Durch die Implementation der beschriebenen digitalen Elemente beinhalten die Lernapps interaktive digitale Tools, welche im SAMR-Modell (Puentedura, 2006) den Stufen *Redefinition* und *Modification* zugeordnet werden können. Die sinnvolle und didaktisch motivierte Nutzung der Vielfalt digitaler Medien ist ein zentrales Ziel bei der Entwicklung der Lernapps. Die Verwendung dieser kann einen positiven Effekt auf den Lernerfolg sowie Anweisungen und Arbeitsaufträge aufweisen (Finkelstein et al., 2005; Hillmayr et al., 2020; Rutten et al., 2012). Zudem begünstigt der Einsatz von Computersimulationen durch Effizienz und Ökonomie den Interventionscharakter der Lernapps (Finkelstein et al., 2005).

Zusammenfassend ist es das Ziel, Lernapps zum Umgang mit Daten und Unsicherheiten zu entwickeln, die von Lernenden selbstreguliert bearbeitet werden können: Die Schüler:innen können im individuellen Tempo die Lernapp bearbeiten und gegebenenfalls Unterstützung bei der Erarbeitung von Inhalten oder der Lösung der Aufgaben anfordern. Durch antwortbasiertes Feedback und die unmittelbare Kontrolle der Eingaben erhalten die Schüler:innen an verschiedenen Stellen der Lernapps eine Rückmeldung über ihren Fortschritt. Die genannten Eigenschaften der Lernapps unterstützen eine selbstregulierte und autonome Bearbeitung der Apps durch die Schüler:innen, welche durch die textbasierten Anweisungen und Eingaben begünstigt wird (Tabbers, 2002) und sich positiv auf die Motivation und den Lernerfolg auswirken kann (Deci & Ryan, 1993; Pfost et al., 2020; Ropohl & Scheuermann, 2018; van de Pol et al., 2010).

Abschließend wurden die Apps hinsichtlicher des Layouts und der Bedienung optimiert. Mittels lauten Denkens wurde der Bearbeitungsprozess der Schüler:innen aufgenommen und Schwierigkeiten behoben (s. Kardas (2019) für eine ausführliche Darstellung).

4.2 Prototypische Darstellung der Apps am Beispiel A1

Die Lernapps zum Umgang mit Daten und Unsicherheiten sollen dazu verwendet werden, um bei den Schüler:innen einen Konzeptaufbau zu fördern (vgl. Abschnitt 4.1.1). Für diesen Zweck wurde die Entwicklung der Lernapps an den Basismodellen von Oser und Patry (1994) orientiert (s. Anhang A). Nachfolgend werden der Aufbau und die Gestaltung der Lernapps prototypisch an der App A1 dargestellt.

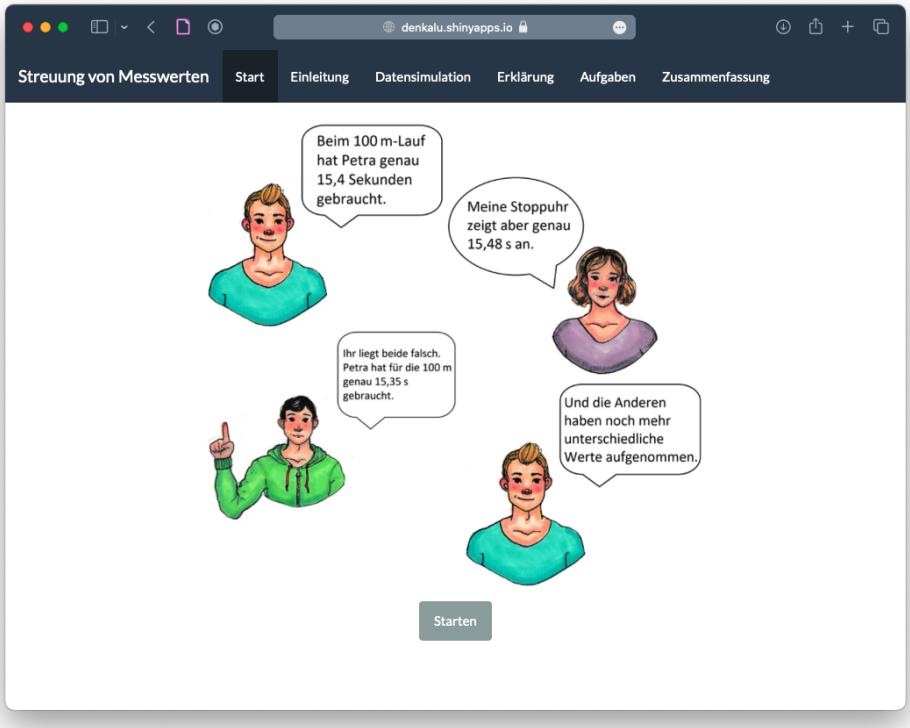


Abb. 4.1: Startseite der Lernapps. Zur Motivation der Thematik wird ein Alltagskontext anhand eines Concept Cartoons verwendet.

Zu Beginn und zum Ende jeder App wird ein Concept Cartoon eingeblendet, welcher die Schüler:innen aus einem alltägliche Kontext heraus für die Inhalte der App motivieren soll (s. Abbildung 4.1). Für die Apps wurden ein 100 m-Lauf, eine Fahrt mit einem E-Scooter oder der Schulweg als Kontext gewählt. Zu Beginn der Lernapp wird über einen kurzen Dialog der Schüler:innen ein Konflikt initiiert, welcher durch die Bearbeitung der App gelöst wird. Der Concept Cartoon am Ende der jeweiligen Lernapp fasst die relevanten Aspekte des neuen Konzepts zusammen und löst anschließend die Konfliktsituation auf.

In Abbildung 4.2 ist die erste Seite der App zu erkennen. Diese ist so aufgebaut, dass den Schüler:innen einerseits die Bedienung der App über die Knöpfe erklärt wird und andererseits für die Bearbeitung der Lernapp nötige Konzepte kurz wiederholt und aufbereitet werden. Im Falle der App A1 (s. Abbildung 4.2 ist keine weitere Darstellung neben den Textanweisungen zu erkennen, da es sich hierbei um das grundlegende Konzept der didaktischen Struktur handelt (s. Kapitel 4.1.1), welches die Existenz von Messunsicherheiten beinhaltet.

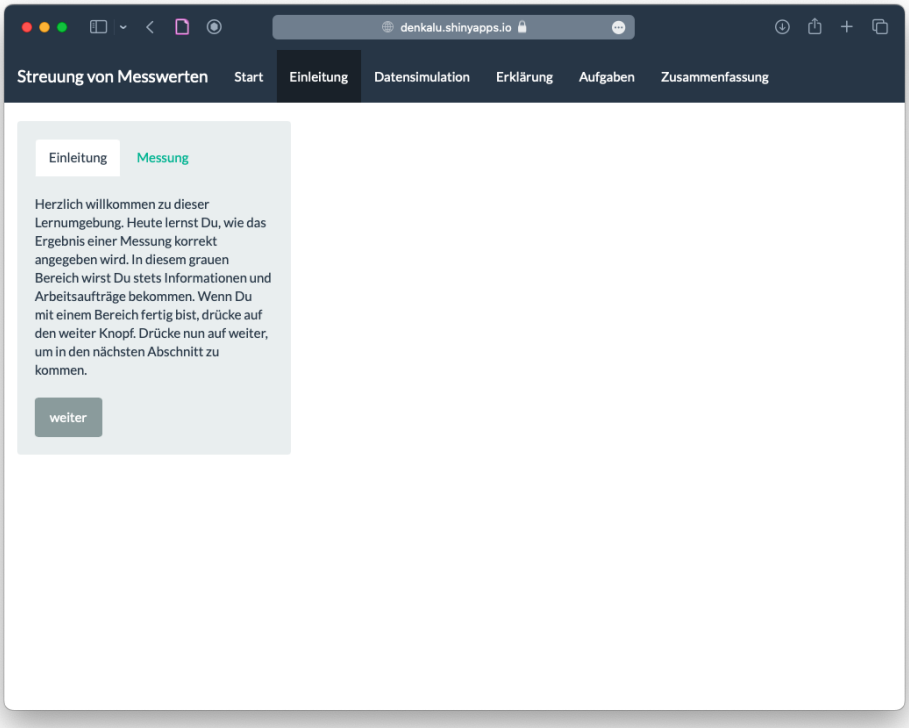


Abb. 4.2: Einleitung der Apps. Auf der linken Seite innerhalb der grauen Bereiche befinden sich Informationstexte und Arbeitsanweisungen.

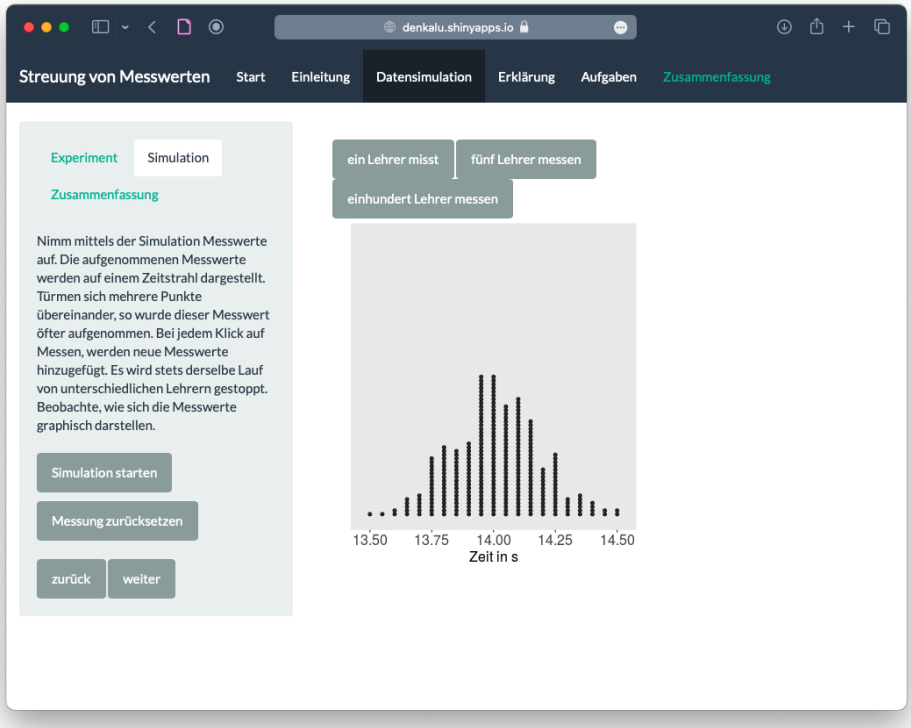


Abb. 4.3: Simulation von Messwerten. Es wird verdeutlicht, dass mit wachsender Anzahl an Messwerten ein Häufungspunkt vorliegt, um den sich die Messwerte symmetrisch streuen.

Für die Sensibilisierung der Schüler:innen bezüglich der Streuung von Messwerten und der Existenz von Messunsicherheiten können die Schüler:innen im Abschnitt Datensimulation eine beliebige Menge von Messwerten generieren. Dabei können die Schüler:innen unterscheiden, ob ein, fünf oder 100 Messwerte dem bisherigen Datensatz hinzugefügt werden sollen. Die Simulation der Messwerte ist zufallsbasiert und kann von den Lernenden jederzeit zurückgesetzt und erneut durchgeführt werden.

Bei den Messwerten handelt es sich um 100-Meter-Lauf-Zeiten, welche von unterschiedlichen Lehrkräften gemessen werden. Den Schüler:innen kann auf diese Weise leicht verdeutlicht werden, wie es zu der Streuung der Messwerte in diesem Fall kommt (s. Abschnitt A für weitere Details). Bei jeder Wiederholung ist zu erkennen, dass es einen mittleren Wert gibt, welcher am häufigsten gemessen wird, und die Werte in beide Richtungen dieses Werts streuen können.

Darauf aufbauend wird das vollständige Messergebnis unter Berücksichtigung des Mittelwerts und der maximalen Abweichung vom Mittelwert als

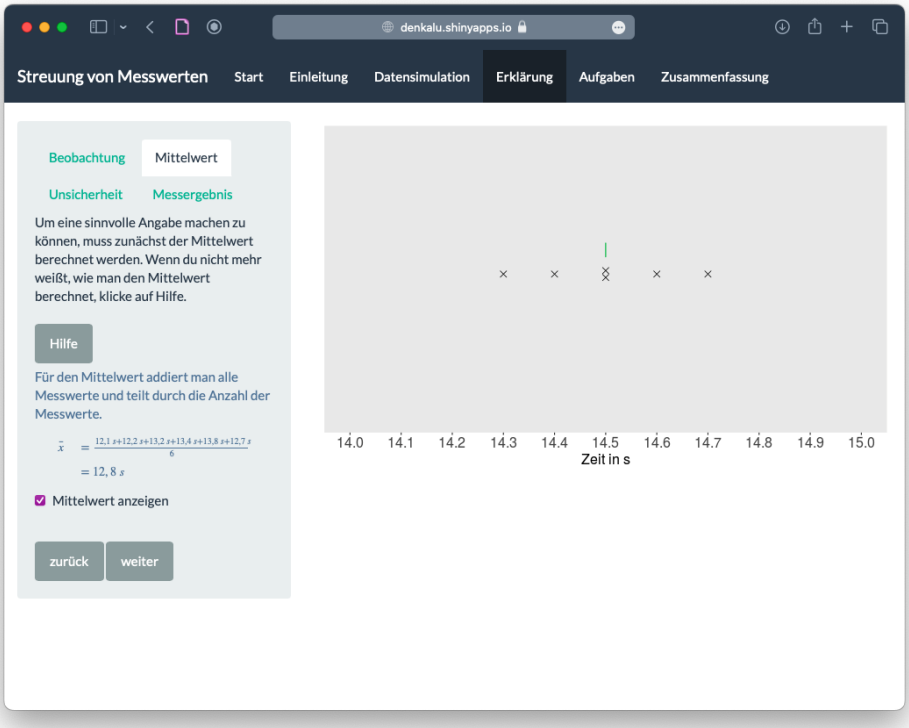


Abb. 4.4: Visualisierung eines vollständigen Messergebnisses, welches sukszessive erstellt wird. Hinweisfelder liefern wesentliche Informationen und Berechnungsvorschriften für die Kenngrößen.

Unsicherheit gemeinsam mit den Schüler:innen ermittelt. Die einzelnen Bestandteile des vollständiges Messergebnisses werden sukszessive in der App dargestellt (s. Abbildung 4.4), wobei die Schüler:innen bei Unkenntnis der statistischen Größen auf die Hilfefelder zurückgreifen können. Dabei werden die rechnerische und grafische Darstellung des Messergebnisses gleichzeitig gezeigt.

Im vorletzten Teil der Lernapp können Schüler:innen das Konzept im Rahmen von verschiedenen Aufgaben wiederholen und auf neue Kontexte anwenden. Dabei variieren die Aufgaben zwischen Rechenaufgaben, Multiple-Choice-Aufgaben, Zeichenaufgaben und Sortieraufgaben. In dieser Phase der Lernapp sollen Schüler:innen das neue Konzept verfestigen und vertiefen. Lernende können ihre Antworten überprüfen lassen und erhalten bei falschen Antworten ein Feedback. Die Abbildung 4.5 zeigt, wie die Überprüfung einer Antwort einer bzw. eines Lernenden funktioniert und eine entsprechende Rückmeldung hinsichtlich der fachlichen Richtigkeit dargestellt wird. Auf diese Weise können die Schüler:innen die Aufgaben selbst kontrollieren.

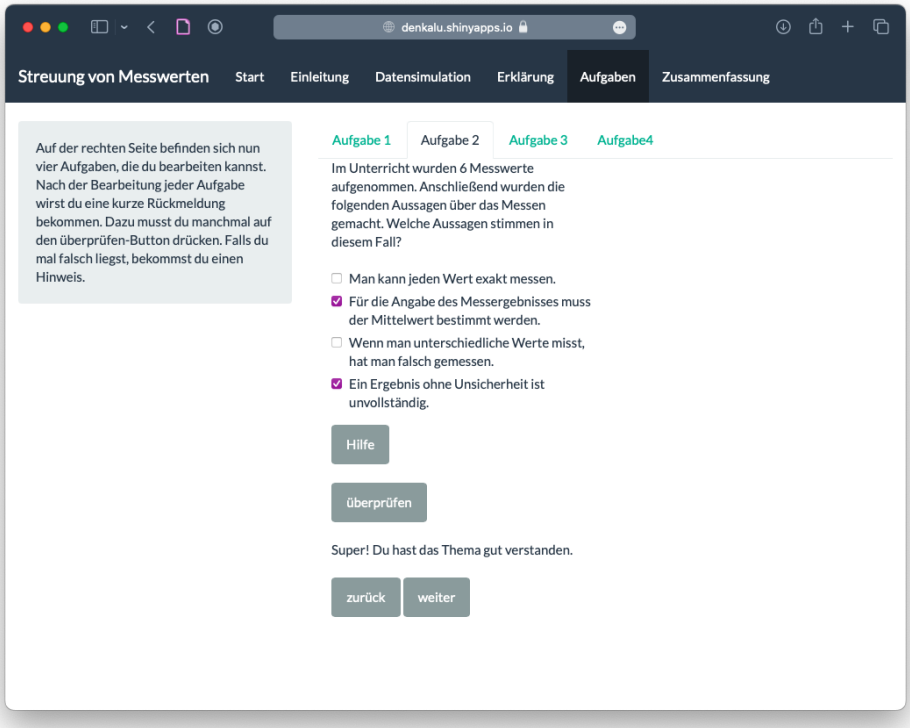


Abb. 4.5: Im Rahmen der Bearbeitung der Aufgaben können Hinweise angefordert und die Lösungen auf Korrektheit überprüft werden.

Die prototypische Darstellung der Lernapp A1 zeigt den grundlegenden Aufbau vieler Apps. Die Apps unterscheiden sich in der Verwendung von interaktiven digitalen Tools (s. Abschnitt 4.1.3) sowie der Aufgabenformate und Kontexte. Eine detaillierte Auflistung der Inhalte der Lernapps sowie ihrer Gestaltung kann Anhang A entnommen werden.

4.3 Zwischenfazit zur Entwicklung der digitalen Lernapps

Im Rahmen der vorliegenden Arbeit sind zehn digitale Lernapps zum Umgang mit Daten und Unsicherheiten entstanden. Bei den Lernapps handelt es sich um ein modernes und didaktisch aufbereitetes Werkzeug, welches durch den Interventionscharakter problemlos in den Unterricht eingebunden werden kann. Durch die Integration der Lernapps ist es möglich, Ergebnisse der Grundlagenforschung in den Unterricht zu integrieren und konsequent den adäquaten Umgang mit Daten und Messunsicherheiten im Physikunterricht zu fördern.

Die Evaluation der entwickelten Apps steht noch aus: Einerseits ist die Wirksamkeit der Apps hinsichtlich des Lernerfolgs noch offen. Im Kapitel 5 wird empirisch untersucht, wie sich die Ausprägung der Datenkompetenz vor und nach der Intervention unterscheidet. Andererseits soll die App im Kontext des Argumentierens evaluiert werden. Dabei soll im Rahmen eines naturwissenschaftlichen Experiments untersucht werden, welchen Einfluss Datenkompetenz auf das Argumentieren der Schüler:innen beim Experimentieren hat. Eine Untersuchung hierzu wird in Abschnitt 6 vorgestellt.

Die digitalen Lernapps liegen aktuell als Open Educational Resources (OER) vor. Die digitalen Lernapps wurden übersetzt und sind somit ebenfalls auf den internationalen Einsatz vorbereitet. Für die Dissemination der Apps wurden Unterrichtsszenarien für den Einsatz skizziert (Kardas & Ludwig, 2021), Workshops angeboten und Fortbildungen durchgeführt, die unter Verwendung der digitalen Lernapps die Relevanz eines adäquaten Umgangs mit Daten und Unsicherheiten im naturwissenschaftlichen Unterricht motivieren (Altenkirch, 2023).

Kapitel 5

Evaluation der Apps hinsichtlich der Wirksamkeit

5.1 Methode

5.1.1 Beschreibung der Studie

Das Ziel der Intervention ist es, die Wirksamkeit der Apps zu evaluieren. Dazu wird überprüft, inwiefern die jeweilige von den Apps adressierte Datenkompetenz gefördert werden kann. Dabei wurde für jeden Kompetenzbereich der didaktischen Strukturierung (vgl. Abschnitt 4.1) eine separate Studie durchgeführt. Es ist zu erwähnen, dass nicht jede entwickelte App evaluiert wird, sondern stellvertretend jeweils die grundlegende App der Kompetenzbereiche Direktes Messen (A2), Indirektes Messen (B1), Grafische Auswertung (C1) und Signifikanz (D1).

Diese Entscheidung wurde aus forschungsökonomischen Gründen getroffen: Alle Apps wurden auf Grundlage derselben Kriterien (vgl. Abschnitt 3) und didaktischen Strukturierungen entwickelt. Der Aufwand, jede App empirisch zu evaluieren, steht nicht im Verhältnis zum Erkenntnisgewinn, da aufgrund der genannten Gemeinsamkeiten im Rahmen der Entwicklung eine ähnliche Wirkweise naheliegend ist. Daher wurde entschieden, aus jedem Kompetenzbereich der didaktischen Strukturierung 4.1.1 eine digitale Lernapp exemplarisch empirisch zu evaluieren.,

Für die Bearbeitung der Apps B1, C1 und D1 sind Grundkompetenzen im Umgang mit Daten notwendig. Da per se nicht davon auszugehen ist, dass diese vorhanden sind, und um Konfundation auszuschließen wurde entschieden, dass die Schüler:innen in jeder Studie zusätzlich vorab die App A1 bearbeiten.

5.1.2 Operationalisierung der Kompetenz im Umgang mit Daten und Messunsicherheiten

Für die Erfassung der Datenkompetenz wird auf die Arbeit von Schulz (2022) zurückgegriffen. Schulz (2022) hat im Rahmen seiner Arbeit Kompetenzen zu den Konzepten aus dem Sachstrukturmodell von Hellwig (2012) entwickelt und darauf aufbauend ein inhaltlich validiertes Testinstrument zur Erfassung der Ausprägung ebendieser Kompetenzen entwickelt.

Zu den zehn Konzepten aus dem Sachstrukturmodell zum Umgang mit Messunsicherheiten hat Schulz (2022) 26 Kompetenzen für Schüler:innen formuliert, zu denen wiederum insgesamt 145 Items entwickelt wurden. Das Testinstrument von Schulz (2022) besteht aus Items im Multiple-Choice-Format, welche entweder im Single-Select- oder Multiple-Select-Format beantwortet werden müssen(Schulz, 2022, S. 35). Im ersten Fall wird das Item als korrekt gewertet, wenn die richtige Aussage ausgewählt wird. Im zweiten Fall müssen Schüler:innen mehrere Teilfragen innerhalb eines Items bearbeiten. Die Lernenden müssen in diesem Fall bei jeder Teilfrage bzw. der dazugehörigen Aussage die Richtigkeit ebendieser bestimmen. Das Item wird nur als korrekt gewertet, wenn alle Teilfragen richtig beantwortet bzw. Teilaussagen richtig bewertet worden sind. Eine Übersicht über die verwendeten Testinstrumente ist in Anhang C und in Tabelle 5.1 enthalten (s. Lauer und Lenz (2021) oder und Abschnitt B.1 für eine ausführliche Diskussion).

Die Validität des Instruments wurde anhand eines Ratings mit sechs Expert:innen beurteilt. Dabei konnte eine sehr gute bis perfekte Übereinstimmung in der Zuordnung der Items zu den formulierten Kompetenzen ermittelt werden ($\kappa_{\text{Fleiss}} \approx 0{,}86$, $\kappa_{\text{Conger}} \approx 0{,}88$, Schulz, 2022, S. 65). Das Forschungsinstrument wurde von Schulz (2022) unter Verwendung der Item-Response-Theorie ausgewertet. Für die in der vorliegenden Arbeit verwendeten Items (s. Abschnitt B.1) konnte dabei eine EAP-Reliabilität zwischen 0,61 und 0,80 sowie eine WLE-Reliabilität zwischen 0,57 und 0,71 bestimmt werden. Damit kann von einer vertretbaren bis guten Reliabilität für das Testinstrument ausgegangen werden (s. Abschnitt E.2 und Schulz, 2022, S. 78 f.).

Vor diesem Hintergrund eignet sich das Testinstrument von Schulz (2022) besonders für die vorliegende Studie: Die Kompetenzen der didaktischen Strukturierung (s. Abschnitt 4.1.1) können eindeutig dem Kompetenzschema von Schulz (2022) zugeordnet werden, wodurch das Testinstrument von Schulz (2022) eine besonders gute Passung zur Evaluation der Wirksamkeit der Lernapps darstellt.

Hinsichtlich der Zielgruppen liegt eine Diskrepanz zwischen den Items von Schulz (2022) (Sekundarstufe II) und den Apps (Sekundarstufe I) vor. Aus diesem Grund wurden die Items zunächst überarbeitet und pilotiert. Auf Grundlage der Ergebnisse mit $n = 120$ Schüler:innen wurde aber deutlich, dass die Items nicht

[712] Bragg-Reflexion

Bei einem Experiment untersucht Nora die Bragg-Reflexion an einem Salzkristall. Beim ersten Durchlauf erhält sie einen Winkle von 6,7°. Da der Versuch laut Experimentieranweisung sechsmal zu wiederholen ist, startet Nora den Versuch erneut und erhält folgende Messwerte:

Messung	Winkel in °
1	6,7
2	6,0
3	7,2
4	7,0
5	5,9
6	7,2

Der aus der Literatur entnommene Referenzwert liegt bei 6,7°. Wie soll Nora vorgehen, um das Ergebnis und die Messunsicherheit anzugeben?

- ☐ Bei der Messung ist die Messunsicherheit 0°, da der erste Messwert mit dem Referenzwert übereinstimmt. Die weiteren Werte sollte Nora vernachlässigen
- ☐ Sie soll den Wert 7,2°, der doppelt vorkommt, als Ergebniswert nehmen und als Messunsicherheit die Differenz zum Referenzwert angeben.
- ☐ Sie soll drei Werte nehmen, die dem Vergleichswert am nächsten liegen, und den Mittelwert bilden. Die Messunsicherheit ist dann der Abstand des kleinsten Werts zum Mittelwert.
- ☐ Sie soll das arithmetische Mittel aus allen sechs Werten bilden und die Standardabweichung für die Messunsicherheit bestimmen

Kraftmesser

Nora zieht beim Experimentieren einen Wagen über den Tisch. Sie misst dabei die Kraft, die sie aufbringen muss. Sie erhält folgende Messwerte

Messung	Kraft in Newton
1	6,7
2	6,0
3	7,2
4	7,0
5	5,9
6	7,2

Ihr bester Freund Paul hat nur eine Messung durchgeführt und erhält 6,7 Newton.

Wie soll Nora vorgehen, um das Ergebnis und die Messunsicherheit anzugeben?

- ☐ Bei der Messung ist die Messunsicherheit **0 Newton**, da der erste Messwert mit dem Wert von Paul übereinstimmt. Die weiteren Werte sollte Nora vernachlässigen.
- ☐ Sie soll den Wert 7,2 Newton, der **zweimal** vorkommt, als Ergebniswert nehmen und als Messunsicherheit die **Differenz** zu Pauls Wert angeben.
- ☐ Sie soll drei Werte nehmen, die Pauls Werten **am nächsten liegen**. Aus diesen Werten sollte sie den **Mittelwert** bilden. Die Messunsicherheit ist dann der Abstand des kleinsten Wertes zum Mittelwert.
- ☐ Sie soll das arithmetische Mittel aus **allen sechs** Werten bilden und den **Maximalfehler** als Messunsicherheit verwenden.

Abb. 5.1: Gegenüberstellung des ursprünglichen Items von Schulz (2022) (links) und der überarbeiteten Fassung (rechts). Das Item wurde auf verschiedenen Ebenen überarbeitet, die anhand der Farbe differenziert werden können. Rot: Änderungen im Kontext, Gelb: Überarbeitung hinsichtlich des Verständnisses, Grün: Hervorhebung von Signalwörtern, Blau: Überarbeitungen bedingt durch unterschiedliche Konzepte der Unsicherheitsabschätzung.

funktionieren. Es konnte keine vertretbare Reliabilität der Messung festgestellt werden (s. Tabelle B.1). Daraufhin wurden die Items erneut überarbeitet. Die Überarbeitung der Apps wird exemplarisch in Abbildung 5.1 dargestellt.

Überarbeitung der Apps

Zunächst wurden die Items dahingehend überarbeitet, dass die Kontexte Teil des Bildungsplans Physik der Sekundarstufe I (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016) oder des Alltags sind. Ebenso wurden die fach- und bildungssprachlichen Formulierungen angepasst, um so dem Niveau der Zielgruppe dieser Arbeit, 8. und 9. Jahrgangsstufe, Rechnung zu tragen (vgl. Abschnitt B.1). Exemplarisch wird die Überarbeitung von Item 712 von Schulz (2022, S. 298) dargestellt (s. Abbildung 5.1).

Tab. 5.1: Übersicht der adressierten Kompetenzen und verwendeten Items (Schulz, 2022) für die Evaluation der Wirksamkeit der jeweiligen App.

App	Kompetenzen	Items
	Die Schüler:innen ...	
A1	... finden in beschriebenen Experimentalsituationen verschiedene Ursachen von Messunsicherheiten [...].	712
	... können Definitionen von Messunsicherheit und Messabweichung angeben und anwenden.	
A2	... können die Definition von Messunsicherheit und Messabweichung angeben und anwenden.	*
B1	... geben das Messergebnis als Zusammenfassung aus Gesamtunsicherheit und Ergebniswert an.	801, 804, 805, 806, 808, 811, 815
	... bestimmen aus mehreren unsicherheitsbehafteten Größen die Gesamtunsicherheit in Summen, Differenzen, Produkten und Quotienten.	
C1	... stellen eine Unsicherheitsbilanz auf und analyieren sowie bewerten diese hinsichtlich ihrer Vollständigkeit und der Stärke der Einflussfaktoren.	402, 404, 405, 407, 409, 412, 414
	... überprüfen Kriterien für eine lineare Regression und kennen Eigenschaften einer Ausgleichsgeraden.	
	... entscheiden in gegebenen Beispielen, ob eine lineare Regression korrekt durchgeführt wurde.	
D1	... prüfen Messwerte auf ihre Verträglichkeit mit einem Referenzwert, indem sie den Referenzwert in Relation zum Messergebnis und dessen Messunsicherheit setzen.	301, 303, 304, 305, 307, 309, 314, 315
	... prüfen zwei oder mehrere Messwerte auf ihre Verträglichkeit, indem sie die Überlappung der Messunsicherheiten der Messergebnisse untersuchen.	

* s. Lauer und Lenz (2021)

In der ursprünglichen Fassung verwendet das Item ein Experiment zur Bragg-Reflexion. Dieses physikalische Experiment ist nicht Teil des Bildungsplans Physik der Sekundarstufe I (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016) und das Item wurde daher in den Kontext einer Kraftmessung eingebettet. Die Aufgabe behandelt nun die Messung mit einem Federkraftmesser. Zudem wurden Fachbegriffe und -konzepte denen aus den Apps angeglichen, um Irritationen bei den Lernenden zu vermeiden. Eine Übersicht über alle durchgeführte Änderungen kann Abbildung 5.1 entnommen werden. Außerdem wurden Signalwörter hervorgehoben, um das Verständnis für die Schüler:innen zu erleichtern. Die beschriebenen Überarbeitungen wurden entsprechend auf den gesamten Itemkatalog angewendet.

5.1.3 Studiendesign

Im Rahmen von Studie 2 wird die Wirksamkeit der entwickelten Lernapps untersucht (FF1). Es wird untersucht, inwiefern sich die Ausprägung der Datenkompetenz durch die Intervention verändert. Für die Untersuchung wird eine Feldstudie mit longitudinalem Design mit einer nicht-probabilistischen Gelegenheitsstichprobe durchgeführt (Döring & Bortz, 2016, S. 305). Die Wirksamkeit der Apps wird analysiert, indem nach der Prä-Post-Erhebung die Datenkompetenz der Proband:innen mit deskriptiv- und inferenzstatistischen Methoden ausgewertet wird (Döring & Bortz, 2016, S. 658). Da die Intervention aus der Bearbeitung der Apps besteht, kann somit unmittelbar auf die Wirksamkeit der Apps geschlossen werden.

5.1.4 Ablauf der Studie

Alle Studien zur Evaluation der Forschungsfrage FF1 wurden nach dem gleichen Design realisiert. Unterschiede gibt es nur in den verwendeten Tests zur Erfassung der Datenkompetenz und Apps (vgl. Abschnitt 5.1.2).

Zunächst gab es eine Motivation und Einführung für die Lernenden bezüglich der Studie durch den Versuchsleiter. Anschließend begann die Arbeitsphase der Schüler:innen, in der sie den digitalen Fragebogen selbstständig bearbeiten konnten. Die Schüler:innen bearbeiteten den digitalen Fragebogen selbstständig an einem Computer oder Tablet. Daraufhin konnten die Schüler:innen die App bearbeiten und die Arbeitsphase durch ein erneutes Beantworten des Fragebogens zur Erfassung der Datenkompetenz beenden.

5.1.5 Pilotierung

Vor dem Hintergrund, dass die Testinstrumente an Proband:innen der achten bis zehnten Klasse angepasst wurden, wurde eine Pilotierung durchgeführt, die die folgende Aspekte untersuchen sollte: Erstens wurde untersucht, inwiefern das vollständig digitale Format praktikabel für die Durchführenden ist und welchen zeitlichen Umfang die vier Studien benötigen. Zweitens sollten die Testinstrumente insbesondere hinsichtlich der Reliabilität evaluiert werden. Drittens sollte eine erste Wirkung der Intervention bezüglich der adressierten Datenkompetenz im Datenkompetenztest beobachtet werden.

Gemäß den untersuchten Apps werden die dazugehörigen Vorstudien jeweils mit den Buchstaben A bis D bezeichnet. Nachfolgend werden die Ergebnisse der Vorstudie zusammengefasst (s. Lauer und Lenz (2021) für ausführliche Ergebnisse).

Im Rahmen der Pilotierung hat sich gezeigt, dass ein digitales Format, bei dem Durchführende nicht anwesend sind, bei den Lernenden nicht gut angekommen ist. Von insgesamt 124 Proband:innen haben nur 58 die Umfrage vollständig, das heißt Prä- und Post-Erhebung der jeweiligen Datenkompetenz und die Apps, bearbeitet. In Tabelle 5.2 zeigt sich, dass die digitale Durchführung der Pilotierung in eine eher geringe Bearbeitungsdauer mündet: Die durchschnittliche Bearbeitungsdauer der Umfrage betrug je nach App zwischen 16 und 22 Minuten. Damit ist die durchschnittliche Bearbeitungsdauer der Lernenden für die Umfrage geringer als die Zeit, die für die Bearbeitung der Apps im Rahmen der Entwicklung (vgl. Abschnitt 4.1) und Pilotierung der Apps (Donner, 2019) festgestellt werden konnte. Es kann daher davon ausgegangen werden, dass die Lernenden die entsprechenden Phasen der Studie nicht adäquat und motiviert bearbeitet haben.

Die geringe Anzahl an vollständigen Bearbeitungen erschwert eine Evaluation des Testinstruments. Für vereinzelte Gruppierungen von Items konnten zwar akzeptable Reliablitäten ermittelt werden (Lauer & Lenz, 2021), die Ergebnisse und die insgesamt sehr geringe Anzahl an vollständigen Bearbeitungen legen aber nahe, dass die Items erneut überarbeitet werden müssen. Die in der Pilotierung verwendeten Testinstrumente stellen eine Zwischenstufe der Überarbeitung dar. Die in Abschnitt 5.1.2 dargestellten Überarbeitungen stellen den gesamten Prozess dar und beinhalten somit die Überarbeitungen, die nach der Pilotierung realisiert wurden.

Bezüglich der Wirksamkeit der Lernumgebungen können nur erste Tendenzen festgestellt werden. Vereinzelt zeigen die Ergebnisse, dass eine längere Bearbeitungsdauer der entsprechenden App mit einer höheren Punktzahl im Posttest einhergeht.

Tab. 5.2: Übersicht des zeitlichen Aufwands in der Pilotierung. Es zeigt sich eine große Differenz zwischen der Bearbeitung zu den entsprechenden Prä- und Postzeitpunkten. Die restliche Zeit entfällt auf Maskenfragen oder Einführungsvideos, die die Schüler:innen im Rahmen der Pilotierung bearbeitet bzw. betrachtet haben.

Vorstudie	Zeit in Minuten							
	Prätest		Posttest		App		Gesamt	
	μ	σ	μ	σ	μ	σ	μ	σ
A: DM	3,40	4,86	1,04	1,90	4,89	10,92	18,14	14,31
B: IM	6,03	3,14	1,82	1,45	3,35	5,94	16,43	8,41
C: GR	7,34	3,49	2,63	2,21	5,33	8,38	18,97	11,87
D: S	8,49	4,88	3,52	3,30	6,41	7,81	21,89	13,46

In den Pilotierungsstudien wurde ein randomisiertes Design verwendet, um mögliche Einflüsse von Schulen oder Klassen vernachlässigen zu können. Die Schüler:innen wurden bei mit Beginn der Umfrage zufällig einer der vier zu evaluierenden Apps und Umfragen zugewiesen. Dieses Vorhaben wurde in der Hauptstudie aus mehreren Gründen verworfen: Im Rahmen der Studie ist eine synchrone Bearbeitung verschiedener Apps innerhalb derselben Klasse für die verschiedenen Apps praktisch schwer realisierbar. Die Apps benötigen unterschiedliche Einleitungen und unterscheiden sich in der zeitlichen Dauer. Außerdem ist die synchrone Durchführung eng mit der Klassenstruktur verbunden, sodass es insbesondere aus forschungsethischer Sicht und aus Perspektive der Lehrkräfte sinnvoller war, innerhalb einer Klasse dieselbe App zu evaluieren. Damit unterscheidet sich das Vorgehen von der Strategie der Pilotierung. Während die synchrone Realisierung der Studie an die Unterrichtszeit gekoppelt war, fand die Bearbeitung während der Pilotierung unter anderem als Hausaufgabe am Nachmittag stattfand.

5.1.6 Kurzbeschreibung der verwendeten statistischen Verfahren

Bei den Forschungshypothesen H^{1a} bis H^{1d}, die im Rahmen von Studie 2 untersucht werden, handelt es sich um Unterschiedshypothesen (Döring & Bortz, 2016, S. 727). Für die Modellierung der Ausprägung der Kompetenzen im Umgang mit Daten und Unsicherheiten zu den unterschiedlichen Erhebungszeitpunkten werden jeweils Mittelwert und Standardabweichung geschätzt. Die Wirksamkeit der Lernapps wird durch die standardisierte Mittelwertdifferenz der Verteilungen evaluiert.

Für den Vergleich der Prä- und Postresultate, die im Rahmen der Datenkompetenztests ermittelt werden konnten, werden probabilistische Verfahren verwendet. Diese unterscheiden sich von frequentistischen inferenzstatistischen Mitteln darin, dass den Verfahren das Bayesianische Wahrscheinlichkeitsverständnis zugrunde liegt (Kruschke, 2010), und weisen den Vorteil auf, dass Aussagen über die Unsicherheiten der geschätzten Parameter getätigt werden könnten (Kruschke, 2011).

Diese Verfahren finden in der naturwissenschaftsdidaktischen Community gerade erst eine größere Beachtung, daher wird im Folgenden kurz auf die grundlegenden Methoden eingegangen und an geeigneter Stelle auf weiterführende Literatur verwiesen. Auf Grundlage der unter Verwendung der bayesianischen Verfahren geschätzten Parameter wird zur Analyse der Wirksamkeit der Interventionen die Effektstärke Cohens d bestimmt, welches die standardisierte Mittelwertdifferenz der Stichproben darstellt (J. Cohen, 1988; Kruschke, 2013).

Für die Evaluation der Wirksamkeit werden zwei Ansätze verfolgt: Für das Testen der Hypothesen und der zugrunde liegenden Modelle werden die

Bayes-Faktoren berechnet. Der zweite Ansatz verwendet die sogenannte „bayesian parameter estimation“ (Kruschke, 2011, S. 300). In diesem Fall werden inferenzstatistisch die Verteilungen der Parameter, in der vorliegenden Arbeit die Mittelwertdifferenz und deren Standardabweichung, ermittelt. Bei den ermittelten Intervallen handelt es sich um Highest-densitiy-Intervalle (HDI). Diese Intervalle zeichnen sich dadurch aus, dass alle Werte innerhalb des HDI eine größere Glaubwürdigkeit aufweisen als die Werte außerhalb des HDI (Kruschke, 2015, S. 342). Zur Evaluation der HDI werden Region-of-practical-equivalence-Intervalle (ROPE) verwendet. Diese Intervalle stellen Bereiche dar, die aus praktischer Perspektive mit einem Nulleffekt gleichgesetzt werden können. Der Abgleich der HDI und ROPE kann somit verwendet werden, um die praktische Relevanz vorgefundener Effekte zu evaluieren (Kruschke, 2021).

Ein Vorteil des bayesianischen Ansatzes ist es, dass die Wahrscheinlichkeit für die Gültigkeit eines Modells bzw. einer Hypothese H_i unter der Bedingung des vorgefundenen Datensatzes d unter Berücksichtigung des Hypothesenraums $\mathcal{H}$ anhand

$$P(H_i|d) = \frac{P(H_i) \cdot p(d|H_i)}{\sum_{j=1}^{J} P(H_j) \cdot p(d|H_j)} \tag{5.1}$$

beschrieben werden kann (Jackman, 2009, S. 36). Dabei bezeichnet p nicht die Wahrscheinlichkeitsdichtefunktion der zugrunde liegenden Verteilung, sondern in diesem Fall die dazugehörigen Likelihood-Funktionen zu den Parametern d und $H_j \in \mathcal{H}$.

In der vorliegenden Arbeit werden die Nullhypothese $H_0 : \mu_{\text{post}} \leq 0$ und die Alternativhypothese $H_1 : \mu_{\text{post}} > 0$ gegeneinander abgeglichen. Setzt man die Likelihood-Funktionen für H_0 und H_1 aus Gleichung 5.1 ins Verhältnis, ergibt sich der sogenannte Bayes-Faktor (BF) (Gelman et al., 2015, S. 183):

$$\text{BF}_{10} = \frac{p(d|H_1)}{p(d|H_0)} \tag{5.2}$$

Der Index der Bezeichnung BF_{10} verdeutlicht, dass in diesem Fall betrachtet wird, wie plausibel die Hypothese H_1 gegenüber H_0 ist. Anhand des BF kann somit evaluiert werden, welche Hypothese eher zu den Daten passt. Ein Wert über 3 kann als Evidenz für die Alternativhypothese gesehen werden (Jackman, 2009, S. 38).

Die ermittelten BF sind unter Umständen abhängig von den angenommenen *Priors* (Kruschke, 2015, S. 346 f.). Bei den Priors handelt es sich um theoretische Vermutungen bezüglich der Verteilung von Parametern. Bei der Bestimmung der BF wird auf den Vorschlag von Morey und Rouder (2011) zurückgegriffen. Es werden keine uninformierten Priors für die Schätzung von Mittelwertdifferenzen verwendet, da diese bei Modellvergleichen, die auf dem BF beruhen, stets die

Tab. 5.3: Interpretation des BF. Es ist zu erkennen, dass Jeffreys (1998) eine zusätzliche Abstufung *sehr stark* beinhaltet. Abgesehen davon stimmen beide Klassifizierungen überein.

Stärke der Evidenz	Jeffreys (1998)	Kass und Raftery (1995)
vernachlässigbar	10^0 bis $10^{0,5}$	10^0 bis $10^{0,5}$
substantiell	$10^{0,5}$ bis 10^1	$10^{0,5}$ bis 10^1
stark	10^1 bis $10^{1,5}$	10^1 bis 10^2
sehr stark	$10^{1,5}$ bis 10^2	-
entscheidend	$> 10^2$	$> 10^2$

Punktlösungen (in diesem Fall eine Mittelwertdifferenz von null) bevorzugen. Es werden daher folgende Modelle verwendet:

$$\begin{aligned} H_0: &\quad \delta = 0 \\ H_1: &\quad \delta \sim \text{Cauchy}(r) = 1 \end{aligned} \tag{5.3}$$

δ bezeichnet dabei die standardisierte Differenz μ/σ der Differenzpunktzahl aus dem Prä- und Posttestzeitpunkt und stellt somit das Effektstärkemaß für diese Untersuchung dar. Für die Analyse wird $r = \sqrt{2}/2$ für die Cauchy-Verteilung verwendet (Morey & Rouder, 2018). Im Gegensatz dazu wird für die Standardabweichung nach Jeffreys (1946) ein Prior der Form

$$p(\sigma^2) \propto \frac{1}{\sigma^2} \tag{5.4}$$

verwendet. Da dieser Parameter in beiden Modellen aus Gleichung 5.3 enthalten ist, ist in diesem Fall ein uninformierter und diffuser Prior erwünscht (Morey & Rouder, 2011). Der BF eignet sich gut als Evidenz, da er wiedergibt, wie die Überzeugungen der Forscher:innen auf Grundlage der Daten angepasst werden sollten (Kass, 1993).

Die Verwendung des Bayes-Faktors ist jedoch auch mit einigen Herausforderungen verbunden. Die Aussagekraft des Bayes-Faktors ist besonders bei diskreten Verteilungen hoch und weniger bei stetigen Parameterschätzungen (Gelman et al., 2015, S. 183). Aus diesem Grund wird für die Evaluation der Mittelwertdifferenz und Effektstärke zusätzlich eine vollständige Parameterschätzung durchgeführt. Die Ergebnisse beider Verfahren werden verwendet, um Forschungsfrage FF1 auszuwerten, da beide Methoden unterschiedliche Schlüsse erlauben (Kruschke, 2011). Es wird jedoch davon abgesehen, sowohl die Bayes-Faktoren als auch die Parameterschätzung in komplexeren Modellen zu vereinheitlichen, wie einige moderne Ansätze es tun (bspw. Baig, 2021; Kruschke, 2011; Rouder et al., 2018). Diese Entscheidung begründet sich damit, dass für die Modellierung eines Mehr-Ebenen-Modells nicht ausreichend Proband:innen vorliegen und keine adäquaten informierten Prior zu den betrachteten Modellen und Hypothesen vorliegen (Baig, 2021).

Gegenüber frequentistischen Verfahren weist die Methode von Kruschke und Liddell (2018) zwei Vorteile auf. Einerseits wird nicht gegen das Ereignis $\Delta\mu = 0$ getestet, welches in einer stetigen Verteilung die Wahrscheinlichkeit null aufweist. Andererseits stellen die HDI Verteilungen der geschätzten Parameter dar und können auch dahingehend interpretiert werden (Kruschke & Liddell, 2018; Morey et al., 2016).

Das HDI stellt einen Bereich dar, in dem der tatsächliche Parameterwert zu 95 % enthalten ist. Es handelt sich somit um diejenigen Werte, die auf Grundlage der Daten am glaubwürdigsten sind. Damit grenzt sich das HDI von frequentistischen Konfidenzintervallen ab, welche nur Aussagen darüber ermöglichen, ob geschätzte Parameter enthalten sind oder nicht (Döring & Bortz, 2016, S. 645). Die ROPE hingegen definiert ein Intervall, welches aus praktischer Sicht den Nulleffekt definiert (Kruschke, 2018).

Für die Evaluation des Effekts gibt es drei Szenarien: Sind das HDI und die ROPE disjunkt, so ist der vorgefundene Effekt nicht null. Stellt das HDI eine echte Teilmenge des ROPE dar, so kann auf Grundlage der Daten ein Nulleffekt angenommen werden. In allen anderen Fällen kann keine Aussage zum Betrag und der Richtung der Parameter getätigt werden (Kruschke, 2015, S. 336).

Für die Bestimmung der ROPE wurde sich an verschiedenen Konventionen orientiert. Für die Untersuchung der Mittelwertdifferenz schlagen McElreath (2020) und Kruschke (2018) das Intervall $[-0{,}1\,\sigma(y); 0{,}1\,\sigma(y)]$ vor, wobei $\sigma(y)$ die Standardabweichung der Differenzpunktzahl y der Proband:innen aus dem Prä- und Postzeitpunkt bezeichnet. Bei der Effektstärke schlägt Kruschke (2018) vor, ein Intervall von $[-0{,}1; 0{,}1]$ als ROPE zu wählen, da nach J. Cohen (1988) eine Effektstärke $|d| \geq 0{,}2$ als kleiner Effekt zu deuten ist.

Für die Auswertung muss zunächst gewährleistet werden, dass die verwendeten Verfahren konvergieren. Zur Beurteilung der Konvergenz wird sich an den Vorgaben von Vehtari et al. (2021) orientiert, welche im Anhang E.1 detailliert dargestellt werden. Zusammenfassend lässt sich sagen, dass für alle vorliegenden Verfahren Konvergenz vorliegt. Die Resultate können somit beibehalten und interpretiert werden.

5.1.7 Stichprobe

Die Studie wurde mit 151 Schüler:innen der 8. bis 10. Jahrgangsstufe durchgeführt, von denen 133 die Umfrage vollständig bearbeitet haben. Die Schüler:innnen besuchten Gymnasien und Realschulen in Baden-Württemberg. Eine genaue Beschreibung der jeweiligen Stichproben für die vier separaten Studien kann Tabelle 5.4 entnommen werden. Bei den Schulen handelt es sich um sogenannte MINT-freundliche Schulen in Baden-Württemberg (MINT Zukunft e.V., 2022). Schulen erhalten dieses Zertifikat, wenn sie gewisse Kriterien erfül-

Tab. 5.4: Übersicht über die Stichproben der Studien A-D.

	n	Klasse in %			Alter	Geschlecht in %		
		8	9	10		m	w	d
Studie A	34	47,0	53,0	-	13,6	26,5	67,6	5,9
Studie B	28	-	100,0	-	14,8	21,4	75,0	3,6
Studie C	44	-	-	100,0	15,7	40,9	54,5	4,6
Studie D	27	-	77,8	22,2	15,0	37,0	63,0	-

len, die einen naturwissenschaftlichen Schwerpunkt bzw. eine Profilierung der Schule erfordern.

Die Entscheidung für die MINT-freundlichen Schulen wurde aus mehreren Gründen getroffen: Die naturwissenschaftliche Profilierung der Schulen geht einher mit einer Offenheit gegenüber außerschulischen Aktivitäten und einem Interesse an fachdidaktischer Forschung. Dadurch konnte gewährleistet werden, dass die Inhalte der Studie auch langfristig in den Unterricht eingebunden werden können. Diesbezüglich versprach sich der Autor eine höhere Erfolgsrate bei der Proband:innenakquise. Die digitale Durchführung der Studie stellt eine hohe Anforderung an die technische Ausstattung der Schule, welche von den MINT-freundlichen Schulen aufgebracht werden konnte. Zudem konnte bei den Schulen angenommen werden, dass ein vergleichsweise höheres Interesse an naturwissenschaftlichen Fächern und anderen Angeboten besteht. Auf diese Weise sollte sichergestellt werden, dass der Einfluss von mangelndem Interesse auf die Studie verringert und eine geringe Abbrecherquote erreicht werden kann.

Stichprobenumfangsplanung

Bei der Stichprobe handelt es sich um eine nicht-probabilistische Gelegenheitsstichprobe (Döring & Bortz, 2016, S. 305). Die Proband:innen wurden geworben, indem zufällig ausgewählte MINT-freundliche Schulen des Landes Baden-Württemberg angeschrieben wurden. Zum Zeitpunkt der Datenerhebung war eine frequentistische Auswertung der Daten geplant. Für einen t-Test für verbundene Stichprobe ist eine Proband:innenzahl von $n \geq 27$ nötig, um eine Teststärke von $1 - \beta = 0{,}8$ (J. Cohen, 1988) zu einem Signifikanzniveau von $\alpha = 0{,}05$ zu erreichen. Die Schulen wurden sukszessive angeschrieben, bis jede der Studien A-D über die entsprechende Anzahl an Proband:innen verfügte.

5.1.8 Evaluation der Testinstrumente

Die verwendeten Testinstrumente werden hinsichtlich ihrer Schwierigkeit und Reliabilität untersucht, um die Güte des Tests zu bestimmen. Die Gütekriterien Objektivität und Validität der verwendeten Testinstrumente wurden bei Schulz (2022) hinreichend untersucht. Nachfolgend werden die mittlere Lösungshäufigkeit der Items als Indikator für die Schwierigkeit sowie Cronbachs α (Cronbach, 1951) und die Trennschärfe r als Maße für die Reliabilität bestimmt. Es wird davon abgesehen, kongenerische Reliabilitätsmaße zu verwenden, da im weiteren Verlauf Mittelwerte auf Grundlage der Items gebildet werden, welche suggerieren, dass alle Items gleich stark mit dem latenten Konstrukt in Beziehung stehen. Bei den untersuchten Testinstrumenten handelt es sich um vergleichsweise kurze Tests ($n < 10$), so dass bei der Analyse berücksichtigt werden muss, dass bei diesen Testlängen Cronbachs α unterschätzt wird (Briggs & Cheek, 1986). Eine Zusammenfassung der Itemkennwerte kann Tabelle 5.5 entnommen werden. Aufgrund der geringen Testlänge wurden einzelne Items nur ausgeschlossen, wenn sie nahezu keine oder eine negative Korrelation zum Konstrukt aufgewiesen haben.

Direktes Messen Das Testinstrument zu diesem Konzept weist mit einer Schwierigkeit von $\bar{x} = 0{,}41$ eine mittlere Schwierigkeit auf. Einzelne Items weichen davon ab, indem sie entweder deutlich leichter (DM21: $\bar{x} = 0{,}71$) oder deutlich schwerer sind (DM24: $\bar{x} = 0{,}24$; DM26: $\bar{x} = 0{,}15$). Die restlichen Items weichen nicht vom Schwierigkeitsgrad des gesamten Instruments ab, sodass Decken- oder Bodeneffekten ausgeschlossen werden können.

Für die interne Konsistenz konnte ein Cronbachs α von $\alpha = 0{,}60$ ermittelt werden, welches im Rahmen von Kompetenztests akzeptabel ist (Taber, 2018). Die Trennschärfe der Items liegt im Durchschnitt bei $r = 0{,}53$ mit einem Minimum bei $r = 0{,}28$ (DM26) und eine Maximum bei $r = 0{,}75$ (DM27). Das Item DM25 wurde für die Skalenbildung nicht berücksichtigt, da es negativ mit der Gesamtpunktzahl korreliert.

Indirektes Messen Im direkten Vergleich zum vorigen Test zeigt sich, dass die Items für den Kompetenzbereich Indirektes Messen im Durchschnitt etwas leichter sind ($\bar{x} = 0{,}67$). Bis auf die Items I26 und I27, die vergleichsweise schwerer ausfallen ($\bar{x} = 0{,}50$ bzw. $\bar{x} = 0{,}49$), weisen alle anderen Items eine ähnliche Schwierigkeit im Bereich von $\bar{x} = 0{,}64$ bis $\bar{x} = 0{,}71$ auf.

Für die Skala konnte ein Cronbachs Alpha von $\alpha = 0{,}82$ ermittelt werden, sodass von einer hohen interne Konsistenz der Skala ausgegangen werden kann. Die minimale Trennschärfe der Items liegt bei $r = 0{,}55$ und steigt bis $r = 0{,}82$ an. Es kann somit davon ausgegangen werden, dass alle Items stark mit der

Skalenpunktzahl korrelieren. Für die weiteren Analysen wurde im vorliegenden Testinstrument kein Item ausgeschlossen.

Grafische Auswertung Die mittlere Schwierigkeit des Testinstruments für die von der App C1 adressierte Datenkompetenz liegt bei $\bar{x} = 0{,}73$ und ist damit eher als mittelleichter Test zu bewerten. Innerhalb des Tests gibt es Schwankungen zwischen den Items. Vier Items sind mit Schwierigkeiten von $\bar{x} = 0{,}82$ bis $\bar{x} = 0{,}91$ deutlich leichter (D21, G22, G23, G25 & G27). Die vier Items G21, G24, G26 und G28 erweisen sich hingegen als deutlich schwerer ($\bar{x} = 0{,}39$ bis $\bar{x} = 0{,}57$).

Alle Items weisen mit Werten von $r > 0{,}3$ eine hohe Trennschärfe auf und können somit für die Bestimmung der Punktzahl verwendet werden. Die interne Konsistenz des Tests kann mit $\alpha = 0{,}63$ als zufriedenstellend eingestuft werden (Taber, 2018).

Signifikanz Der Schwierigkeitsgrad des Testinstruments für die Erfassung der Datenkompetenz im Kompetenzbereich Signifikanz weist einen durchschnittlichen Schwiergikeitsgrad von $\bar{x} = 0{,}80$ auf und ist damit im Vergleich der verwendeten Testinstrumente am einfachsten. Auffällig ist in diesem Fall, dass es erneut viele Items gibt, die leichter (S22, S23 & S25, $\bar{x} = 0{,}59$ bis $\bar{x} = 0{,}63$) oder schwerer (D21, S21, S24, S26 & S28, $\bar{x} = 0{,}89$ bis $\bar{x} = 0{,}93$) als der durchschnittliche Schwierigkeitsgrad sind.

Die Interitemkorrelation beträgt mindestens $r = 0{,}39$ und liegt im Durchschnitt bei $r = 0{,}60$. Der Test weist somit eine hohe Trennschärfe auf. Die interne Konsistenz beträgt ein Cronbachs $\alpha = 0{,}76$.

Alle verwendeten Testinstrumente verfügen somit über eine hohe Trennschärfe und eine ausreichende bis gute interne Konsistenz. Es wird daher davon ausgegangen, dass die Datenkompetenz anhand der Testinstrumente reliabel erfasst werden kann.

5.2 Ergebnisse

Die Ergebnisse der in Abschnitt 5.1.6 beschriebenen Verfahren werden nacheinander berichtet und anschließend im Abschnitt 5.3 gemeinsam diskutiert. Zunächst wird darauf eingegangen, welches Modell aus Gleichung 5.3 auf Grundlage der Daten plausibler erscheint und folglich angenommen wird. Zusätzlich wird durch die konkrete Schätzung der Parameter und der daraus resultierenden Verteilungen evaluiert, inwiefern die Ergebnisse der ersten Analyse eine praktische Relevanz aufweisen.

Tab. 5.5: Darstellung der Itemkennwerte für die verwendeten Testinstrumente in Studie 1.

	Items									
Direktes Messen	D21	DM21	DM22	DM23	DM24	DM26	DM27			$\bar{x}$
Schwierigkeit	0,71	0,50	0,44	0,50	0,24	0,15	0,35			0,41
Trennschärfe	0,61	0,69	0,52	0,40	0,44	0,28	0,75			0,53
Cohens α										0,60 [0,40; 0,80]
Indirektes Messen	D21	I21	I22	I23	I24	I25	I26	I27		$\bar{x}$
Schwierigkeit	0,68	0,64	0,68	0,71	0,71	0,64	0,50	0,49		0,67
Trennschärfe	0,61	0,69	0,82	0,55	0,71	0,78	0,59	0,60		0,67
Cohens α										0,82 [0,73; 0,92]
Grafische Auswertung	D21	G21	G22	G23	G24	G25	G26	G27	G28	$\bar{x}$
Schwierigkeit	0,91	0,57	0,86	0,82	0,59	0,91	0,61	0,91	0,43	0,73
Trennschärfe	0,46	0,30	0,64	0,60	0,51	0,72	0,49	0,55	0,46	0,53
Cohens α										0,63 [0,46; 0,80]
Signifikanz	D21	S21	S22	S23	S24	S25	S26	S27	S28	$\bar{x}$
Schwierigkeit	0,91	0,57	0,86	0,82	0,59	0,91	0,61	0,91	0,43	0,80
Trennschärfe	0,49	0,81	0,67	0,58	0,74	0,61	0,39	0,55	0,53	0,60
Cohens α										0,76 [0,63; 0,89]

5.2.1 Modellevaluation

Im Gegensatz zu klassischen Signifikanztestungsverfahren erlaubt die Bestimmung der BF, sowohl Evidenz für die Nullhypothese H_0 als auch für die Alternativhypothese H_1 zu finden. Die Evaluation der Ergebnisse für BF wird anhand der Vorgaben von Kass und Raftery (1995) sowie Jeffreys (1998) durchgeführt (s. Tabelle 5.3).

Auf Grundlage der verschiedenen Kompetenztests für die Apps A2, B1, C1 und D1 konnten Bayes-Faktoren von $\mathrm{BF}_{10}^{A} \approx 10{,}9$; $\mathrm{BF}_{10}^{B} \approx 166{,}7$; $\mathrm{BF}_{10}^{C} \approx 23{,}2$ und $\mathrm{BF}_{10}^{D} \approx 4537{,}4$ ermittelt werden. Hinsichtlich der Wirksamkeit der Apps A2 und C1 findet sich somit starke Evidenz für die Alternativhypothese: Es ist daher davon auszugehen, dass die Mittelwerte des Prä- und Posterhebungszeitpunkts nicht gleich sind. Die Bayes-Faktoren BF^{B} und BF^{D} können nach Tabelle 5.3 als entscheidende Evidenz für einen Mittelwertunterschied gesehen werden. Die vorangehende Analyse hat verdeutlicht, dass in allen vier Fällen die Nullhypothese H_0 abgelehnt werden kann, da starke bis entscheidende Evidenz für die Alternativhypothese vorliegt. Im weiteren Verlauf wird nun untersucht, welche Parameterschätzungen auf Grundlage dieser Ergebnisse am plausibelsten sind. Dabei wird die Mittelwertdifferenz explizit und standardisiert in Form eines Effektstärkemaßes untersucht.

5.2.2 Parameterschätzung

Im Rahmen des in Abschnitt 5.1.6 beschriebenen Verfahrens werden die Mittelwertdifferenz $\Delta\mu$ und die Standardabweichung σ^2 der Differenzpunktzahl aus dem Prä- und Posttest geschätzt. Um die praktische Relevanz der Ergebnisse zu evaluieren, werden die 95 % HDI mit den ROPE-Intervallen verglichen. Die Analyse wurde für die Datenkompetenztests der Apps A2, B1, C1 und D1 durchgeführt.

Für den Datenkompetenztest zum Kompetenzbereich Direktes Messen zeigt sich, dass die Mittelwertdifferenz $\Delta\mu$ im 95 % HDI echt von null verschieden ist. Ein Abgleich mit dem ROPE-Intervall zeigt, dass es keine Überschneidung gibt. Es kann somit davon ausgegangen werden, dass die Mittelwertdifferenz auch eine praktische Relevanz aufweist. Abbildung 5.2a veranschaulicht die Wahrscheinlichkeitsdichte des geschätzten Parameters. Es ist zu erkennen, dass rund die Hälfte der Dichte oberhalb des Werts $\frac{1}{7}$ liegt. Somit lässt sich folgern, dass die Schüler:innen im Posttest mit hoher Wahrscheinlichkeit im Durchschnitt mehr als eine Frage korrekt beantwortet haben. Für die standardisierte Mittelwertdifferenz wurde ein Wert von $d \approx 0{,}55$ ermittelt. Damit liegt ein mittlerer

Tab. 5.6: Übersicht über die Ergebnisse der vier Studien. Dargestellt sind die HDIs und ROPEs für die Differenz der Mittelwerte und die Effektstärke.

	$\overline{x}$		HDI	ROPE	
	Mittelwert	Median	95 % HDI	ROPE	% in ROPE
Direktes Messen					
$\Delta\mu$	0,14	0,14	[0,05; 0,22]	[-0,02; 0,02]	0,76
Cohens d	0,55	0,55	[0,16; 0,92]	[-0,10; 0,10]	0,88
Indirektes Messen					
$\Delta\mu$	0,18	0,18	[0,10; 0,27]	[-0,02; 0,02]	0,01
Cohens d	0,86	0,85	[0,39; 1,35]	[-0,10; 0,10]	0,03
Grafische Auswertung					
$\Delta\mu$	0,08	0,08	[0,03; 0,12]	[-0,02; 0,02]	0,49
Cohens d	0,56	0,56	[0,21; 0,92]	[-0,10; 0,10]	0,45
Signifikanz					
$\Delta\mu$	0,26	0,26	[0,17; 0,36]	[-0,02; 0,02]	0,00
Cohens d	1,15	1,14	[0,59; 1,68]	[-0,10; 0,10]	0,00

positiver Effekt vor (J. Cohen, 1988). Die Analyse des 95 % HDI und ROPE zeigt, dass die Effektstärke einerseits echt von null verschieden ist und andererseits außerhalb der ROPE liegt und somit eine praktische Relevanz aufweist. Die Posteriorverteilung des Parameters kann Abbildung 5.2b entnommen werden. Es ist zu erkennen, dass im Rahmen der Unsicherheiten mindestens von einem kleinen Effekt ausgegangen werden kann.

Die Auswertung des Datenkompetenztests für die App B1 zeigt, dass der Mittelwert der ermittelten Mittelwertdifferenz bei $\Delta\mu = 0{,}18$ liegt. Da der Datenkompetenztest aus acht Items besteht, bedeutet dies, dass Schüler:innen im Durchschnitt mehr als eine Frage im Posttest zusätzlich korrekt beantwortet haben ($\Delta\mu = 0{,}18 > \frac{1}{8}$). Abbildung 5.2c zeigt, dass der Nullwert nahezu vollständig aus der Verteilung der Mittelwertdifferenz ausgeschlossen werden kann. Ebenso zeigt sich, dass das 95 % HDI und die ROPE sich nicht überschneiden. Somit weist die ermittelte Differenz eine praktische Relevanz auf. Bezüglich der Effektstärke bestätigt sich dieser Eindruck. Der Mittelwert der ermittelten Verteilung für die standardisierte Mittelwertdifferenz liegt bei $d \approx 0{,}86$ und kann somit als großer positiver Effekt kategorisiert werden (J. Cohen, 1988). Das 95 % HDI und die ROPE sind disjunkt, somit ist die ermittelte Effektstärke auch praktisch substanziell. In Abbildung 5.2d wird deutlich, dass die Verteilung stark von der null abweicht. Es kann davon ausgegangen werden, dass mit einer sehr hohen Wahrscheinlichkeit ein mittlerer Effekt und weiterhin mit über 50 % ein großer Effekt vorliegt.

Der Datenkompetenztest für den Kompetenzbereich Grafische Regression besteht aus neun Items. Eine Verbesserung um ein Item spiegelt sich somit in einem Zuwachs von $\frac{1}{9}$ im Mittelwert wider. Für die Mittelwertdifferenz wurde ein Mittelwert von $\Delta\mu = 0{,}08$ ermittelt. Das 95 % HDI ist von null verschieden und überschneidet sich nicht mit der ROPE. Der gefundene Unterschied kann somit als praktisch relevant angesehen werden. Im Vergleich zu den bisherigen Ergebnissen zeigt sich, dass mit einer Wahrscheinlichkeit von über 50 % weniger als ein Item mehr im Posttest korrekt beantwortet worden sind. Die standardisierte Differenz lässt mit einem Mittelwert von $d \approx 0{,}56$ auf einen moderaten positiven Effekt schließen, der praktische Relevanz aufweist (vgl. Tabelle 5.6). Die Verteilung des Parameters in Abbildung 5.3b zeigt eine hohe Dichte um den Bereich eines moderaten Effekts. Zudem ist zu erkennen, dass mit einer Wahrscheinlichkeit von 97 % mindestens von einem kleinen positiven Effekt ausgegangen werden kann.

Abschließend wird der Datenkompetenztest für den Kompetenzbereich Signifikanz evaluiert. Dieser Test weist ebenfalls neun Items auf. Eine Erhöhung der Punktzahl im Posttest um $\frac{1}{9}$ bedeutet, dass im Durchschnitt im Posttest ein Item zusätzlich korrekt gelöst wurde. Auf Grundlage der Analyse konnte für die Verteilung der Mittelwertdifferenz $\Delta\mu$ ein Mittelwert von 0,26 ermittelt werden. Dies entspricht einem Zuwachs von über zwei Items im Posttest. Es gibt

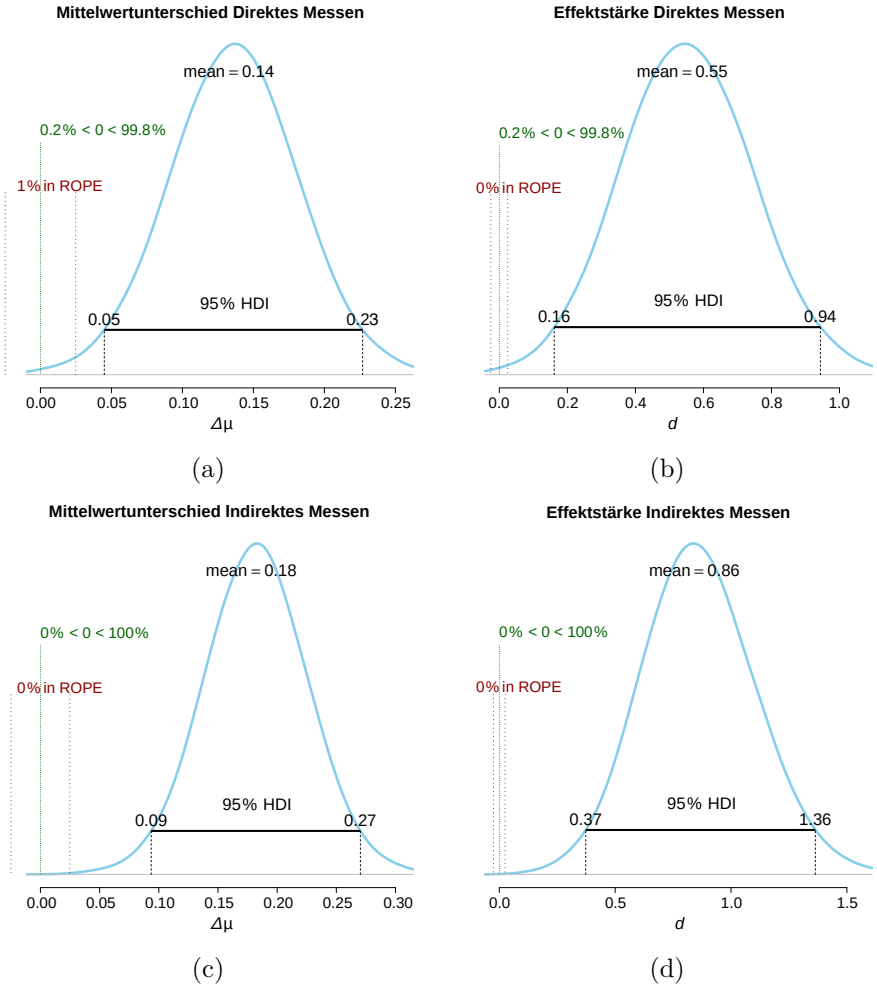


Abb. 5.2: Darstellung der Parameterschätzung für die Mittelwertdifferenz und die Effektstärke für die Apps bezüglich des Direkten Messens (a + b) und Indirekten Messens (c + d).

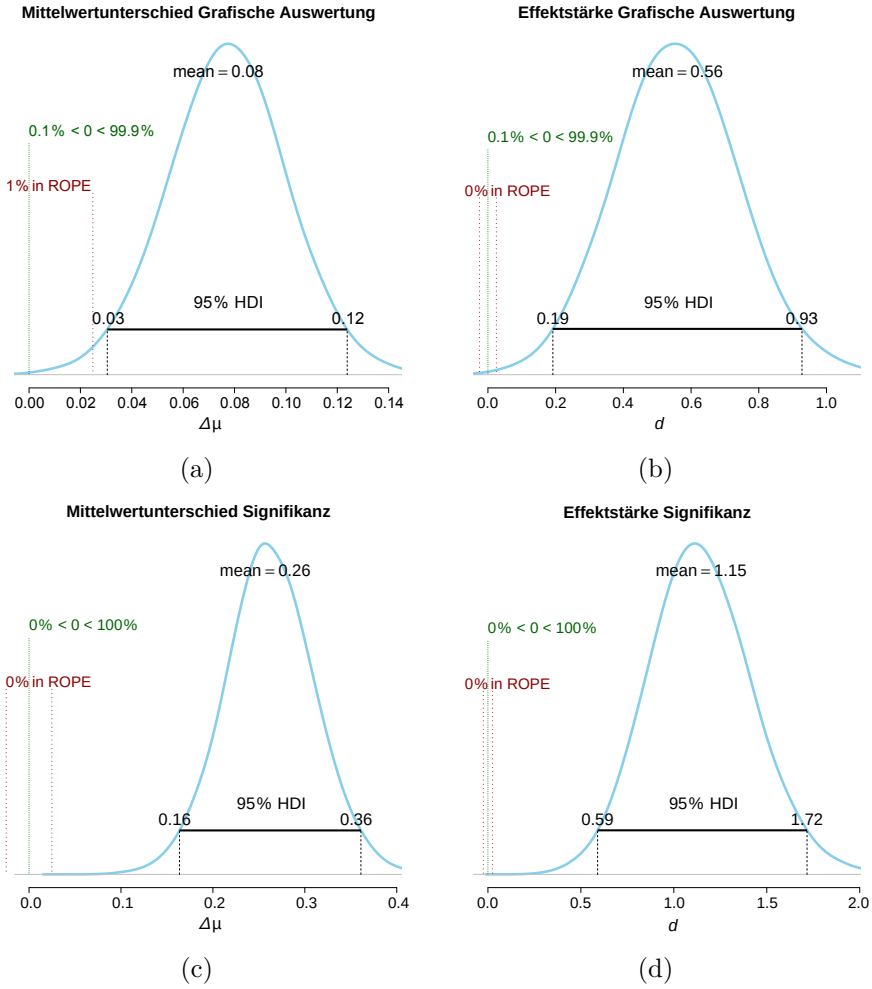


Abb. 5.3: Darstellung der Parameterschätzung für die Mittelwertdifferenz und die Effektstärke des Kompetenztests für die Apps bezüglich der Grafischen Regression (a + b) und Signifikanz (c + d).

keine Überschneidung des 95 % HDI und der ROPE, somit ist der gefundene Unterschied auch für die Praxis als substanziell anzusehen. Bezüglich der Effektstärke konnte ein Mittelwert von $d \approx 1{,}15$ ermittelt werden. Dies entspricht einem großen Effekt (J. Cohen, 1988). Die Posteriorverteilung der Effektstärke zeigt, dass das 95 % HDI und ROPE disjunkt sind. Abbildung 5.3d zeigt, dass die Verteilung deutlich höhere Werte als die Effektstärken der vorigen drei Kompetenzbereiche aufweist. Auf Grundlage der Daten kann mindestens von einem moderaten positiven Effekt und mit einer Wahrscheinlichkeit von nahezu 90 % von einem starken positiven Effekt ausgegangen werden.

5.3 Diskussion

Die vorangehende Analyse hat gezeigt, dass die Apps mit mittleren (A2 & C1) sowie starken (B1 & D1) Effekten die adressierten Datenkompetenzen im Umgang mit Daten und Unsicherheiten fördern. Die Analyse liefert mit Wahrscheinlichkeiten von über 95 % Evidenz für die Gültigkeit der Forschungshypothesen 1a, 1b, 1c und 1d.

In Bezug auf die Forschungsfrage, inwiefern die jeweils adressierten Kompetenzen im Umgang mit Daten und Unsicherheiten gefördert werden können (FF1), lässt sich somit konstatieren, dass die Datenkompetenzen unter Verwendung der Apps gefördert werden können. Das Kernelement der Intervention ist eine interaktive und selbstregulierte Lernphase, in der sich die Lernenden mit in der jeweiligen App adressierten Konzepten und Kompetenzen auseinandersetzen. Innerhalb von kurzen Zeitintervallen können somit Schüler:innen entsprechende Kompetenzen vermittelt werden, um experimentelle Daten physikalisch adäquat auszuwerten. Durch den Interventionscharakter der Apps und die kurze Bearbeitungsdauer eignen sich die Apps insbesondere für die Verwendung im Unterricht. Die Orientierung am Alltagskontext weist eine geringe Handlungsschwelle für die Lernenden auf, da Schüler:innen sich nicht auf zwei Fachkontexte (Unterricht & Intervention) einlassen müssen.

Auffällig ist die unterschiedliche Wirksamkeit der Apps. Die Apps B1 und D1 weisen höhere Mittelwerte bezüglich der Effektstärken auf als die Apps A2 und C1. Ein Vergleich der jeweils verwendeten Kompetenztests zeigt, dass die ermittelte Reliabilität des Tests für die Apps A2 und C1 geringer ist ($\alpha_A = 0{,}60$, $\alpha_C = 0{,}63$) als für die Apps B1 und D1 ($\alpha_B = 0{,}82$, $\alpha_D = 0{,}76$). Im Rahmen der 95 %-Konfidenzintervalle überschneiden sich die ermittelten Reliabilitäten zwar, aber es besteht dennoch die Möglichkeit, dass die Datenkompetenz, die von den Apps A2 und C1 adressiert wurde, mit den Tests schlechter erfasst wurde als in den zu den Apps B1 und D1 zugehörigen Kompetenztests. Einerseits ist denkbar,

dass die von den Apps A2 und C1 adressierten Kompetenzen und Konzepte schwerer sind als in den Apps B1 und D1. Die Items zur Evaluation von App A2 sind im Vergleich zu den restlichen verwendeten Tests schwieriger (vgl. Abschnitt 5.1.2). Ein möglicher Grund für die erhöhte Schwierigkeit des Tests könnte sein, dass die Items in diesem Fall selbstständig entwickelt worden sind, da die von App A2 adressierten Konzepte und Kompetenzen nur in sehr wenigen Items von Schulz (2022) enthalten waren. Die restlichen Tests unterscheiden sich nicht stark in der Schwierigkeit, was als Indiz dafür angesehen werden kann, dass die jeweils adressierten Konzepte und Kompetenzen von den Schüler:innen als gleich schwer aufgefasst wurden. Andererseits ist denkbar, dass die Effektstärke der Apps A1 und C1 durch die geringe Reliabilität der verwendeten Tests unterschätzt wird, weil das zugrunde liegende latente Konstrukt nicht ausreichend reliabel erfasst wurde und die inhaltliche Validität der Apps nicht gewährleistet ist (vgl. Abschnitt 5.2).

Zudem ist zu berücksichtigen, dass die unterschiedlichen Apps an verschiedenen Schulen und mit verschiedenen Klassenstufen evaluiert worden sind. Die Apps wurden jeweils von zwei bis drei Klassen bearbeitet, sodass Klassen- bzw. Schuleffekte durchaus denkbar sind. Ein Quervergleich der ermittelten Effektstärken für die Wirksamkeit der Apps sollte daher vorsichtig erfolgen. Dies stellt einen weiteren möglichen Punkt dar, welcher von zukünftigen Forschungsprojekten untersucht werden könnte, um die Wirksamkeit der Apps hinsichtlich der Klassenstufe und des Schultyps präziser bestimmen zu können.

Die größte Effektstärke weist die App D1 des Kompetenzbereichs Signifikanz auf ($d \approx 1{,}15$). Im Gegensatz zu den übrigen Apps lernen die Schüler:innen auf grafische Weise, dass die Überlappung der Unsicherheitsintervalle bedeutet, dass die Messergebnisse nicht voneinander unterscheidbar sind. Die übrigen Apps adressieren rechnerische Kompetenzen wie die Ermittlung von Typ A und Typ B Unsicherheiten oder das Addieren von Messergebnissen. Das Konzept der Verträglichkeit von Messergebnissen weist somit eine geringere Komplexität auf, was der Grund für die hohe Wirksamkeit gewesen sein könnte. Auf einer ähnlichen Stufe bezüglich des Abstraktionsgrads ist die App C1. Diese adressiert die Eigenschaften einer Ausgleichsgeraden. In dem darauffolgenden Datenkompetenztest müssen die Schüler:innen entsprechend Ausgleichsgeraden einzeichnen oder auswerten. Die Effektstärke ist hier mit $d \approx 0{,}56$ deutlich geringer als bei der App, welche den Kompetenzbereich Signifikanz adressiert. Denkbar wäre, dass die unterschiedlichen kognitiven Vorgehensweisen beim Zeichnen und Rechnen zu den verschiedenen Resultaten geführt haben. Es muss jedoch berücksichtigt werden, dass die Reliabilität des Tests im Vergleich geringer ist. Eine Überarbeitung des Datenkompetenztests der App C1 könnte in einer weiteren Studie verwendet werden, um zu evaluieren, inwiefern die rechnerische Komplexität der Apps A1 und B1 sich auf die Effektstärken der Intervention auswirkt. Ein weiterer praktischer Einsatz der App D1 wird in Abschnitt 6.2 evaluiert.

Das aktuelle längsschnittliche Design der Untersuchung ermöglicht die Evaluation der Apps hinsichtlich ihrer Wirksamkeit. In dieser Form ist es jedoch nicht möglich zu evaluieren, wie zeitlich stabil der Lernerfolg ist. Durch eine weitere Studie mit einem Follow-up-Test könnte dieser Aspekt untersucht werden.

Im Rahmen dieser Arbeit wurde nicht jede App aus Kapitel 4.1 evaluiert. Vor dem Hintergrund der Testökonomie und der Tatsache, dass alle Apps auf denselben Prinzipien beruhend entwickelt wurden, wurde entschieden, stellvertretend jeweils eine App der vier Kompetenzbereiche Direktes Messen, Indirektes Messen, Grafische Auswertung und Signifikanz zu evaluieren. Die Apps unterscheiden sich somit lediglich in den adressierten Kompetenzen und Konzepten. Außerdem gibt es Unterschiede in den implementierten interaktiven Elementen innerhalb der Apps, die nur eingesetzt wurden, wenn der Einsatz dieser didaktisch sinnvoll erscheint. Für die evaluierten Apps konnte eine effektive und effiziente Wirksamkeit gezeigt werden, sodass es naheliegend ist, dass die übrigen Apps ebenso wirksam sind und für die Förderung der Datenkompetenz für die Auswertung experimenteller Daten verwendet werden können. An dieser Stelle könnten zukünftige Forschungsprojekte ansetzen und untersuchen, wie sich die Wirksamkeit der Apps innerhalb eines Kompetenzbereichs verhält.

Zusammenfassend lässt sich sagen, dass die kurzen und modularen Interventionen verwendet werden können, um die Datenkompetenz von Schüler:innen zu adressieren und zu fördern. Es konnte empirisch gezeigt werden, dass Kompetenzen von Schüler:innen selbstreguliert und in wenigen Minuten mithilfe von spezifischen Lernapps mit mittleren bis großen Effekten gefördert werden können. Die mindestens moderate positive Wirkung der Apps ist unabhängig von den adressierten Konzepten und Kompetenzen. Diese vielversprechenden Ergebnisse gemeinsam mit der modularen Ausrichtung der Apps sowie der Orientierung am Alltagskontext ermöglichen es, die Apps flexibel im Unterricht einzusetzen.

Kapitel 6

Wirkung der Kompetenzen im Umgang mit Daten und Unsicherheiten auf das Argumentieren und den Lernerfolg

6.1 Methode

Das Ziel der vorliegenden Arbeit ist die Untersuchung des Einflusses von Datenkompetenz auf das Argumentieren beim Experimentieren. Nachdem in Studie 1 und Studie 2 die Entwicklung sowie Evaluation der Apps dargelegt worden sind, untersucht die vorliegende Studie nun die Wirkung der Apps auf das Argumentieren beim Experimentieren. Es wird analysiert, welchen Einfluss a) die Förderung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien und die Richtigkeit der Hypothese und b) die Zustimmung zu den Argumentkategorien auf die Richtigkeit der Hypothese hat. Dabei wird sich an dem Vorgehen von Ludwig (2017) orientiert. Eine ausführliche Beschreibung erfolgt nur für die Methoden und Punkte, in denen sich die vorliegende Studie und die Arbeit von Ludwig (2017) unterscheiden.

6.1.1 Beschreibung der Experimentiersituation

Die Studie wurde unter Verwendung des Umfragetools LimeSurvey digital durchgeführt. In diesem Rahmen wurde davon abgesehen, die Schüler:innen selbstständig experimentieren zu lassen. Wie in Abschnitt 2 dargelegt, ist davon auszugehen, dass die beobachteten Zusammenhänge durch die Bereitstellung der Daten nicht beeinflusst werden (bspw. Ludwig, 2017; Pfeiler, 2019). Für den Einsatz eines vorgefertigten Datensatzes sprechen einige praktische Gründe: Einerseits kann der zeitliche Rahmen der Studie besser kontrolliert werden

und andererseits eine gleiche Datenqualität für alle Schüler:innen gewährleistet werden. Es wurde entschieden, einen experimentellen Kontext anhand eines Alltagskontexts zu motivieren und anschließend das entsprechende physikalische Experiment ausführlich zu beschreiben. Den Schüler:innen wurden die zum Versuch gehörenden experimentellen Daten anschließend vorgelegt.

Als Kontext wurde sich wie Ludwig (2017) auf das Fadenpendel bezogen. Für die Auswahl dieses Experiments sprechen einige Gründe, auf die im Folgenden detailliert eingegangen wird.

Relevanz Dieser Punkt trifft auf zwei Ebenen zu: Einerseits ist das Fadenpendel in den fakultativen Themen des Bildungsplans der Gymnasien für das Bundesland Baden-Württemberg in der Sekundarstufe I vorgesehen (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016). Andererseits werden verschiedene Messreihen ermittelt und miteinander verglichen. Diese prozessbezogenen Kompetenzen der Lernenden sind für viele Physikexperimente im Unterricht relevant (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016, S. 8 ff.).

Vergleichbarkeit Die Verwendung eines anderen Kontextes bzw. eines anderen Experiments erschwert die Vergleichbarkeit von Studien. Die Forschungsidee zu der vorliegenden Arbeit basiert auf den Ergebnissen von Ludwig (2017) und die Resultate werden daher in erster Linie mit den dortigen Ergebnissen verglichen. Käme es zu unterschiedlichen Ergebnissen unter Verwendung eines anderen Kontexts, könnte der Kontext nicht als mögliche Ursache ausgeschlossen werden. Daher wird auf den Kontext *Fadenpendel* zurückgegriffen.

Starke Präkonzepte Hinsichtlich des Zusammenhangs der Periodendauer und der schwingenden Pendelmasse besitzen Lernende Präkonzepte. In diversen Arbeiten konnte gezeigt werden, dass diese in diesem Kontext stark ausgeprägt sind und rund 90 % den falschen Zusammenhang zwischen der Pendelmasse und der Periodendauer vermuten (Kanari & Millar, 2004). Dieser Wert konnte in der Arbeit von Ludwig (2017, S. 94) mit 88 % falschen Eingangshypothesen bestätigt werden. Die starke Dominanz der Präkonzepte im Rahmen dieses Kontextes führt dazu, dass der Kontext sich besonders eignet, um Schüler:innen mit nicht hypothesenkonformen Daten zu konfrontieren. Für die vorliegende Arbeit ist diese Eigenschaft von großer Relevanz, da so eine argumentative Auseinandersetzung und ein eventueller Konzeptwechsel herbeigeführt werden können (Kind et al., 2011).

Die von Ludwig beschriebenen Effekte bezüglich der Schwerpunktverschiebung und Änderung der effektiven Fadenlänge (Ludwig, 2017, Vgl. S. 58 f.),

können in den vorgefertigten Messwerten nicht unmittelbar ausgeschlossen werden. Anders als bei Ludwig wurde nicht mit einem Gewichtsteller und Schlitzgewichten gearbeitet, sondern mit drei 50 g Gewichten, die je nach Messreihe aneinander gehangen wurden. Eine Berechnung der Mittelwerte zeigt, dass augenscheinlich mit jedem zusätzlichen angehängten Massestück die Periodendauer minimal zunimmt.

Im Rahmen der Studie ist dieser Effekt für die Untersuchung des Zusammenhangs zwischen Periodendauer T und Pendelmasse m unproblematisch, da die Unsicherheit der Messwerte über die größte Abweichung des Mittelwerts zu den Messwerten bestimmt wird. Eine Gegenüberstellung verschiedener Möglichkeiten zur Abschätzung der Unsicherheit ist in Tabelle 6.1 dargestellt. Allerdings könnte diese Eigenschaft der Messwerte Schüler:innen in Präkonzepten bestätigen, wenn das Konzept von Messunsicherheiten und eines vollständigen Messergebnisses nicht bekannt sind. Die Proband:innen könnten dann darin bestärkt werden, dass es eine positive Abhängigkeit zwischen den zu untersuchenden Größen gibt.

Tab. 6.1: Dargestellt sind arithmetische Mittel der Periodendauer T und verschiedene Unsicherheiten für unterschiedliche Pendelmassen m (vgl. Abschnitt D). Für weitere angehängte Pendelmassen könnte die Verwendung der Standardabweichung zu nicht verträglichen Messergebnissen führen.

		Abschätzung der Unsicherheit	
m	$\overline{T}$	Maximalfehler	Standardabweichung
50 g	2,03 s	0,02 s	0,02 s
100 g	2,04 s	0,03 s	0,02 s
150 g	2,05 s	0,02 s	0,01 s
200 g	2,06 s	0,04 s	0,02 s

6.1.2 Beschreibung der Intervention

Ein wesentlicher Unterschied zum Vorgehen von Ludwig (2017) besteht darin, dass anstelle des Fachwissens der Schüler:innen die Kompetenzen im Umgang mit Daten und Unsicherheiten untersucht werden. Zu diesem Zweck werden die Lernumgebungen aus Kapitel 5.2 eingesetzt. Die Apps eignen sich insbesondere für das zuvor beschriebene Experiment, da die Schüler:innen Messergebnisse für unterschiedliche Messreihen auswerten und vergleichen müssen. Für den Vergleich der Messergebnisse müssen die Schüler:innen folgende Kompetenzen beherrschen:

Die Schüler:innen können...

1. ... ein vollständiges Messergebnis bestehend aus dem arithmetischen Mittel sowie einer Messunsicherheit bestimmen und die größte Abweichung des Messwerts vom arithmetischen Mittel als Messunsicherheit verwenden.

2. ... Messergebnisse hinsichtlich ihrer metrologischen Verträglichkeit bzw. Unterscheidbarkeit vergleichen.

Diese Kompetenzen werden durch die Apps A1 und D1 abgedeckt (vgl. Abschnitt 4.1.1). Die Apps werden in der Studie als Teil 1 und Teil 2 bezeichnet und wurden von den Lernenden hintereinander in maximal 45 Minuten bearbeitet. Für detaillierte Ausführungen zu den Eigenschaften und die ausführlichen didaktischen Konzepte der Lernumgebungen wird auf Kapitel 5.2 sowie Anhang A verwiesen.

6.1.3 Operationalisierung der Konstrukte

Der folgende Abschnitt beschreibt, wie die Konstrukte *Kompetenzen im Umgang mit Daten und Messunsicherheiten*, *Zustimmung zu den Argumentkategorien* und *Lernerfolg* operationalisiert und erfasst wurden. Dabei wird auf die grundlegenden verwendeten Testinstrumente (Ludwig, 2017; Schulz, 2022) sowie auf eventuelle Überarbeitungen der Items eingegangen. Detaillierte Beschreibungen zu den Überarbeitungen der Testinstrumente, eine Übersicht der Testinstrumente sowie die stetige Genese des globalen Messmodells befinden sich im Anhang.

Erfassung der Kompetenzen im Umgang mit Daten und Unsicherheiten

Die Kompetenzen im Umgang mit Daten und Unsicherheiten wurden mit dem überarbeiteten Testinstrument aus Kapitel 5.1 erfasst. Verwendet wird der Test für die App D1, welche die Verträglichkeit bzw. Unterscheidbarkeit von Messergebnissen adressiert.

Das Testinstrument umfasst die Items 301, 303, 304, 305, 307, 309, 314, 315 von Schulz (2022), die in der gemäß Kapitel 5.1 überarbeiteten Fassung verwendet werden. Die Validität der verwendeten Items konnte von Schulz (2022) durch ein Expertenrating sichergestellt werden. Für das Testinstrument, aus dem die zuvor genannten Items stammen, werden eine EAP-Reliabilität von 0,80 und eine WLE-Reliabliität von 0,70 berichtet (Schulz, 2022, S. 108). In der vorliegenden Arbeit konnte eine WLE-Reliabilität von 0,59 ermittelt werden. Die geringere Reliabilität im Vergleich zu Schulz (2022) lässt sich durch zwei mögliche Faktoren erklären: Einerseits ist der verwendete Test deutlich kürzer als im Fall von Schulz (2022). Andererseits adressiert der vorliegende Test eine spezifische Kompetenz zum Vergleich von Messergebnissen. Beide Aspekte

Zeitmessung
Ein Schüler erhält bei einer Zeitmessung ein Messergebnis von $(3{,}5 \pm 0{,}5)$ s. Der Vergleichswert für die Messung ist 3,2 s.

Entscheide und begründe, ob das Messergebnis und der Vergleichswert voneinander unterscheidbar sind.

- ○ Die Messergebnisse und der Vergleichswert sind **nicht** voneinander **unterscheidbar**, denn das Intervall der Unsicherheit geht von 3,0 s bis 4,0 s.
- ○ Die Messergebnisse und der Vergleichswert sind **nicht** voneinander **unterscheidbar**, denn das Intervall der Unsicherheit geht von 3,2 s bis 3,8 s.
- ○ Die Messergebnisse und der Vergleichswert sind voneinander **unterscheidbar**, denn das Intervall der Unsicherheit geht von 3,5 s bis 4,0 s.
- ○ Die Messergebnisse und der Vergleichswert sind voneinander **unterscheidbar**, denn das Intervall der Unsicherheit geht von 3,5 s bis 4,5 s.

Abb. 6.1: Beispielitem aus dem Testinstrument für die Erfassung der Datenkompetenz. Es kann dem Konzept Vergleich von Messwerten zugeordnet werden (Hellwig, 2012; Schulz, 2022).

können nach Bond und Fox (2007, S. 58) Ursachen für eine geringe Reliabilität sein. Außerdem berichtet Kauertz (2008), dass Deckeleffekte die ermittelte Reliabilität verringern können (vgl. Abb. E.4). Aufgrund der Trennschärfe der Items sowie der guten Verteilung der Itemschwierigkeiten im Vergleich zu den Schüler:innenfähigkeiten wird von einer ausreichend reliablen Messung der Datenkompetenz ausgegangen. Eine detaillierte Evaluation der Items und des Tests kann Abschnitt E.2 entnommen werden.

Insgesamt ergibt sich eine Länge von zehn Items, wobei das erste Item sich auf den ersten Anstrich der zuvor genannten Kompetenzen bezieht. Die übrigen neun Items adressieren den zweiten Anstrich der Schüler:innen bezüglich der Verträglichkeit von Messergebnissen. Anhand des ersten Items soll beobachtet werden, inwiefern die Lernenden ein vollständiges Messergebnis formulieren können. Schüler:innen, die nicht in der Lage sind, die Bestandteile eines vollständigen Messergebnisses zu nennen, können nur schwer die Unterscheidbarkeit der Messergebnisse auf Grundlage der zugrunde liegenden Messunsicherheiten bewerten. Für die Schätzung der Schüler:innenfähigkeiten werden jedoch nur die restlichen neun Items zur Verträglichkeit von Messergebnissen verwendet (s. Abbildung 6.1 für ein Beispielitem).

Erfassung der Zustimmung zu den Argumentkategorien

Die Zustimmung zu den Argumentkategorien wird mit dem Testinstrument von Ludwig (2017) ermittelt. Dabei werden insgesamt 20 Items verwendet. Die Items besitzen eine fünfstufige Likert-Skala und können jeweils einer der Ar-

gumentkategorien Intuition, Expertenwissen, Messunsicherheiten und Daten als Evidenz zugewiesen werden. Die Zustimmung zu jeder Argumentkategorie kann somit über fünf Items ermittelt werden. Bei dem Testinstrument werden Post-hoc-Analysen bzw. Selbsteinschätzungen der Lernenden anhand vorgelegter Aussagen erhoben. Die Vorteile des Verfahrens („Vermeidung von Effekten sozialer Erwünschtheit, leichte Quantifizierung und die Testökonomie“, Ludwig, 2017, S. 66) sprechen für den Einsatz des Testinstruments. Es handelt sich dabei um das bisher einzige quantitative Verfahren, um die Zustimmung zu Argumentkategorien bzw. Begründungen in situ zu erfassen. Aus diesem Grund findet dieses Verfahren Verwendung. Für alternative Möglichkeiten zur Evaluation von Argumenten wird auf den Abschnitt 2.1.4 verwiesen.

Hypothese bezüglich des Experiments

Die Hypothese der Lernenden wurde erfasst, indem den Schüler:innen Aussagen in Form eines Mutliple-Choice-Items vorgelegt wurden. Im Anschluss an die Abfrage der Hypothese gab es ein offenes Begründungsfeld. Durch die zusätzliche schriftliche und offene Begründung sollte eine intensive Auseinandersetzung mit der Hypothese gewährleistet werden. Die offenen und schriftlichen Begründungen wurden im Rahmen dieser Arbeit nicht ausgewertet. Der Lernerfolg ist daher dichotom operationalisiert.

Ein Vergrößern der Pendelmasse bewirkt, dass die Schwingungsdauer …

… größer wird.
… gleich wird.
… kleiner wird.

Eine derartige Operationalisierung der Hypothese zieht eine Ratewahrscheinlichkeit von $\frac{1}{3}$ nach sich. Dem kann jedoch entgegengestellt werden, dass in diesem Experiment Präkonzepte sehr dominant sind und die nicht vorhandene Abhängigkeit der Pendelmasse m und Periodendauer T von Schüler:innen nicht als plausibel angenommen wird (Kanari & Millar, 2004).

Im Verlauf der Studie wurde die Hypothese der Lernenden mehrfach erfasst (vgl. Abschnitt 6.1.4). Bei der erneuten Erfassung der Hypothese wurden die Lernenden gefragt, ob sie bei ihrer bisherigen Hypothese bleiben oder wechseln. Eine erneute Auswahl der Hypothese ist daher nur möglich gewesen, wenn die Lernenden ihre Hypothese zwischen den Erfassungszeitpunkten gewechselt haben.

6.1.4 Studiendesign

Bei den Forschungshypothesen zu den Forschungsfragen FF2 bis FF4 handelt es sich um Zusammenhangshypothesen, daher eignet sich in diesem Fall eine explanative Studie (Döring & Bortz, 2016, S. 192). Es wird wie in Studie 1 ein quasi-experimentelles Ein-Gruppen-Prätest-Posttest-Design als Forschungsdesign verwendet. Eine Diskussion zu dieser Methode wird daher an dieser Stelle nicht durchgeführt (vgl. Abschnitt 5.1).

Im Unterschied zur vorigen Studie wurden in diesem Fall mehrere unabhängige Variablen erfasst. Durch die Erfassung multipler Konstrukte kann die interne Validität des Designs erhöht werden. Dies ermöglicht es eher, kausale Zusammenhänge zu beobachten.

Die Datenkompetenz und die Zustimmung zu den Argumentkategorien wurden jeweils vor und nach der Intervention erhoben. Die Hypothese der Schüler:innen wurde ebenfalls vor und nach der Intervention erfasst. Zusätzlich hierzu wurde die Eingangshypothese der Lernenden zu Beginn erhoben. Auf diese Weise soll der Einfluss der Daten auf die Hypothese der Lernenden kontrolliert werden. Eine detaillierte Beschreibung des Studienablaufs findet sich im kommenden Abschnitt.

Die Studie wurde in einem digitalen synchronen Format durchgeführt. Einerseits konnte den Schüler:innen so der Zugang stark vereinfach werden, da kein Wechsel der Plattform bzw. des Mediums notwendig war. Die Testinstrumente und Apps konnten in einer abgeschlossenen Einheit bearbeitet werden. Das digitale Format verringert das Risiko von Übertragungsfehlern oder fehlerhafter Zuordnung von Prä- und Posttests. Außerdem konnten durch die synchrone Durchführung parallel der Fortschritt der Lernenden beobachtet und ein reibungsfreier Ablauf gewährleistet werden.

6.1.5 Ablauf der Studie

Zunächst gab es eine kurze Einführungsphase seitens des Studienleiters. Anschließend wurde den Schüler:innen Zugang zu den Testinstrumenten gewährt, welche in LimeSurvey aufbereitet worden sind. Den Schüler:innen wurde ein Zeitungsartikel präsentiert, welcher den physikalischen Kontext des Fadenpendels aus dem Alltag heraus motiviert. Anschließend folgte eine Beschreibung des Versuchs und der Durchführung. Nun wurde zum ersten Mal die Hypothese der Lernenden erfasst. Daraufhin wurde den Lernenden ein Datensatz präsentiert (s. Anhang D). Hierbei handelt es sich um eine Versuchsreihe, bei der die Massen $m_1 = 50\,\mathrm{g}$, $m_2 = 100\,\mathrm{g}$, $m_3 = 150\,\mathrm{g}$ und $m_4 = 200\,\mathrm{g}$ an das Pendel gehängt wurden. Die

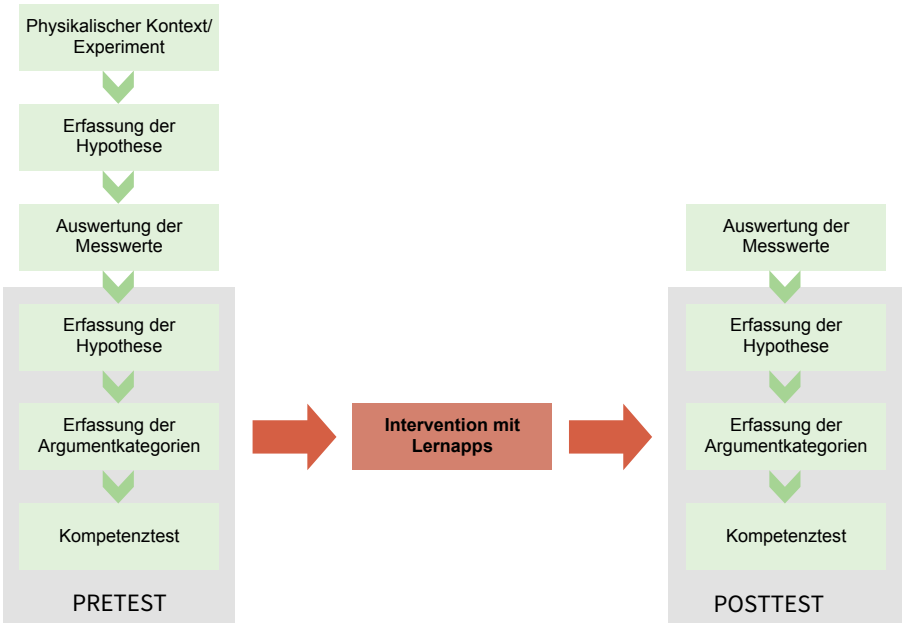


Abb. 6.2: Übersicht zum Ablauf der Studie.

Periodendauer wurde für jede Masse m_i sechsmal erfasst. Nach Betrachtung bzw. Auswertung der Daten, wurden die Lernenden erneut aufgefordert, ihre Hypothese aufzugeben. Diesmal wurde zuvor gefragt, ob die Hypothese beibehalten oder gewechselt wird. Nach Erfassung der Hypothese erfolgte der Einsatz der Testinstrumente zur Erfassung der Kompetenzen im Umgang mit Daten und Unsicherheiten sowie der Zustimmung zu den Argumentkategorien. Danach erfolgte die Intervention in Form der Apps, welche die Schüler:innen digital bearbeitet haben. Anschließend wurden den Lernenden dieselben Messdaten wie zuvor präsentiert. Nach erneuter Betrachtung bzw. Analyse wurden final die Hypothese der Lernenden zum Experiment, die Zustimmung zu den Argumentkategorien und die Kompetenz im Umgang mit Daten und Messunsicherheiten erfasst. Die Studie wurde mit den Proband:innen im Rahmen von Doppelstunden durchgeführt. Die Bearbeitungsdauer der Studie liegt im Bereich von 65 bis 90 Minuten. Eine Übersicht des Ablaufs findet sich in Abbildung 6.2.

6.1.6 Pilotierung

Der Einsatz der Instrumente wurde in zwei Studien pilotiert. Das Instrument zur Erfassung der Datenkompetenz wurde für den Einsatz zur Evaluation der

Apps bereits pilotiert. Die Ergebnisse der Pilotierungen können in Kapitel 5.1 nachgelesen werden. Zusammenfassend lässt sich sagen, dass die Items für die jüngere Stichprobe hinsichtlich Layout und Formulierungen angepasst werden mussten.

Die Apps, das Testinstrument zur Erfassung der Zustimmung zu den Argumentkategorien und das grundlegende Studiendesign wurden mit $n = 75$ Schüler:innen pilotiert. Dabei sollte die grundsätzliche Wirksamkeit der Apps anhand eines Prototyps untersucht werden. Die Pilotierung konnte zeigen, dass die Effekte der App mit dem Testinstrument von Ludwig (2017) erfasst werden können (Donner, 2019). Ebenso konnte ein zeitlicher Rahmen von rund 70 Minuten abgesteckt werden (Donner, 2019). Somit konnte davon ausgegangen werden, dass der zusätzliche Kompetenztest für die Erfassung der Datenkompetenz innerhalb des Zeitfensters der Studie von 90 Minuten bearbeitet werden kann.

6.1.7 Beschreibung der verwendeten statistischen Verfahren

Der folgende Abschnitt beinhaltet die grundlegende verwendete Methode für die Analyse der Konstrukte. Detaillierte Analysen bzw. Ausführungen zu den Gütekriterien können im Anhang nachgelesen werden.

Analyse der Datenkompetenz mittels Rasch-Analyse

Die Auswertung des Kompetenztests wird anhand der *Item-response*-Theorie durchgeführt. Unter Verwendung der Rasch-Analyse können die Schüler:innenfähigkeit, in der vorliegenden Arbeit die Datenkompetenz, und die Itemschwierigkeiten des Testinstruments latent modelliert werden (Brown, 2006). Neben der latenten Modellierung kommt hinzu, dass die beiden Parameter stochastisch unabhängig voneinander modelliert werden (Rasch, 1966). Auf diese Wiese kann nicht nur die Passung des Modells zu den Daten untersucht werden, sondern auch die Passung des Testinstruments zu den ermittelten Schüler:innenfähigkeiten.

Auf Grundlage der Antworten jedes Lernenden wird für jedes Item eine Wahrscheinlichkeitsverteilung generiert. Mit dieser Verteilung kann für eine fixierte Schüler:innenfähigkeit eine Wahrscheinlichkeit angegeben werden, mit der die Aufgabe korrekt gelöst wird. Dieser Zusammenhang lässt sich mit θ_n als Fähigkeit von Schüler:in n und ξ_i als Schwierigkeit von Item i als

$$P(1|\theta_n, \xi_i) = \frac{e^{\theta_n - \xi_i}}{1 + e^{\theta_n - \xi_i}} \tag{6.1}$$

formalisieren (Brown, 2006, S. 397). Ein Vorteil dieses Ansatzes ist, dass die Parameter der Itemschwierigkeit und Schüler:innenfähigkeit mit derselben Metrik abgebildet werden. Die Schüler:innenfähigkeiten und Itemschwierigkeiten können dabei Werte zwischen $-\infty$ und ∞ annehmen. Der Zusammenhang dieser

Parameter wird in der sogenannten Item Characteristic Curve (ICC) dargestellt. Mathematisch handelt es sich bei der ICC um ein logistisches Wachstum, deren Wachstum im Fall einer Rasch-Analyse auf 1 begrenzt ist. Für ein ideales Item würde gelten, dass für eine bzw. einen Schüler:in mit der Fähigkeit $\theta_i = 0$ die Lösungswahrscheinlichkeit 50 % beträgt. Allgemein kann im Fall von Gleichung 6.1 ebenfalls von einem sogenannten 1PL-Modell ausgegangen werden, wenn der Diskriminationsparameter nicht auf einen Wert von 1 fixiert wird (Wu et al., 2016, S. 228).

Die Schüler:innenfähigkeiten werden zunächst separat zum Prä- und Postzeitpunkt evaluiert. Diese Vorgehensweise wurde gewählt, um zunächst sicherzustellen, dass die Prä- und Posttests eine gute Passung zum Rasch-Modell aufweisen und zu beiden Zeitpunkten gleich funktionieren (vgl. Abschnitt E.2.2). Anschließend wird ein gesamtes Messmodell erstellt, welches in Abschnitt 6.1.9 ausführlich dargelegt ist. Durch die gemeinsame Modellierung des Prä- und Postzeitpunkts kann besser analysiert werden, wie sich die Schüler:innenfähigkeit durch die Intervention verändert hat (Wright, 1996, 2003).

Die Verwendung des Rasch-Modells hat als Voraussetzung die lokale stochastische Unabhängigkeit der Aufgabenbearbeitung (Wu et al., 2016, S. 192). Ist diese Voraussetzung verletzt, kann die spezifische Objektivität des Modells (Wu et al., 2016) oder die Robustheit der Parameterschätzung beeinträchtigt sein. In der vorliegenden Arbeit handelt es sich um wiederholte Messungen derselben Schüler:innen, so dass dieser Aspekt in der Diskussion (s. Abschnitt 6.3) aufgegriffen werden muss. Eine Diskussion zu dieser Limitierung und eine Diskussion zu anderen Nachteilen der IRT finden sich beispielsweise in Becker (2004).

Analyse der Zustimmung zu den Argumentkategorien

Die Analyse der Zustimmung zu den Argumentkategorien wird latent modelliert. Latente Verfahren weisen unter anderem die Vorteile auf, dass sowohl Messfehler als auch eine Gewichtung des Einflusses, den jedes Item auf das Konstrukt besitzt, in der Modellierung berücksichtigt werden kann (sog. τ-kongenerische Variablen, vgl. Eid et al., 2010, S. 835). Dadurch können die Parameter und Effekte messfehlerbereinigt geschätzt werden (Geiser, 2010, S. 41).

Das Testinstrument von Ludwig (2017) hat die Zustimmung der Schüler:innen zu den Argumentkategorien Intuition, Expertenwissen, Messunsicherheiten und Daten als Evidenz erfasst. Bei den verwendeten Items handelt es sich um fünfstufige Likert-Skalen, welche für die Analyse numerisch codiert wurden (0 = „trifft gar nicht zu“ bis 4 = „trifft voll zu“). Eine Diskussion zur Umcodierung der Ordinalskala in numerische Werte findet sich in Abschnitt E.3.2 und bei verschiedenen Autoren (bspw. Knapp, 1990).

Die Zustimmung zu den Argumentkategorien wird jeweils durch eine latente Variable modelliert. Ein latentes Messmodell berechnet latente Variablen auf Grundlage mehrerer manifester Variablen, die als Indikator fungieren. Dabei

wird untersucht, inwiefern das latente Konstrukt F die Indikatoren X_i erklären kann. Formal lässt sich der Zusammenhang als

$$\vec{\alpha} \cdot F + \vec{\delta} = \vec{X} \tag{6.2}$$

darstellen, wobei $\vec{\alpha}$ die Faktorladungen beinhaltet, $\vec{\delta}$ die Varianz für jeden Indikator beinhaltet und $\vec{X}$ die beobachteten Indikatoren darstellt. In dem vorliegenden Fall der Studie wird die Zustimmung zu den Argumentkategorien jeweils durch fünf Indikatoren erfasst. Eine ausführliche Darstellung erfolgt in Abschnitt E.3.3.

Wird das Modell um Zusammenhänge zwischen den latenten Faktoren ergänzt, spricht man von einem Strukturgleichungsmodell. Dabei können sowohl gerichtete (latente Pfadanalysen) als auch ungerichtete (kofirmatorische Faktorenanalysen) betrachtet werden (Geiser, 2010, S. 42).

Eine detaillierte Bestimmung der Messmodelle für die Zustimmung zu den Argumentkategorien kann dem Anhang E.3.1 entnommen werden. Die Charakterisierung des globalen Messmodells ist im Abschnitt 6.1.9 ausführlich dargelegt.

Die Passung von latenten Modellen wird durch die χ^2-Statistik und sog. Fit-Indizes beurteilt. Dabei wird untersucht, inwiefern die empirische und theoretische Varianz-Kovarianzmatrix übereinstimmen. Neben der exakten Testung (χ^2-Test) gibt es Goodness-of-fit-Indizes, die die aufgedeckte Abweichung zwischen empirischer und theoretischer Varianz-Kovarianzmatrix auf empirische Relevanz untersuchen. Eine Übersicht zu den verwendeten Gütekriterien und den jeweiligen Cut-off-Werten ist im Anhang E.3.1 und in Tabelle 6.2 aufgeführt. Die verwendeten Indizes stellen dabei die wichtigsten Indizes für die Evaluation der Modellgüte für das verwendete Programm dar (Geiser, 2010; Muthén & Muthén, 2017).

Der χ^2-Test weist im Fall der Faktorenanalyse einige Schwächen und Restriktionen auf, weswegen bei widersprüchlichen Ergebnissen das Ergebnis der Fit-Indices gewichtet wird. Die unterschiedlichen Fit-Indizes können verwendet werden, um einige Schwächen des χ^2-Tests zu umgehen (Yu, 2002). Im Rahmen dieser Arbeit werden zusätzlich zum χ^2-Test die Goodness-of-fit-Indizes Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA) und Standardized Root Mean Square (SRMR) für die Evaluation der Modellpassung berücksichtigt.

Für die Festlegung des Schätzverfahrens müssen die Daten auf multivariate Normalverteilung geprüft werden (s. Abschnitt E.3.2). Für die Evaluation der multivariaten Normalverteilung wird auf den Energy-Test und die Teststatistik von Mardia zurückgegriffen (Ebner & Henze, 2020). Der Energy-Test überprüft unter Verwendung bestimmter Metriken die *energetische Distanz* zwischen den theoretischen multivariat normalverteilten Zufallsvariablen und den empirisch ermittelten Schätzungen. Eine explizite Überprüfung der Schiefe und Kurtosis der Daten wird durch die Teststatistik von Mardia realisiert. Die Ergebnisse

Tab. 6.2: Übersicht über die Fit-Indizes zur Evaluation der Modellgüte im Rahmen der latenten Modellierung. Die χ^2-Statistik ist nicht aufgeführt. Diese sollte für die Modellpassung nicht signifikant sein. Bei Werten unterhalb bzw. oberhalb der aktzeptablen Passung wird das entsprechende Modell abgelehnt (Cut-off-Kriterium).

Fit-Indizes	Gute Passung	Akzeptable Passung
CFI	>0,95	>0,90
RMSEA	<0,05	<0,08
90 % CI *	< 0,05	-
SRMR	<0,05	<0,10

* Evaluation anhand eines Signifikanztests

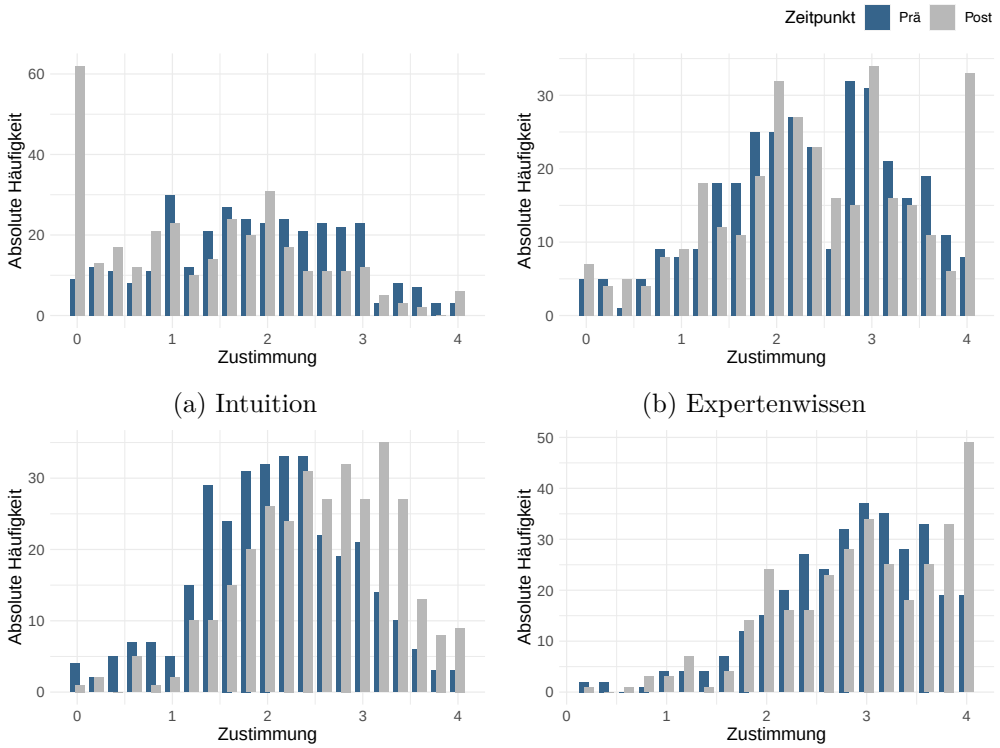
Tab. 6.3: Überprüfung der multivariaten Normalveteilung der Indikatoren. Überprüft werden die Energiedistanz der empirischen und theoretischen Verteilung (Energie-Test) und die Kurtosis und Schiefe der empirischen Verteilung (Mardias Teststatistik).

ArgKat		Energy-Test		Mardias Teststatistik					
				Schiefe			Kurtosis		
		E	p	$\hat{\beta}$	κ	p	$\hat{\beta}$	κ	p
Int	pre	3,64	<0,05	1,87	101,58	<0,05	40,23	5,63	<0,05
	post	19,79	<0,05	3,68	199,16	<0,05	56,55	12,22	<0,05
Exp	pre	4,62	<0,05	2,39	129,46	<0,05	40,17	5,57	<0,05
	post	9,40	<0,05	2,62	141,99	<0,05	45,79	11,62	<0,05
MU	pre	2,41	<0,05	1,08	58,66	<0,05	37,83	3,05	<0,05
	post	6,57	<0,05	3,55	192,42	<0,05	42,34	7,90	<0,05
Evi	pre	6,87	<0,05	4,83	261,61	<0,05	44,13	9,53	<0,05
	post	15,05	<0,05	5,69	308,23	<0,05	49,56	15,69	<0,05

der Tests können in Tabelle 6.3 eingesehen werden. Unter Berücksichtigung der Ergebnisse sowie der Histogramme in Abbildung 6.3 sind für die vorliegende Arbeit robuste Schätzverfahren zu bevorzugen, da für keine der Skalen eine multivariate Normalverteilung der zugrunde liegenden Indikatorvariablen angenommen werden kann. Wie in Abschnitt E.3.2 dargelegt, wird daher ein robustes ML-Schätzverfahren (MLR) verwendet (Muthén & Muthén, 2017).

6.1.8 Stichprobe

Die Studie wurde mit 325 Schüler:innen der neunten (45,2 %) und zehnten (54,8 %) Klasse durchgeführt. Dabei waren 42,4 % der Proband:innen weiblich,



(a) Intuition

(b) Expertenwissen

(c) Messunsicherheiten

(d) Daten als Evidenz

Abb. 6.3: Darstellung der Zustimmung zu den Argumentkategorien. Es zeigen sich Unterschiede in den Prä- und Postdarstellungen. Ebenso ist zu erkennen, dass eine Normalverteilung für keine der Argumentkategorien vorzuliegen scheint.

55,9 % männlich und 1,7 % divers. Alle Schüler:innnen besuchen Gymnasien des Bundeslands Baden-Württemberg. Das Durchschnittsalter der Proband:innen beträgt 15,4 Jahre ($SD = 0{,}8$ Jahre). Bei den Schulen handelt es sich erneut um MINT-freundliche Schulen in Baden-Württemberg. Für eine detaillierte Beschreibung der Stichprobe und der Eigenschaften wird daher auf Abschnitt 5.1.7 verwiesen.

Verwendetes Schätzverfahren Bei der Stichprobe handelt es sich um eine nicht-probabilistische Stichprobe. Die Proband:innen wurden geworben, indem alle MINT-freundlichen Schulen des Landes Baden-Württemberg angeschrieben wurden. Insgesamt wurden rund 250 Schulen angeschrieben, von denen neun Schulen mit 16 Klassen akquiriert werden konnten. Eventuelle Effekte der Profilierung können anhand der Ergebnisse zum Präzeitpunkt evaluiert werden.

Untersuchung der Teststärke

Eine ausreichende Teststärke ist notwendig, um mögliche Effekte mit einer vorher festgelegten Sicherheit untersuchen und aufdecken zu können. Für den weiteren Verlauf der Studie wird eine Teststärke nach J. Cohen (1988) von $1 - \beta = 0{,}80$ festgelegt, wobei β die Wahrscheinlichkeit für den Fehler zweiter Art bezeichnet (Eid et al., 2010, S. 240). Die Teststärke ist von vielen Faktoren, wie beispielsweise von der Anzahl der manifesten Variablen, dem Modell oder der Stichprobengröße, abhängig (Wolf et al., 2013). Zur Abschätzung der Stichprobengröße gibt es verschiedene Methoden und „Daumenregeln“ (bspw. Bentler & Chou, 1987; Comrey & Lee, 1992; Kline, 2015), die vor dem Hintergrund der zuvor genannten Abhängigkeiten und weiterer systematischer Untersuchungen (MacCallum et al., 1996) jedoch mit Vorsicht zu betrachten sind, da die fehlenden Spezifikationen des Modells bei einer Poweranalyse im Kontext von Strukturgleichungsmodellen zu einer Über- bzw. Unterschätzung der Teststärke führen können (Jak et al., 2021). Die Bestimmung der Teststärke im Rahmen von Strukturgleichungsmodellen ist weiterhin Bestandteil der aktuellen Forschung und resultiert in verschiedenen systematischen Zugängen und Tools zur Abschätzung der benötigten Stichprobengröße (bspw. Jak et al., 2021; Wolf et al., 2013). Insbesondere die ersten Autoren diskutieren den Ansatz von Monte-Carlo-Simulationen, da diese verwendet werden können, um die Teststärke in Abhängigkeit des Forscher-spezifischen Modells ermittelt werden kann.

Für die Analyse der Teststärke wurde von der Verwendung von Faustformeln abgesehen und das Vorgehen von MacCallum et al. (1996) verwendet. Das Verfahren bestimmt auf Grundlage des RMSEA Fit-Index und vorgegebener Cut-off-Werte ($\varepsilon_0 = 0{,}05$; $\varepsilon_a = 0{,}01$ für einen not-close-fit-Test, MacCallum et al., 1996, S. 139) den Stichprobenumfang . Dabei wird in Abhängigkeit des Stichprobenumfangs n, der Freiheitsgrade df des zu untersuchenden Modells und der Eigenschaft, dass die Abweichung der Modellpassung annähernd eine χ^2-Verteilung aufweist, die Teststärke ermittelt.

Einerseits weist die vorgeschlagene Vorgehensweise weiterhin eine große Relevanz für die Beurteilung der Teststärke auf (Jobst et al., 2021), andererseits wurde das Testinstrument von Ludwig (2017) verwendet. Die Stichprobengröße, die Betrachtung einer Gruppe und die Verwendung von 40 manifesten Variablen resultiert in einer vergleichbaren Teststärke (Ludwig, 2017, S. 80 f.).

Für die vorliegende Arbeit wird die erreichte Teststärke post hoc evaluiert. Durch einen Stichprobenumfang von $n = 325$ und $df = 698$ Freiheitsgraden des globales Messmodells (vgl. Tabelle 6.4) kann eine hohe Teststärke von $> 0{,}95$ bezüglich der Modellpassungstestung erreicht werden (MacCallum et al., 1996, S. 142).

6.1.9 Analyse der Messmodelle

Bevor mit der Analyse begonnen werden kann, wird die Güte der Messmodelle untersucht. Dabei wird betrachtet, wie gut die Passung zwischen den Daten und dem Modell zur Erfassung der zugrunde liegenden Konstrukte ist. Dieser Schritt ist notwendig, um die Validität und Reliabilität der Messmodelle zu evaluieren. Da es sich um eine longitudinale Untersuchung handelt und das Testinstrument vor und nach der Intervention verwendet wurde, muss gewährleistet werden, dass die Funktionsweise des Tests sich nicht verändert hat und dasselbe Konstrukt gemessen wird. Dieser Schritt ist insbesondere notwendig, da es sich um ein situatives Konstrukt handelt (Ludwig, 2017, S. 135), welches vor und nach einer Intervention ermittelt wird. Zu diesem Zweck wird das Modell zusätzlich auf Messinvarianz untersucht (Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). Auf diese Weise werden die Vergleichbarkeit und Validität der Konstrukte gewährleistet. Die Evaluation der Messmodelle wird differenziert und detailliert in Abschnitt 6.1.7 dargestellt. Die Reliabilität des Modells wird anhand der messfehlerfreien und empirischen Varianz der latenten Faktoren geschätzt (Eid et al., 2010, S. 841).

Passung des Rasch-Modells

Zunächst werden die Messmodelle separat zu den Erhebungszeiträumen evaluiert. Anschließend wird untersucht, inwiefern sich die Funktionsweise der Items zwischen den Zeiträumen geändert hat. Der verwendete Datenkompetenztest erfasst einen spezifischen Bereich im Umgang mit Daten und Messunsicherheiten (s. Abschnitt 5.1). Den Schüler:innen sind im Prätest die Items und die benötigten Konzepte nicht bekannt. Die Intervention adressiert diese Kompetenzen ebenso wie das Konzept und Verständnis von Messunsicherheiten im Kontext vom Experimentieren. Die Lernenden erfahren durch die Intervention somit ein Treatment. Es muss daher sichergestellt werden, dass die möglichen Effekte der Intervention nicht die Funktionsweise des Testinstruments beeinflussen. Dies

wird anhand einer Differential-item-function (DIF)-Analyse durchgeführt. Diese ist Voraussetzung, um einen validen Vergleich der ermittelten Parameter zu erlauben.

Messmodell Für jeden Messzeitraum wurde anhand der Daten die Datenschwierigkeit für die Items sowie die Schüler:innenfähigkeit geschätzt. Die Berechnungen erfolgen unter Verwendung des Pakets *TAM* (Robitzsch et al., 2021) für die Programmiersprache *R* (R Core Team, 2021).

Für die Evaluation der Messmodelle wird auf die MSQ-Itemfit-Werte, die ICCs und die Wright-Map zurückgegriffen (Bond & Fox, 2007). Detaillierte Ausführungen zu diesen Schritten können in Anhang E.2 betrachtet werden.

Die Analyse der Messmodelle hat ergeben, dass eine gute Passung zwischen dem Modell und den Daten vorliegt. Es gibt einige statistische Auffälligkeiten bezüglich der Itemfit-Werte, die jedoch unterhalb gängiger Cut-off-Grenzen liegen (Bond und Fox, 2007; Wu et al., 2016). Die Wright-Maps zeigen, dass der Schwierigkeitsgrad der Items für die Tests zum Prä- und Postzeitpunkt gleichmäßig verteilt ist. Die Testinstrumente weisen Reliabiltiäten von $\rho_{\text{pre}} = 0{,}59$ und $\rho_{\text{post}} = 0{,}72$ auf.

DIF-Analyse Für den Vergleich der ermittelten Ergebnisse muss zunächst gewährleistet sein, dass die Testinstrumente sowohl vor als auch nach der Intervention gleichwertig funktionieren. Dieser Schritt ist notwendig, da die Konzepte und Kompetenzen im Umgang mit Daten und Messunsicherheiten durch die Intervention adressiert werden und sich dadurch verändert haben könnten. Dabei handelt es sich um einen Sonderfall der DIF-Analyse, da zwischen den zu vergleichenden Zeitpunkten eine Intervention stattfindet. Durch die Analyse soll sichergestellt werden, dass das gemessene Konstrukt sich nicht verändert hat und zeitlich stabil ist. Das Ergebnis der DIF-Analyse hat gezeigt, dass der Test die Kompetenzen vor und nach der Intervention gleichwertig erfasst und die beobachteten Veränderungen der Itemschwierigkeiten auf den Kompetenzzuwachs der Schüler:innen zurückzuführen sind (vgl. E.2.2). Für die weiteren Verfahren werden die Resultate der dort vorgestellten *stacked* Methode verwendet (Wright, 2003). Diese Methode erlaubt es, den Fokus der Untersuchung auf die Veränderung der Schüler:innenfähigkeit zu legen (Cunningham & Bradley, 2010).

Single-Item-Indikator in der latenten Modellierung Für den weiteren Verlauf der Analyse werden die anhand der Rasch-Analyse geschätzten Fähigkeiten der Schüler:innen als sog. *Single-item*-Indikator in der latenten Modellierung berücksichtigt. Auf diese Weise können die Ergebnisse der Rasch-Analyse in die latenten Strukturgleichungsmodelle eingebunden werden. Die Einbindung der

Schüler:innenfähigkeit in die latenten Pfadanalysen ist dahingehend wichtig, dass die Mittelwertstruktur und die Effekte messfehlerbereinigt ermittelt werden.

Die Verwendung von Single-item-Indikatoren bringt einige Nachteile mit sich. Eine übersichtliche Diskussion diesbezüglich findet sich bei Petrescu (2013, S. 102 f.). Insbesondere die Abschätzung der Residualvarianz und Faktorladung nimmt einen hohen Stellenwert in diesem Fall ein (Anderson & Gerbing, 1988). Da es sich in dem vorliegenden Fall um die auf Grundlage der Messungen ermittelten Schüler:innenfähigkeiten handelt, kann die Residualvarianz durch

$$\delta_x = \sigma_x^2(1 - \rho) \tag{6.3}$$

abgeschätzt werden (Brown, 2006, S. 137). Dabei bezeichnet δ_x die Residualvarianz des latenten Schüler:innenfähigkeitsfaktors, σ_x die Varianz der ermittelten Schüler:innenfähigkeiten und ρ die Reliabilität des *stacked*-Modells. Derart konstruirrte latente Faktoren werden auch Pseudofaktoren genannt und weisen einen perfekten Fit auf (Brown, 2006, S. 143 f.).

Zustimmung zu den Argumentkategorien

In diesem Abschnitt wird das globale Messmodell diskutiert, welches aus allen vier Messmodellen der Begründungskategorien *Intuition, Expertenwissen, Messunsicherheiten* und *Daten als Evidenz* besteht. Die ausführliche Diskussion der einzelnen Messmodelle für die Begründungskategorien kann Abschnitt E.3 entnommen werden.

Messmodell Die Güte des Messmodells kann anhand von Fit-Indizes beurteilt werden. Die Fit-Indizes für das globale Messmodell können Tabelle 6.4 entnommen werden. Es zeigt sich ein auffälliger χ^2-Wert. Dem stehen gute bis sehr gute Goodnes-of-fit-Indizes gegenüber. Das Modell weist somit trotz der Abweichung der theoretischen und datenbasierten Varianz-Kovarianzmatrix, die der χ^2-Test aufdeckt, eine gute Passung zu den Daten auf. Für den weiteren Verlauf der Analyse wird das globale Messmodell (`gloMeMo`) verwendet. Eine weitere Analyse für das globale Messmodell ist an dieser Stelle nicht notwendig, da die zugrunde liegenden Modelle der einzelnen Faktoren bereits ausführlich in Abschnitt E.3.1 diskutiert wurden.

Messinvarianz Im Laufe der weiteren Analyse sollen Mittelwerte zwischen den Erhebungszeiträumen miteinander verglichen werden. Da es sich einerseits um unterschiedliche Erhebungszeiträume handelt und um Gruppen, die im Laufe der Untersuchung eine Intervention durchleben, muss das globale Messmodell auf Messinvarianz untersucht werden. Dadurch kann gewährleistet werden, dass die Erfassung der Konstrukte und die ermittelten Parameter frei von Verzerrungen sind. Die Evaluation der Messinvarianz wurde dabei mit Verfahren nach Byrne (2011) und Geiser (2010) durchgeführt. Diese greifen auf folgende Arten der

Tab. 6.4: Fit-Indizes für das globale Messmodelle, basierend auf dem Testinstrument von Ludwig (2017). Das Messmodell weist eine gute Passung zu den Daten auf. Es kann daher von einer validen Erfassung der Konstrukte ausgegangen werden.

Modell	χ^2	df	p	CFI	RMSEA [90% CI]; p	SRMR
`gloMeMo`	988,5	698	<0,001	0,95	0,04 [0,03;0,04]; 1	0,05

Messinvarianz nach Meredith (1993) zurück:

konfigurale MI Konfigurale Messinvarianz liegt vor, wenn die Struktur der latenten Faktoren übereinstimmt. Im vorliegenden Fall ist somit darunter zu verstehen, dass zum Prä- und Posterhebungszeitpunkt die latenten Faktoren jeweils von denselben manifesten Indikatoren erfasst werden.

schwache MI Im Fall der schwachen Messinvarianz stimmen die Faktorladungen zwischen den jeweiligen latenten Faktoren zum Prä- und Postzeitpunkt überein. Nach Gleichung 6.2 bedeutet dies, dass die Messmodelle für beide Faktoren dieselben α_i aufweisen. Diese Stufe der Messinvarianz gewährleistet die Konstruktvalidität des Messmodells und stellt sicher, dass dasselbe Konstrukt zu unterschiedlichen Erhebungszeitpunkten erfasst wird.

starke MI Für die nächste Stufe der Messinvarianz werden die Achsenabschnitte der Faktorladungen für die jeweiligen Gruppen fixiert. Dieser Schritt ist notwendig, um einen Vergleich von skalaren Werten zu vollziehen.

Die Testung der verschiedenen Stufen für die Messinvarianz wird in der vorliegenden Reihenfolge realisiert. Dieses Vorgehen richtet sich nach Geiser (2010) und Byrne (2011), stellt aber nicht die einzige Möglichkeit für die Testung der Messinvarianz dar (Vandenberg & Lance, 2000). Die verschiedenen Stufen der Messinvarianz entsprechen hierarchischen Modellen, sodass beim Modellvergleich die Passung der Modelle verglichen wird (Geiser, 2010, S. 108). Für das vorliegende globale Messmodell (`gloMeMo`) konnte partielle starke Messinvarianz nachgewiesen werden (vgl. Anhang E.3.3), da für einzelne Indikatoren der latenten Faktoren die Annahme der konstanten Achsenabschnitte über die

Erhebungszeitpunkte nicht gehalten werden konnte. Dabei handelt es sich um 3 von 20 Indikatoren (Byrne et al., 1989).

Auswertungen auf Basis partieller starker Messinvarianz müssen unter großer Sorgfalt erfolgen. Einerseits kann sie zu signifikanten Abweichungen bei den beobachteten latenten Faktoren führen (Steinmetz, 2013). Andererseits muss modellspezifisch auf die Auswirkungen der partiellen starken Messinvarianz untersucht werden (Byrne et al., 1989; Millsap & Kwok, 2004). Im vorliegenden Fall führt die partielle Messinvarianz tendenziell zu einem negativen Bias, sodass die erhaltenen Schätzungen konservativer Natur sind (Hsiao & Lai, 2018; Putnick & Bornstein, 2016). Für die weiteren Analysen wird daher angenommen, dass ein Vergleich der ermittelten Mittelwerte und Effektstärken zwischen den unterschiedlichen Erhebungszeitpunkten valide ist. Auf den Einfluss der partiellen Messinvarianz wird in Abschnitt 6.3 eingegangen werden.

Reliabilität der Skalen Jede Skala des Tests zur Erfassung der Zustimmung zu den Argumentkategorien wurde hinsichtlich ihrer Reliabilität evaluiert. Die Reliabilität wird als Verhältnis der *wahren* Varianz zur *gemessenen Varianz* ermittelt (Brown, 2006, S. 338). Die Skalen wurden über kongenerische Modelle bestimmt, sodass die Reliabilität über ein kongenerisches Reliabilitätsmaß ermittelt werden muss (Cho, 2016). Für die Bestimmung können die unstandardisierten Faktorladungen λ_i und die dazugehörigen Residualvarianzen θ_i der Indikatoren verwendet werden (Brown, 2006; Cho, 2016).

$$\rho = \frac{\mathrm{Var}(T)}{\mathrm{Var}(X)} = \frac{(\Sigma\lambda_i)^2}{(\Sigma\lambda_i)^2 + \Sigma\theta_i} \tag{6.4}$$

Bei den Skalen Expertenwissen und Daten als Evidenz wurden korrelierte Residualvarianzen modelliert. Gleichung 6.4 muss dann um die entsprechenden kovariierenden Residualvarianzen erweitert werden (Brown, 2006, S. 339). Es ergibt sich die Gleichung

$$\rho = \frac{\mathrm{Var}(T)}{\mathrm{Var}(X)} = \frac{(\Sigma\lambda_i)^2}{(\Sigma\lambda_i)^2 + \Sigma\theta_i + 2\Sigma\theta_{ij}}. \tag{6.5}$$

Die Ergebnisse der Reliabilitätsbestimmung basierend auf Gleichungen 6.4 und 6.5 können in Tabelle 6.5 eingesehen werden. Mit einer Reliabilität von $0{,}85 < \rho < 0{,}93$ kann von einer zuverlässigen Messung der Skalen ausgegangen werden.

Tab. 6.5: Übersicht zur Reliabilität der Skalen

	Reliabilität ρ	
Skala	Prä	Post
Intuition	0,89	0,92
Expertenwissen	0,85	0,90
Messunsicherheiten	0,90	0,93
Daten als Evidenz	0,90	0,92

6.2 Ergebnisse

Bevor die Ergebnisse der Strukturgleichungsmodelle präsentiert werden, erfolgt eine Darstellung der Mittelwertstrukturen der latenten Resultate sowie der deskriptiven Ergebnisse für den Datenkompetenztest und den Lernerfolg. Auf diese Weise kann ein erster Eindruck bezüglich der Daten und der Stichprobe gewonnen werden. Darauf aufbauend werden die Ergebnisse der latenten Pfadanalysen präsentiert. Im Rahmen der latenten Analyse wurden vier verschiedene Modelle berechnet. Die Eigenschaften der Modelle können Tabelle 6.10 entnommen werden. Die Beschreibung und Bewertung der Güte für die jeweiligen Modelle erfolgen separat im Abschnitt E.3.3.

6.2.1 Mittelwertstruktur

Kompetenztest

Items wurden als richtig (= 1) codiert, wenn die korrekte Antwortmöglichkeit ausgewählt worden war bzw. im Fall der Multiple-Choice-Items alle Teilfragen bzw. -aussagen korrekt bewertet wurden. Andernfalls wurde das Item als falsch (= 0) codiert. Im Pretest wurden 44 % aller Fragen korrekt beantwortet, wobei die Spannweite von 6 % bis 66 % reicht. Im Posttest wurden 58 % aller Fragen korrekt beantwortet. In diesem Fall betragen der Minimalwert 20 % und der Maximalwert 80 %.

Zustimmung zu den Argumentkategorien

Die fünfstufige Likert-Skala wurde umcodiert (stimme überhaupt nicht zu = 0, stimme voll zu = 4) und für jede Argumentkategorie wurden entsprechende latente Mittelwerte gebildet. Der Mittelwert des Präzeitpunkts wurde auf

Tab. 6.6: Vergleich der Ausprägungen der latenten Faktoren zum Prä- und Postzeitpunkt. Es werden sowohl standardisierte als auch unstandardisierte Skalen dargestellt. Bei der standardisierten Darstellung wird die Ausprägung auf null fixiert.

	Prä			Post		
Faktor	Unst.	SE	Stand.	Unst.	SE	Stand.
Intuition						
Varianz	0,84	0,07	1	1,09	0,09	1
Mittelwert	0	-	0	−0,46	0,05	−0,44
Expertenwissen						
Varianz	0,54	0,06	1	0,72	0,10	1
Mittelwert	0	-	0	0,05	0,04	0,05
Messunsicherheiten						
Varianz	0,61	0,08	1	0,67	0,08	1
Mittelwert	0	-	0	0,53	0,06	0,65
Daten als Evidenz						
Varianz	0,38	0,06	1	0,50	0,06	1
Mittelwert	0	-	0	0,18	0,04	0,26

null fixiert, so dass die standardisierten Mittelwerte des Posterhebungszeitraums als Effektstärken interpretiert werden können. Eine Übersicht findet sich in Tabelle 6.6.

Es zeigt sich, dass die Zustimmung zur Argumentkategorie Expertenwissen keine Veränderung durch die Intervention erfährt. Der ermittelte Mittelwert zum Postzeitpunkt ist nicht signifikant, sodass die Nullhypothese beibehalten wird. Somit gibt es keinen Unterschied in der Zustimmung zur Argumentkategorie Expertenwissen zum Prä- und Posterhebunszeitpunkt.

Die Zustimmung zur Argumentkategorie Intuition nimmt zwischen den Erhebungszeitpunkten ab ($\mu_{\text{int,post}} = -.44$). Der ermittelte Unterschied ist substanziell und die Intervention weist einen signifikanten mittleren negativen Effekt auf die Zustimmung zur Argumentkategorie Intution auf.

Bezüglich der Zustimmung zu den Argumentkategorien Messunsicherheiten und Daten als Evidenz zeigt sich eine verstärkte Zustimmung. Die Zustimmung

erhöhte sich dabei auf $\mu_{\text{mu,post}} = 0{,}65$ für die Argumentkategorie Messunsicherheiten bzw. auf $\mu_{\text{evi,post}} = 0{,}26$. Die geschätzten Parameter sind signfikant zu Signifikanzniveau $\alpha = 5\,\%$ und es liegt ein mittlerer positiver Effekt ($d_{\text{mu}} = 0{,}53$) für die Argumentkategorie Messunsicherheiten und ein kleiner Effekt für die Zustimmung zur Argumentkategorie Daten als Evidenz vor.

Lernerfolg

Der Lernerfolg wird für die Auswertung binär codiert (richtig = 1, falsch = 0). Es zeigt sich, dass rund 27 % der Schüler:innen in der Eingangshypothese die korrekte Hypothese abgeben. Dieser Anteil steigt nach erstmaligem Betrachten der Daten auf 48 % an. Nach Bearbeitung der App sowie dem erneuten Betrachten der Daten stiegt die Anzahl der korrekten Hypothesen auf einen Anteil von rund 60 %. Der hohe Anteil an korrekten Hypothesen nach erstmaligem Betrachten der Daten könnte ein Indiz dafür sein, dass das Experiment einem gewissen Anteil der Schüler:innen bereits bekannt ist. Auf diesen Aspekt wird im späteren Verlauf dieses Abschnitts sowie im Abschnitt 6.3 genauer untersucht.

6.2.2 Beschreibung der latenten Faktoren

Veränderung der latenten Faktoren

Für den latenten Faktor Zustimmung zur Argumentkategorie Expertenwissen (Exp) konnte im Rahmen der latenten Modellierung keine Veränderung festgestellt werden. Für die Faktoren ergab sich jeweils eine signifikante Änderung. Der Faktor Zustimmung zur Argumentkategorie Intuition (Int) ging um 0,44 zurück, wohingegen die Faktoren Zustimmung zur Argumentkategorie Messunsicherheiten (MU) und Zustimmung zur Argumentkategorie Daten als Evidenz (Evi) um 0,65 bzw. 0,26 gestiegen sind. Die Beobachtungen der deskriptiven Ergebnisse lassen sich somit bestätigen und es zeigt sich, dass die Modellierung des Messfehlers zu stärkeren beobachteten Veränderungen führt.

Zusammenhänge zwischen den latenten Faktoren

Für die Darstellung der Zusammenhänge zwischen den latenten Faktoren werden die Korrelationen ausgewertet (vgl. Tabelle 6.7). Zum Präzeitpunkt zeigt sich eine negative leichte Korrelation zwischen dem Faktor Int und den Faktoren MU ($r = -0{,}22$) und Evi ($r = -0{,}17$). Die Zustimmung zur Argumentkategorie Expertenwissen weist zum Präzeitpunkt eine mittlere positive Korrelation zum Faktor Evi auf. Abschließend zeigt sich, dass die Faktoren und MU eine Korrelation von $r = 0{,}15$ im Prätest aufweisen. Alle weiteren Zusammenhänge sind nicht signifikant.

Nach der Intervention ändert sich grundsätzlich nicht die Richtung der Effekte, jedoch die Stärke. Intuition weist nun eine mittlere negative Korrelation zu den Faktoren MU ($r = -0{,}34$) und Evi ($r = -0{,}35$) auf. Zusätzlich zu dem

Tab. 6.7: Übersicht der Korrelationen zwischen den Argumentkategorien zum Präzeitpunkt (obere Dreiecksmatrix) und Postzeitpunkt (untere Dreiecksmatrix) ohne Kompetenz als Prädiktor.

	Int	Exp	MU	Evi
Int	1	0,04	$-0{,}22^*$	$-0{,}17^*$
Exp	0,04	1	0,02	$0{,}35^*$
MU	$-0{,}34^*$	$0{,}16^*$	1	$0{,}15^*$
Evi	$-0{,}35^*$	$0{,}34^*$	$0{,}47^*$	1

Zusammenhang zwischen Exp und Evi ($r = 0{,}34$) besteht nun ebenfalls ein leichter Zusammenhang zwischen Exp und MU mit $r = 0{,}16$. Der Zusammenhang zwischen den Zustimmungen zu den Kategorien MU und Evi ist deutlich gestiegen und liegt nun bei $r = 0{,}47$. Diese Korrelationen sind insbesondere bei der späteren Analyse zu beachten, da es bei starken Korrelationen zu Varianzdiebstahl kommen kann. Die Faktoren weisen dann keinen signifikanten Einfluss auf, weil die Varianz des Modells über den korrelierenden Prädiktor erläutert wird. Dieser Punkt wird im späteren Verlauf und in Abschnitt 6.3 adressiert.

6.2.3 Latente Analysen

Nachfolgend werden die Ergebnisse der latenten Analysen vorgestellt. Zu Beginn jedes Abschnitts wird das zugrunde liegende Modell erläutert und auf den Schwerpunkt der Modellierung eingegangen. Eine Übersicht der verwendeten Modelle findet sich in Tabelle 6.10.

Untersuchung von Forschungsfrage 2

Die Modelle `MoKompArg` und `MoKompArgLCSM` dienen der detaillierten Analyse zum Einfluss der Kompetenz auf die Zustimmung zu den Argumentkategorien. Während das Modell `MoKompArg` einen Schwerpunkt auf die jeweilige Betrachtung zum Prä- oder Postzeitpunkt legt, kann das Modell `MoKompArgLCSM` verwendet werden, um die Veränderung zwischen den Messzeitpunkten mit den Ausprägungen der latenten Faktoren in Beziehung zu setzen. Bei letzterem Modell handelt es sich um eine spezielle latente Modellierung, welche in der Literatur unter Latent-change-score-model (LCSM) zu finden ist (vgl. Geiser, 2010; Ghisletta & McArdle, 2012; Kievit et al., 2018; Klopack & Wickrama, 2020). Das Modell `MoKompArg` stellt das globale Messmodell `gloMeMo` dar, welches durch die Regression der Datenkompetenz auf die Zustimmung zu den Argumentkategorien zu einem Strukturgleichungsmodell erweitert worden ist.

Tab. 6.8: Darstellung der standardisierten Regressionskoeffizienten für die Datenkompetenz als Prädiktor für die Zustimmung zu den Argumentkategorien.

Prädiktor:	Datenkompetenz					
	Pre			Post		
Argumentkategorie	β	SE	R^2	β	SE	R^2
Intuition	$-0{,}35^{***}$	0,06	$0{,}12^{**}$	$-0{,}43^{***}$	0,05	$0{,}18^{***}$
Expertenwissen	0,10	0,07	0,01	0,08	0,06	0,01
Messunsicherheit	$0{,}23^{**}$	0,07	0,06	$0{,}54^{***}$	0,06	$0{,}29^{***}$
Daten als Evidenz	$0{,}26^{***}$	0,07	$0{,}07^{*}$	$0{,}38^{***}$	0,07	$0{,}15^{**}$

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Wirkung der Kompetenz auf die latente Zustimmung

Die Resultate werden jeweils für die beiden Erhebungszeitpunkte dargestellt. Eine Übersicht der Ergebnisse kann Tabelle 6.8 entnommen worden. Es zeigt sich, dass die Datenkompetenz ein signifikanter Prädiktor für die latenten Faktoren Int, MU und Evi ist. Ein Zusammenhang zwischen der Datenkompetenz und der Zustimmung zur Argumentkategorie Expertenwissen lässt sich weder im Prä- noch im Posttest erkennen.

Ermittelt wurden die standardisierten Regressionskoeffizienten β, welche als Effektstärke interpretiert werden können und dieselbe Einteilung wie das Effektstärkemaß r aufweisen. Zum Präerhebungszeitpunkt weist die Ausprägung der Kompetenz im Umgang mit Daten und Messunsicherheiten einen negativen mittleren Effekt auf den latenten Faktor Int auf. Positive kleine Effekte können bei der Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien Messunsicherheiten und Daten als Evidenz beobachtet werden.

Die Effektstärken zeigen starke Zusammenhänge zwischen der Datenkompetenz und der Ausprägung der Zustimmung zu den Argumentkategorien auf. Die Betrachtung der aufgeklärten Varianz R^2 zeigt jedoch, dass der Prädiktor Datenkompetenz nur wenig Varianz der latenten Faktoren MU und Evi erklärt. Für die Zustimmung zur Argumentkategorie Intuition können rund 12 % der Varianz aufgeklärt werden.

Zum zweiten Erhebungszeitpunkt zeigen sich ähnliche Effekte. Die Datenkompetenz wirkt mit einem mittleren negativen Effekt ($r = -0{,}43$) auf die Zustimmung zur Argumentkategorie Intuition. Der Zusammenhang zwischen der Datenkompetenz und den latenten Faktoren Evi und MU ist hingegen gestiegen. Während für die Zustimmung zur Argumentkategorie Daten als Evidenz nun als mittlerer positive Effekt ($r = 0{,}38$) vorliegt, steigt der Effekt der Datenkompetenz auf den latenten Faktor MU auf einen starken positiven Effekt ($r = 0{,}54$).

Der Anteil der aufgeklärten Varianz steigt im Posttest ebenfalls an. Datenkompetenz als Prädiktor erklärt nun 18 % (Int), 29 % (MU) und 15 % (Evi) der Varianz. Insbesondere der Anteil der aufklärten Varianzen für die Zustimmung

Tab. 6.9: Korrelation zwischen dem Achsenabschnitt $Komp_0$ und der Veränderungsrate Δ-Komp der Datenkomptenz mit den entsprechenden Größen der Zustimmung zu den Argumentkategorien.

Argumentkategorie		Kompetenz	
Kategorie	Parameter	$Komp_0$	Δ-Komp
INT	INT_0	−0,31	-
	Δ-Int	-	−0,36
EXP	EXP_0	-	-
	Δ-EXP	-	-
MU	MU_0	0,21	-
	Δ-MU	0,28	-
EVI	EVI_0	0,24	-
	Δ-EVI	-	-

zu den Argumentkategorien Daten als Evidenz und Messunsicherheiten steigt nach der Intervention stark an.

Zusammenhänge zwischen den Veränderungen der latenten Faktoren

Das Modell `MoKompArgLCSM` stellt eine longitudinale Modellierung der latenten Faktoren dar und bestimmt für die Datenkompetenz und die Zustimmung zu den Argumentkategorien jeweils einen Achsenabschnitt, dargestellt mit dem Index null, und eine Veränderungsrate vom Prä- zum Posterhebungszeitpunkt, dargestellt durch Δ und den entsprechenden Faktor.

Einerseits kann nun untersucht werden, wie sich die explizite Änderung der Kompetenz, im Allgemeinen der Zuwachs, auf die Veränderung der Zustimmung zu den Argumentkategorien auswirkt. Andererseits erlaubt das Modell Rückschlüsse darüber, wie die Veränderungen der latenten Faktoren Int, MU und Evi untereinander zusammenhängen. Der Faktor Exp wurde aufgrund der bisherigen vorliegenden Resultate nicht weiter im Modell berücksichtigt.

Die Ergebnisse für die Zusammenhänge zwischen der Datenkompetenz und den latenten Faktoren für die Zustimmung zu den Argumentkategorien können Tabelle 6.9 entnommen werden. Die Zusammenhänge zwischen dem Achsenabschnitt $Komp_0$ und den Achsenabschnitten der Zustimmung zu den Argumentkategorien zeigen ein Ergebnis, welches den Resultaten des vorigen Abschnitts ähnelt. Es wird an dieser Stelle daher nicht weiter ausgeführt. Zusätzlich zeigt sich nun ein Zusammenhang zwischen dem Achsenabschnitt der Datenkompetenz und dem Zuwachs des latenten Faktors MU, der den Zuwachs der Zustimmung zur Argumentkategorie Messunsicherheit schätzt. Es zeigt sich ein mittlerer positive Zusammenhang der beiden Parameter. Das bedeutet, dass eine hohe Ausprägung der Kompetenz zum Präzeitpunkt mit einem größeren Zuwachs Δ-MU zusammenhängt.

Außerdem zeigt sich, dass die Veränderungen Δ-Komp der Datenkompetenz sowie Δ-Int der Zustimmung zur Argumentkategorie Intuition negativ miteinander korrelieren. Je größer der Zuwachs an Kompetenzen ist, desto kleiner wird die Veränderung des Faktors Int. An dieser Stelle ist zu beachten, dass Δ-Int ein negatives Vorzeichen aufweist und somit eine Verkleinerung dieses Parameters einer stärkeren Abnahme der Zustimmung zur Argumentkategorie Intuition entspricht.

Tab. 6.10: Übersicht der verwendeten Modelle für die latente Analyse. Jedes Modell ist im Anhang E.4 detailliert aufgeführt.

Modell	Forschungsfrage
`MoKompArg` Strukturgleichungsmodell zum Prä- und Postzeitpunkt zum Einfluss der Kompetenz auf die Zustimmung zu den Argumentkategorien	FF2
`MoKompArgLCSM` Strukturgleichungsmodell zum Einfluss der Kompetenz auf die Zustimmung zu den Argumentkategorie mit Schwerpunkt auf der Änderung der jeweiligen Ausprägungen	FF2
`MoArgHyp` Strukturgleichungsmodell zum Einfluss der Zustimmung zu den Argumentkategorien auf den Lernerfolg	FF3
`MoKompArgHyp` Strukturgleichungsmodell zum Einfluss der Datenkompetenz auf den Lernerfolg	FF4

Einfluss der Zustimmung zu den Argumentkategorien auf den Lernerfolg

Um zu untersuchen, welchen Einfluss die Zustimmung zu den Argumentkategorien auf den Lernerfolg hat, wurde eine logistische Regression durchgeführt. Der Lernerfolg ist eine dichotome Variable und daher kategorial und nicht kontinuierlich. Es ist nicht möglich, Werte zwischen richtiger und falscher Hypothese anzunehmen. Durch die logistische Regression kann evaluiert werden, wie sich eine höhere Zustimmung in den Argumentkategorien auf die Chance für den Lernerfolg, hier die richtige Hypothese, auswirkt. Das Modell `MoKompArg` wurde um den endogenen latetenten Faktor der Hypothese für die jeweiligen Erhebungszeitpunkte ergänzt. Zusätzlich wurde eine multiple logistische Regression in das Strukturgleichungsmodell eingefügt, die die Zustimmung zu den Argumentationskategorien in Form der latenten Faktoren Int, Exp, MU und Evi auf

den Lernerfolg regregiert. Das dazugehörige Modell wird im weiteren Verlauf der Arbeit als `MoArgHyp` bezeichnet.

Die Bedeutung der Prädiktoren kann anhand des Odds Ratio evaluiert werden (vgl. Anhang E). Das Odds Ratio stellt einen multiplikativen Faktor dar, der angibt, wie sich die Chancen auf den Lernerfolg ändern, wenn die Ausprägung eines Prädiktors unter Konstanthaltung der restlichen Prädiktoren um eins erhöht wird. Das Odds Ratio entspricht somit dem Effektstärkemaß der logistischen Regression.

Im Nachfolgenden wird die Notation R^2 verwendet, um zu evaluieren, wie gut die Prädiktoren bzw. die zugrunde liegenden Modelle die abhängige Variable erklären. Dabei handelt es sich jedoch nicht um das R^2 der linearen Regression, sondern um das Pseudo-R^2, welches verwendet werden kann, um die aufgeklärte Varianz bei logistischen Regressionen zu evaluieren. Die vorliegende Arbeit verwendet das Pseudo-R^2 von McKelvey und Zavoina (1975), welches wie das gewöhnliche R^2 interpretiert werden kann (Veall & Zimmermann, 1994). Die Evaluation der aufgeklärten Varianz erfolgt daher nach J. Cohen (1988).

Die Ergebnisse der Regression können den Tabellen 6.11 und 6.12 entnommen werden. Letztere Tabelle beinhaltet die Kompetenz als Prädiktor für die Zustimmung zu den Argumentkategorien. Auf diese Weise wird die zur Datenkompetenz gehörige Varianz bezüglich der Hypothese herauspartialisiert. Ein Vergleich der Resultate kann Aufschluss darüber geben, ob starke Korrelationen untereinander zu Varianzdiebstahl führen.

Es zeigt sich, dass die Zustimmung zur Argumentkategorie Messunsicherheiten sowohl zum Prä- ($\beta = 1{,}02, p < 0.05$) als auch zum Postzeitpunkt ($\beta = 0{,}70, p < 0{,}05$) einen signifikanten Einfluss auf den Lernerfolg besitzt. Die Zustimmung zur Argumentkategorie Intuition weist im Posttest einen kleinen negativen Effekt auf den Lernerfolg auf ($\beta = -0{,}63, p < 0{,}05$), klärt jedoch keine nennenswerte Varianz auf (R^2). Die Zustimmung zu den übrigen Argumentkategorien Intuition, Expertenwissen und Daten als Evidenz weist keinen signifikanten Einfluss auf den Lernerfolg auf und klärt keinen wesentlichen Anteil der Varianz auf.

Ein ähnliches Bild zeigt sich im mehrfaktoriellen Modell. Die Regressionsparameter für die Zustimmung zu den Argumentkategorien sind im Prä- und Posttest nur für Messunsicherheiten signifikant (pre: $\beta = 0{,}96, p < 0{,}05$, post: $\beta = 0{,}70, p < 0{,}05$). Der signifikante Einfluss der Zustimmung zur Argumentkategorie Intuition ist nun nicht mehr zu sehen. Die Zustimmung zu den übrigen Kategorien weist weder im Prä- noch im Posttest einen Einfluss auf den Lernerfolg auf.

Insbesondere fällt an dieser Stelle auf, dass die aufgeklärte Varianz des gesamten Modells nahezu deckungsgleich mit der aufgeklärten Varianz des einfaktoriellen Modells Messunsicherheiten ist. Die übrigen Argumentkategorien tragen somit bei der Erklärung der Varianz nur einen sehr geringen Beitrag

Tab. 6.11: Einfluss der Zustimmung zu den Argumentkategorien auf die Richtigkeit der Hypothese. Dargestellt sind die Parameter der logistischen Regression. Der Einfluss wurde separat und in einem kombinierten Modell untersucht.

	Prä				Post			
Modelle	β	SE	OR [95 % CI]	R^2	β	SE	OR [95 % CI]	R^2
einfaktoriell								
Int	−0,19	0,14	0,83 [0,64; 1,08]	0,01	−0,36*	0,12	0,70 [0,55; 0,88]	0,04
Exp	−0,02	0,18	0,98 [0,69; 1,41]	0,00	−0,02	0,15	0,97 [0,73; 0,1,32]	0,00
MU	1,02*	0,21	2,77 [1,84; 4,2]	0,16*	0,70*	0,17	2,02 [1,45; 2,79]	0,09*
Evi	0,49	0,31	1,64 [0,89; 3,01]	0,02	0,42	0,24	1,52 [0,94; 2,44]	0,01
mehrfaktoriell								
Int	0,03	0,15	1,03 [0,77; 1,38]	0,16	−0,23	0,14	0,80 [0,61; 1,04]	0,11
Exp	−0,12	0,22	0,89 [0,58; 1,35]		−0,06	0,18	0,94 [0,67; 1,33]	
MU	0,96*	0,21	2,62 [1,72; 3,99]		0,70*	0,19	2,01 [1,39; 2,91]	
Evi	0,26	0,25	1,29 [0,79; 2,10]		−0,22	0,23	0,80 [0,51; 1,25]	

Tab. 6.12: Einfluss der Zustimmung zu den Argumentkategorien auf die Richtigkeit der Hypothese. Dargestellt sind die Parameter der logistischen Regression. Der Einfluss wurde separat und in einem kombinierten Modell untersucht. Zusätzlich wird der Einfluss der Datenkompetenz auf die Zustimmung zu den Argumentkategorien berücksichtigt.

	Prä				Post			
Modelle	β	SE	OR [95 % CI]	R^2	β	SE	OR [95 % CI]	R^2
einfaktoriell komp→ArgKat								
Int	−0,17	0,14	0,83 [0,65; 1,10]	0,01	−0,37 *	0,13	0,70 [0,54; 0,87]	0,04
Exp	−0,03	0,18	0,97 [0,68; 1,38]	0,00	−0,02	0,15	0,98 [0,73; 1,32]	0,00
MU	1,06	0,21	2,87 [1,90; 4,35]	0,17*	0,67	0,16	2,00 [1,42; 2,69]	0,08*
Evi	0,44*	0,22	1,56 [1,01; 2,41]	0,02	0,29	0,18	1,33 [0,93; 1,91]	0,01
mehrfaktoriell komp→ArgKat								
Int	0,03	0,15	1,03 [0,77; 1,38]	0,16	−0,25	0,14	0,78 [0,60; 1,01]	0,10
Exp	−0,12	0,22	0,89 [0,58; 1,35]		−0,07	0,17	0,94 [0,67; 1,30]	
MU	0,96*	0,21	2,62 [1,72; 3,99]		0,63*	0,18	1,88 [1,31; 2,70]	
Evi	0,26	0,25	1,29 [0,79; 2,10]		−0,17	0,21	0,85 [0,56; 1,28]	

Tab. 6.13: Einfluss der Datenkompetenz auf die Richtigkeit der Hypothese. Dargestellt sind die Parameter der logistischen Regression. Der Einfluss wurde zunächst selbstständig und anschließend im Rahmen einer multiplen Analyse unter Berücksichtigung der Zustimmung zu den Argumentkategorien untersucht.

	Prä				Post			
Modelle	β	SE	OR [95 % CI]	R^2	β	SE	OR [95 % CI]	R^2
einfaktoriell								
Komp	0,46*	0,22	1,58 [1,02; 2,45]	0,03	0,86*	0,20	2,37 [1,60; 3,49]	0,14*
mehrfaktoriell								
Int	0,05	0,18	1,05 [0,75; 1,48]	0,18*	−0,05	0,15	0,96 [0,71; 1,29]	0,16*
Exp	−0,17	0,22	0,84 [0,55; 1,29]		−0,11	0,17	0,90 [0,64; 1,27]	
MU	1,02*	0,23	2,78 [1,77; 4,38]		0,24	0,23	1,27 [0,80; 2,00]	
Evi	0,34	0,27	1,41 [0,84;2,37]		−0,40	0,23	0,67 [0,43; 1,05]	
Komp	0,05	0,31	1,05 [0,57; 1,92]		0,87*	0,34	2,38 [1,23; 4,61]	

bei. Insgesamt besitzt das Modell zum Präzeitpunkt eine moderate und zum Postzeitpunkt eine geringe Varianzaufklärung (J. Cohen, 1988).

Auffällig ist außerdem, dass die Effektstärke für die Zustimmung zur Argumentkategorie Messunsicherheiten im Prätest stärker ausfällt und der Anteil der aufgeklärten Varianz durch die Zustimmung zu den Argumentkategorien im Posttest deutlich geringer ist als im Prätest.

Die Hinzunahme der Kompetenz als Prädiktor für die Zustimmung zu den Argumentkategorien weist nur kleine Effekte auf. In den einfaktoriellen Modellen zeigt sich, dass die aufgeklärte Varianz im Modell Messunsicherheiten im Postzeitraum zurückgeht. Zudem geht die aufgeklärte Varianz des mehrfaktoriellen Modells zum Postzeitpunkt um 0,01 zurück.

Einfluss der Datenkompetenz auf den Lernerfolg

Abschließend wird untersucht, welchen direkten Einfluss die Datenkompetenz auf den Lernerfolg aufweist, nachdem durch die vorangagenen Abschnitte eine indirekte Wirkung untersucht worden ist.

Dazu wird die multpiple logistische Regression des Modells `MoHypArg` um den latenten Faktor der Datenkompetenz ergänzt. Dieses Modell wird im Folgenden mit `MoKompArgHyp` bezeichnet (vgl. Tab. 6.10). Der Einfluss der Datenkompetenz wird aus den zuvor genannten Gründen ebenfalls mit einer logistischen Regression untersucht. Dabei wird zunächst mit einfaktoriellen Modellen betrachtet, welchen Einfluss der Prädiktor Kompetenz aufweist, bevor ein mehrfaktorielles Modell mit der Datenkompetenz und der Zustimmung zu den Argumentkategorien eingeführt wird. Die Ergebnisse können Tabelle 6.13 entnommen werden.

Für das einfaktorielle Modell zeigt sich, dass der Prädiktor Datenkompetenz sowohl vor ($\beta = 0{,}46, p < 0{,}05$) als auch nach ($\beta = 0{,}86, p < 0{,}05$) der Intervention ein signifikanter Prädiktor für den Lernerfolg ist. Dabei handelt es sich zum Präzeitpunkt um einen kleinen positiven Effekt und zum Postzeitpunkt um einen moderaten positiven Effekt. Ein großer Unterschied zeigt sich bezüglich der aufgeklärten Varianz. Zum Präzeitpunkt kann durch den Faktor Kompetenz kaum Varianz aufgeklärt werden ($R^2 = 0{,}03$), während das einfaktorielle Modell im Posttest eine moderate Varianzaufklärung aufweist.

Die Hinzunahme des Prädiktors Komp in das mehrfaktorielle Modell ändert an dem Resultat der logistischen Regression für den Präzeitpunkt wenig. Die Zustimmung zur Argumentkategorie Messunsicherheit ist weiterhin ein signifikanter Prädiktor $\beta = 1{,}02, p < 0{,}05$ und weist einen mittleren positiven Effekt auf den Lernerfolg auf (OR = 2,78), während die Zustimmung zu den übrigen Argumentkategorien sowie die Datenkompetenz keinen signfikanten Einfluss besitzen. Die aufgeklärte Varianz bleibt mit $R^2_{\text{pre}} = 0{,}18$ im Vergleich zum multifaktoriellen Modell zur Zustimmung der Argumentkategorien unverändert für den Präzeitpunkt.

Zur Posterhebung zeigt sich, dass die Zustimmung zu den Argumentkategorien keinen signifikanten Einfluss auf den Lernerfolg auf den Lernerfolg aufweist.

Der latente Faktor Komp weist nun einen positiven mittleren Effekt auf den Lernerfolg auf ($\beta = 0{,}87, p < 0{,}05$, OR = 2,38). Die Hinzunahme dieses Prädiktors hat die aufgeklärte Varianz um 6 % erhöht. Das mehrfaktorielle Modell weist somit zum Postzeitpunkt eine moderate Varianzaufklärung auf.

6.2.4 Wirkung der Datenkompetenz auf den Lernerfolg

Die FF4 untersucht den Einfluss der Datenkompetenz auf die Richtigkeit der Hypothese. Zu diesem Zweck wurde in der Analyse in ein- und mehrfaktoriellen Modellen mit der Zustimmung zu den Argumentkategorien der Einfluss auf den Lernerfolg untersucht. Die Ergebnisse können in Tabelle 6.13 eingesehen werden.

Die Ergebnisse zeigen, dass die Datenkompetenz einen signifikanten Effekt auf die Richtigkeit der Hypothese aufweist. Im Prätest handelt es sich um einen kleinen positiven Effekt (OR = 1,58, $p < 0{,}05$) und im Posttest um einen mittleren positiven Effekt (OR = 2,37, $p < 0{,}05$). In diesem Rahmen muss somit die Nullhypothese verworfen werden und für den weiteren Verlauf wird von der Gültigkeit von Forschungshypothese H^4 ausgegangen. Das einfaktorielle Modell weist im Prätest kaum Varianzaufklärung auf ($R^2 = 0{,}03$), während es im Posttest über eine mittlere Varianzaufklärung verfügt ($R^2 = 0{,}14$). Dieser Wert ist im Vergleich zum vorigen Abschnitt der höchste Wert, der im Rahmen von einfaktoriellen Modellen mit der Hypothese als abhängige Variable erzielt werden kann. Die Datenkompetenz ist somit nach der Intervention der aussagekräftigste Prädiktor für die Richtigkeit der Hypothese.

Im mehrfaktoriellen Modell mit der Zustimmung zu den Argumentkategorien zeigen sich etwas andere Einflüsse der Datenkompetenz auf die Richtigkeit der Hypothese. Im Prätest gibt es keinen signifikanten Zusammenhang zwischen der Datenkompetenz und der Richtigkeit der Hypothese ($\beta_{\text{pre}} = 0{,}05$, $p > 0{,}05$). Dieser Zusammenhang ändert sich im Posttest, wo die Datenkompetenz einen mittleren positiven Effekt auf den Lernerfolg aufweist ($\beta_{\text{post}} = 0{,}87$, $p < 0{,}05$, OR = 2,38). Auffällig ist, dass gleichzeitig die Effektstärke der Zustimmung zur Argumentkategorie Messunsicherheiten abnimmt und nun nicht mehr von einem Nulleffekt unterscheidbar ist. Dies deutet auf geteilte Varianz der Prädiktoren Zustimmung zur Argumentkategorie Messunsicherheiten und Datenkompetenz hin.

Bezüglich der aufklärten Varianz weist das mehrfaktorielle Modell sowohl im Prä- ($R^2 = 0{,}18$) als auch im Posttest ($R^2 = 0{,}16$) eine mittlere Varianzaufklärung auf. Im Vergleich zum mehrfaktoriellen Modell mit der Zustimmung zu den Argumentkategorien zeigt sich eine Zunahme der aufgeklärten Varianz. Besonders stark ist die Zunahme im Posttest zu sehen.

Im einfaktoriellen Modell kann die Datenkompetenz als Prädiktor 14 % der Varianz des Lernerfolgs erklären. Im mehrfaktoriellen Modell gemeinsam

mit der Zustimmung zu den Argumentkategorien liegt der Anteil jedoch bei 16 %. Dieser Anteil ist geringer als erwartet, da das mehrfaktorielle Modell der Argumentkategorien bereits 10 % aufklärt. Der Zuwachs an aufgeklärter Varianz könnte ein weiteres Indiz für eine gemeinsame Varianz der Prädiktoren Zustimmung zur Argumentkategorie Messunsicherheiten und Datenkompetenz sein. Dafür spricht ebenfalls, dass sich die Ergebnisse der Analysen in den Tabellen 6.11 und 6.12 kaum verändert haben.

Auf Grundlage dieser Diskussion wird eine Post-hoc-Analyse durchgeführt. Diese hat zum Ziel, Ursachen für den geringer werdenden Einfluss der Zustimmung zu den Argumentkategorien auf den Lernerfolg und die abnehmende aufgeklärte Varianz im Posttest aufzudecken.

Im Vergleich zur Arbeit von Ludwig (2017) zeigt sich, dass die Lernenden in der vorliegenden Arbeit bereits zum Nullzeitpunkt, das heißt bei der erstmaligen Erfassung der Hypothese vor der ersten Auswertung der Daten, zu 27 % bereits die richtige Hypothese ausgewählt haben. Dieser Anteil betrug bei Ludwig (2017) nur 12 % und bei Donner (2019) ebenfalls 27 %. Auffällig ist jedoch, dass nach der ersten Auswertung der Daten der Anteil der richtigen Hypothesen unter den Lernenden in der vorliegenden Arbeit steigt, während der Anteil bei Donner (2019) weiter zurückgeht. Das könnte ein weiterer Beleg für die unterschiedlichen Eingangsvoraussetzungen der Lernenden für diese Studie sein. Bei Ludwig steigt der Anteil der richtigen Hypothesen nach dem Experimentieren auf 54 % und ist damit in ähnlichen Bereichen wie nach der Intervention in der vorliegenden Arbeit (60 %) und bei Donner (2019) (54 %). Aus diesem Grund wird der Einfluss der Richtigkeit der Hypothese auf die Wirkung der Kompetenzen in der Post-hoc-Analyse ebenfalls aufgegriffen.

6.2.5 Post-hoc-Analyse

Die offenen Fragen aus dem vorigen Abschnitt werden mit neuen Analysen ausgewertet. Zunächst wird die Analyse bezüglich Forschungsfrage FF2 um eine Gruppenanalyse erweitert. Es wird verglichen, inwieweit sich die Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien ändert, wenn die Proband:innen bezüglich ihrer Prähypothese differenziert werden.

Darauf aufbauend wird untersucht, wie sich die Zustimmung zur Argumentkategorie Messunsicherheiten im Laufe der Studie entwickelt hat. Die Ergebnisse aus Abschnitt 6.2 deuten darauf hin, dass die Zustimmung zur Argumentkategorie Messunsicherheiten an Aussagekraft bezüglich der Richtigkeit der Hypothese verliert. Um diesen Effekt genauer zu untersuchen, wird eine Gruppenenalyse durchgeführt, die für jeden Erhebungszeitpunkt der Hypothese eine Differenzierung der Stichprobe basierend auf der Hypothese generiert. Auf diese Weise kann für drei unterschiedliche Konstellationen von Gruppen mit richtiger bzw. falscher

Hypothese die Zustimmung zur Argumentkategorie Messunsicherheiten evaluiert werden. Das Ziel der Analyse ist es, aufzudecken, warum die Varianzaufklärung des mehrfaktoriellen Modells nach der Intervention abnimmt bzw. selbst unter Zunahme der Datenkompetenz unter dem Anteil aufgeklärter Varianz aus dem Präzeitpunkt bleibt.

Abschließend wird die Untersuchung auf alle Argumentkategorien erweitert. Es erfolgt eine erneute Untersuchung von Forschungsfrage FF3, wobei hier zunächst alle Proband:innen aus der Analyse entfernt werden, welche die korrekte Eingangshypothese angeben. Es wird davon ausgegangen, dass diese Schüler:innen den Versuch bereits kennen. Schwerpunkt der vorliegenden Studie ist, zu untersuchen, wie sich Schüler:innen mit Daten in einem unbekannten Kontext auseinandersetzen. Für Proband:innen, welche bereits zum ersten Erhebungszeitpunkt die korrekte Hypothese angeben, kann diese Tatsache nicht gewährleistet werden. Zusammenfassend dient die Post-hoc-Analyse der Untersuchung der folgenden Forschungsfragen:

1. Inwiefern beeinflusst die Wahl der Prähypothese die Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien?

2. Kann anhand der Unterscheidung zwischen richtiger und falscher Hypothese die unterschiedliche Wirkung der Zustimmung zur Argumentkategorie Messunsicherheiten auf die Richtigkeit der Hypothese erklärt werden?

3. Inwiefern unterscheidet sich die Wirkung der Zustimmung zur Argumentkategorie Messunsicherheiten auf die Richtigkeit der Hypothese in Abhängigkeit der zuvor gewählten fachlichen Hypothese?

Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien in Abhängigkeit der Richtigkeit der Hypothese

Für diese Analyse wird die Stichprobe anhand der Hypothesen zum Präzeitpunkt in zwei Gruppen aufgeteilt. Anschließend wird die Analyse zur Forschungsfrage FF2 wiederholt. Die Ergebnisse der Analyse sind in Tabelle 6.14 dargestellt.

Die Ergebnisse zeigen, dass die Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien im Prätest zwischen den Gruppen unterschiedlich ausfällt. Die Datenkompetenz beeinflusst die Zustimmung zur Argumentkategorie Intuition für beide Gruppen gleich und weist einen mittleren negativen Effekt auf ebendiese Zustimmung auf ($\beta_{\text{pre}} = -0{,}33$, $p < 0{,}05$ bzw. $\beta_{\text{pre}} = -0{,}38$, $p < 0{,}05$). Für die Proband:innen, die die richtige Hypothese zum Präzeitpunkt gewählt haben, ist kein weiterer Effekt zu beobachten. Für die Proband:innen mit falscher Hypothese zeigt sich, dass die Datenkompetenz mit einem kleinen positiven Effekt auf die Zustimmung zur Argumentkategorie Expertenwissen ($\beta_{\text{pre}} = 0{,}21$, $p < 0{,}05$) und die Zustimmung zur Argumentkategorie Messunsicherheiten ($\beta_{\text{pre}} = 0{,}29$, $p < 0{,}05$) sowie mit einem moderaten

Tab. 6.14: Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien vor und nach der Intervention aufgeteilt in Abhängigkeit der zum Präzeitpunkt abgegebenen Hypothese.

Prädiktor:	Datenkompetenz					
	Pre			Post		
Falsche Prähypothese	β	SE	R^2	β	SE	R^2
Intuition	$-0{,}33^{***}$	0,08	$0{,}11^{*}$	$-0{,}47^{***}$	0,07	$0{,}22^{***}$
Expertenwissen	0,21	0,08	0,05	0,09	0,09	0,01
Messunsicherheit	$0{,}29^{*}$	0,11	0,08	$0{,}59^{***}$	0,05	$0{,}34^{***}$
Daten als Evidenz	$0{,}41^{***}$	0,09	$0{,}17^{*}$	$0{,}46^{***}$	0,07	$0{,}21^{**}$
Richtige Prähypothese	β	SE	R^2	β	SE	R^2
Intuition	$-0{,}38^{***}$	0,08	$0{,}14^{*}$	$-0{,}38^{***}$	0,07	$0{,}14^{**}$
Expertenwissen	$-0{,}02$	0,10	0,00	0,09	0,09	0,01
Messunsicherheit	0,19	0,12	0,04	$0{,}50^{***}$	0,08	$0{,}25^{**}$
Daten als Evidenz	0,09	0,09	0,01	$0{,}30^{**}$	0,11	0,09

$^{*}p < 0,05$ $^{**}p < 0,01$ $^{***}p < 0,001$

positiven Effekt auf die Zustimmung zur Argumentkategorie Daten als Evidenz ($\beta_{\text{pre}} = 0{,}41$, $p < 0{,}05$) wirkt.

Im Posttest zeigt sich, dass die Datenkompetenz auf beide Proband:innengruppen gleich wirkt. Im Detail gibt es Unterschiede in den standardisierten Regressionskoeffizienten, jedoch lassen sich die Regressionkoeffizienten für die Zustimmung zu den Argumentkategorien bezüglich der Effektstärken gleich charakterisieren. Die Datenkompetenz weist einen mittleren negativen Effekt auf die Zustimmung zur Argumentkategorie Intuition auf ($\beta_{\text{post}} = -0{,}47$, $p < 0{,}05$ vs. $\beta_{\text{post}} = -0{,}38$, $p < 0{,}05$), einen großen positiven Effekt auf die Zustimmung zur Argumentkategorie Messunsicherheiten ($\beta_{\text{post}} = 0{,}59$, $p < 0{,}05$ vs. $\beta_{\text{post}} = 0{,}50$, $p < 0{,}05$) und einen moderaten positiven Effekt auf Zustimmung zur Argumentkategorie Daten als Evidenz ($\beta_{\text{post}} = 0{,}46$, $p < 0{,}05$ vs $\beta_{\text{post}} = 0{,}30$, $p < 0{,}05$).

Folgerung

Der erste Teil der Post-hoc-Analyse hat gezeigt, dass es vor der Intervention einen Unterschied bezüglich der Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien in Abhängigkeit der Richtigkeit der Hypothese der Lernenden gibt. Die Datenkompetenz erklärt für die Proband:innen mit falscher Hypothese deutlich mehr Varianz.

Weiterführend hat die Analyse gezeigt, dass dieser Unterschied nach der Intervention nicht mehr zu erkennen ist. Nach der Intervention wirkt die Datenkompetenz in beiden Teilgruppen gleich auf die Zustimmung zu den Argumentkategorien und deckt sich mit den Ergebnissen aus Abschnitt 6.2.

An dieser Stelle wird somit erkenntlich, dass alle Proband:innen von der Intervention gleich profitieren. Es macht in dieser Hinsicht keinen Unterschied, ob die Schüler:innen sich hinsichtlich ihrer Leistung im Lernerfolg unterscheiden oder nicht. Dieser Effekt ist überraschend, da sich im Allgemeinen das Vorwissen von Lernenden positiv auf die Wirkung von Lernangeboten auswirken kann (bspw. Bos, 2003). Die unterschiedlichen Eingangsvoraussetzungen der Stichproben in der Arbeit von Donner (2019) und des Autors gegenüber der Arbeit von Ludwig (2017) haben somit keinen Einfluss auf die Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien nach der Intervention.

Analyse der Argumentkategorie Messunsicherheiten

Die beobachteten Ausprägungen der Argumentkategorie Messunsicherheiten wurden sowohl deskriptiv als auch inferenzstatistisch miteinander verglichen. Diese Verfahren unterliegen zwar den in Abschnitt 6.1.7 dargestellten Schwächen, da jedoch Werte gleicher Erhebungszeitpunkte miteinander verglichen werden, sind die entsprechenden Einflüsse gleich zu bemessen.

Für die Analyse wird die Stichprobe in drei unterschiedlichen Gruppierungen untersucht. Es findet eine Zweiteilung auf Grundlage der Eingangshypothese, der Prähypothese und der Posthypothese statt. Für jede Teilanalyse gibt es somit zwei Gruppen (richtige und falsche Hypothese). Eine Gegenüberstellung der Resultate kann Tabelle 6.15 entnommen werden.

Grundsätzlich zeigt sich für alle Gruppierungen, dass die Zustimmung zur Argumentkategorie Messunsicherheiten vom Prä- zum Postzeitpunkt steigt. Vergleicht man jedoch die Gruppen Eingangs- und Prähypothese zur Posthypothese, fällt auf, dass die ersten beiden Gruppen zum Postzeitpunkt nicht mehr anhand der Zustimmung zur Argumentkategorie differenziert werden können. Durch die Intervention hat sich sowohl für die Proband:innen sowohl mit richtiger als auch falscher Hypothese die Zustimmung zur Argumentkategorie Messunsicherheiten erhöht, wobei der Effekt bei Proband:innen mit falscher Hypothese stärker ausfällt. Beide Teilgruppen stimmen nun in gleichem Maße zur Argumentkategorie Messunsicherheiten zu.

Die Gruppierung anhand der Hypothese zum Postzeitpunkt zeigt, dass es zwischen den Proband:innen mit richtiger und falscher Hypothese in der Zustimmung zur Argumentkategorie Messunsicherheit sowohl im Prä- als auch im Postzeitpunkt signifikante Unterschiede gibt. Der Unterschied besteht bereits im Prätest, wird jedoch zum Postzeitpunkt geringer.

Tab. 6.15: Darstellung der Zustimmung zur Argumentkategorie Messunsicherheiten für verschiedene Gruppierungen und Erhebungszeiträume. Auffällig ist, dass für über alle Gruppierungen hinweg die durchschnittliche Zustimmung zur Argumentkategorie für Proband:innen mit richtiger Hypothese gleich ausfällt.

	Prezeitpunkt		Postzeitpunkt	
Gruppierung	Punktzahl	t-Statistik	Punktzahl	t-Statistik
Eingangshyp				
falsch	2,05	$t(168{,}7) = -2{,}12$	2,54	$t(170{,}77) = -0{,}31$
richtig	2,25	$p < 0{,}04$	2,57	$p = 0{,}76$
Prähyp				
falsch	1,89	$t(322) = -4{,}92$	2,54	$t(323) = -0{,}29$
richtig	2,32	$p < 0{,}0001$	2,56	$p = 0{,}77$
Posthyp				
falsch	1,84	$t(274{,}7) = -4{,}76$	2,41	$t(250{,}96) = -2{,}58$
richtig	2,28	$p < 0{,}001$	2,64	$p = 0{,}01$

Folgerung

Anhand der vorangehenden Ergebnisse lässt sich nun erklären, wie es zum Rückgang der Effektstärke im Posttest kommt. Sowohl bei Schüler:innen mit richtiger als auch mit falscher Hypothese kommt es durch die Intervention zu einem Zuwachs in der Zustimmung zur Argumentkategorie Messunsicherheit. Der Zuwachs für Proband:innen mit falscher Hypothese fällt in allen Fällen stärker aus als für Proband:innen mit richtiger Hypothese und ist somit nicht unbedingt an den Lernerfolg gekoppelt. Ferner hervorzuheben sind die Ausprägungen anhand der Gruppierung Prähypothese zum Präzeitpunkt und die Gruppierung Posthypothese zum Postzeitpunkt. In beiden Fällen weisen die Proband:innen mit richtiger Hypothese eine höhere Zustimmung zur Argumentkategorie Messunsicherheit auf, welche den positiven Effekt der Zustimmung zu dieser Argumentkategorie auf den Lerneffekt erklärt. Der Unterschied zwischen den Proband:innen mit richtiger und falscher Hypothese ist im Postzeitpunkt jedoch geringer, was zu einer geringeren Effektstärke dieser Kategorie auf den Lernerfolg führt.

Einfluss der Zustimmung zur Argumentkategorie Messunsicherheiten in Abhängigkeit der zuvor gewählten Hypothese

Es kann beobachtet werden, dass die Zustimmung für die Argumentkategorie Messunsicherheiten ebenfalls für Proband:innen gesteigert werden kann, die bereits zum Präzeitpunkt die korrekte Hypothese gewählt haben. Durch die Förderung der Datenkompetenz kann die Zustimmung zu datenbasierten Argumentkategorien für Schüler:innen gesteigert werden, die bereits im Prätest eine

Tab. 6.16: Einfluss der Zustimmung zur Argumentkategorie Messunsicherheiten auf den Lernerfolg. Gruppiert wurde danach, ob die Proband:innen zu den verschiedenen Erhebungszeiträumen die Hypothese gewechselt haben oder nicht.

	Prehyp		Posthyp		Hyp-wechsel	
	OR	95% CI	OR	95% CI	OR	95% CI
Hyp gewechselt						
nein	3,4***	[2,1; 5,5]	1,7**	[1,2; 2,5]	-	-
ja	1,2	[0,4; 3,7]	3,1*	[1,1; 8,3]	1,8*	[1,1; 2,8]

$^{*}p < 0,05$ $^{**}p < 0,01$ $^{***}p < 0,001$

überdurchschnittlich hohe Zustimmung zur Argumentkategorie Messunsicherheiten aufweisen und die korrekte Hypothese aufgestellt haben.

Andererseits stellt sich die Frage, inwiefern die Schüler:innen, die bereits zum Präzeitpunkt die richtige Hypothese aufweisen, ihre Entscheidung im Posttest anhand der Daten begründen. Wird die Studie über alle Proband:innen ausgewertet, kann der Aspekt nicht vollständig außen vor gelassen werden, da die Zustimmung zur Argumentkategorie Messunsicherheiten unabhängig von der Richtigkeit der Hypothesen durch die Förderung der Datenkompetenz gesteigert werden kann. Aus den vorangehend genannten Überlegungen werden sowohl für den Prä- als auch für den Postzeitpunkt Odds Ratios für Probanden berechnet, die ihre Prähypothese verwerfen und eine neue Hypothese wählen. Dabei wird nicht unterschieden, ob die Schüler:innen von einer falschen zu einer richtigen Hypothese wechseln oder umgekehrt.

Tabelle 6.16 zeigt das Odds Ratio als Effektmaß für die Ergebnisse der logistischen Regression. Die Resultate zeigen, dass die Zustimmung zur Argumentkategorie Messunsicherheiten für beide Gruppen unterschiedliche Effekte zu den Erfassungszeitpunkten aufweist. Für Proband:innen, die ihre Hypothese über beide Erfassungszeitpunkte hinweg beibehalten, liegt im Präzeitpunkt ein mittlerer positiver Effekt der Zustimmung zur Argumentkategorie Messunsicherheiten auf den Lernerfolg vor. Im Postzeitpunkt liegt weiterhin ein positiver Effekt vor, der jedoch nun einem kleinen Effekt entspricht. Diese Beobachtung deckt sich mit den Überlegungen aus dem vorigen Abschnitt.

Für Proband:innen, die die Prähypothese verwerfen und sich für eine neue Hypothese entscheiden, zeigt sich, dass im Präzeitpunkt die Zustimmung der Argumentkategorie Messunsicherheiten keinen Einfluss auf die Hypothese hat. Nach der Intervention steigt die Wirkung der Zustimmung zur Argumentkategorie und es liegt ein mittlerer positiver Effekt auf den Lernerfolg vor. Diese Gruppe der Proband:innen wechselt größtenteils von der fachlich falschen zur richtigen Hypothese.

Folgerung

Wird durch die Trennung der Stichproben die im Vergleich zum Prätest sehr hohe Zustimmung zur Argumentkategorie bereinigt, zeigt sich, dass die Effekststärke der Zustimmung zur Argumentkategorie Messunsicherheiten konstant über die Erhebungszeitpunkte bleibt. Es zeigt sich somit, dass die Zustimmung zu den Argumentkategorien gut geeignet ist, um den Wechsel der Hypothese zu erklären. Diese Tatsache stellt somit auch eine Limitierung der Studie dar, da durch den hohen Anteil der Schüler:innen mit richtiger Hypothese zu den unterschiedlichen Zeitpunkten die Chance für den kognitiven Konflikt weniger gegeben ist. Nachfolgend werden daher systematisch die Proband:innen aussortiert, welche die richtige Hypothese bereits zu Versuchsbeginn kennen.

Einfluss der Zustimmung zu den Argumentkategorien auf den Lernerfolg für Proband:innen mit falscher Eingangshypothese

Das verwendete Forschungsinstrument von Ludwig (2017) hatte als Entwicklungshintergrund die Untersuchung, inwiefern die Verwendung bestimmter Argumentkategorien sich auf den Wechsel der fachlichen Hypothese auswirkt (Ludwig, 2017, S. 49 f.). Im vorigen Abschnitt konnte gezeigt werden, dass die Wirkweise der Zustimmung zu den Argumentkategorien auf den Lernerfolg eine Abhängigkeit vom Wechseln der Hypothese aufweist.

Vor diesem Hintergrund soll in einer letzten Analyse untersucht werden, wie die Zustimmung zu den Argumentkategorien auf den Lernerfolg wirkt, wenn die Proband:innen, welche bereits die richtige Hypothese zum vorherigen Erfassungszeitpunkt aufweisen, nicht mehr in der Analyse betrachtet werden. Diese Analyse ermöglicht zu untersuchen, wie sich ein Wechsel der Hypothese durch die Zustimmung zu einer bestimmten Argumentkategorie erklären lässt. Nachfolgend wird nur das mehrfaktorielle Modell in diesem Zusammenhang berichtet. In der Gesamtheit werden zwei Analysen durchgeführt. Zunächst wird untersucht, wie sich die Hereingabe der Daten auf den Wechsel der Hypothese auswirkt. Im nächsten Schritt wird die Analyse erneut durchgeführt, wobei nun auch alle Proband:innen entfernt werden, welche vor der Intervention bereits die korrekte Hypothese aufweisen. Auf diese Weise können die Wirkung der experimentellen Daten und der Intervention auf die Hypothese bzw. den Wechsel der Hypothese gegenübergestellt werden. In allen Modellen wurde die Datenkompetenz als Prädiktor für die Zustimmung zu den Argumentkategorien berücksichtigt.

Die Ergebnisse der Analyse sind in den Tabellen 6.17 und 6.18 dargestellt. Die Resultate der Analysen weichen in diesem Zusammenhang von den Ergebnisse aus Abschnitt 6.2 ab.

Tabelle 6.17 zeigt die Resultate der Analysen, wenn alle Proband:innen aus der Stichprobe entfernt werden, die eingangs bereits die richtige Hypothese gewählt haben. Es zeigt sich, dass in den einfaktoriellen Modellen die Zustimmung zu den Argumentkategorien Intuition ($\beta_{\text{pre}} = 0{,}07$, $\beta_{\text{post}} = -0{,}21$, $p > 0{,}05$) und Expertenwissen ($\beta_{\text{pre}} = 0{,}07$, $\beta_{\text{post}} = 0{,}07$, $p > 0{,}05$) im Prä- und Posttest

Tab. 6.17: Gruppierung nach Eingangshypothese ($n = 237$ mit falscher Hypothese).

	Präzeitpunkt				Postzeitpunkt			
Modell	β	SE	OR [95 % CI]	R^2	β	SE	OR [95 % CI]	R^2
einfaktoriell								
Int	0,07	0,17	1,07 [0,77; 1,49]	0,00	−0,21	0,13	0,81 [0,62; 1,05]	0,02
Exp	0,07	0,22	1,08 [0,70; 1,65]	0,00	0,07	0,16	1,07 [0,78; 1,47]	0,00
MU	0,90*	0,24	2,45* [1,52; 3,95]	0,12*	0,68*	0,19	1,98* [1,36; 2,87]	0,08
Evi	1,23*	0,30	3,40* [1,88; 6,15]	0,16*	0,54*	0,21	1,72* [1,14; 2,59]	0,04
mehrfaktoriell								
Int	0,40	0,22	1,49 [0,96; 2,31]	0,28*	−0,01	0,16	0,99 [0,73; 1,35]	0,08*
Exp	−0,27	0,26	0,77 [0,46; 1,28]		−0,04	0,17	0,96 [0,68; 1,35]	
MU	0,82*	0,27	2,28* [1,34; 3,88]		0,56*	0,21	1,75* [1,16; 2,65]	
Evi	1,28*	0,35	3,60* [1,82; 7,11]		0,27	0,26	1,31 [0,79; 2,17]	

Tab. 6.18: Gruppierung nach Prähypothese ($n = 168$ mit falscher Hypothese).

Modell	Postzeitpunkt			
	β	SE	OR [95 % CI]	R^2
einfaktoriell				
Int	$-0{,}54^*$	0,19	$0{,}58^*$ [0,40; 0,85]	0,09
Exp	$-0{,}24$	0,24	0,79 [0,49; 1,26]	0,01
MU	$1{,}24^*$	0,33	$3{,}44^*$ [1,79; 6,60]	$0{,}23^*$
Evi	$0{,}78^*$	0,33	$2{,}18^*$ [1,13; 4,19]	0,09
mehrfaktoriell				
Int	$-0{,}04$	0,24	0,96 [0,60; 1,54]	$0{,}24^*$
Exp	$-0{,}42$	0,29	0,66 [0,37; 1,16]	
MU	$1{,}07^*$	0,41	$2{,}92^*$ [1,32; 6,47]	
Evi	0,27	0,48	1,31 [0,51; 3,36]	

keinen signifikanten Einfluss auf die Richtigkeit der Hypothese hat.

Die Zustimmung zur Argumentkategorie Messunsicherheiten weist einen signifikanten mittleren positiven Effekt auf die Richtigkeit der Hypothese auf ($\beta_p re = 0{,}90$, $p < 0{,}05$, OR $= 2{,}45$). Das Modell weist eine geringe Varianzaufklärung auf ($R^2 = 0{,}12$). Ein signifikanter Zusammenhang lässt sich im Posttest ebenso beobachten, wobei die Effektstärke zurückgeht (OR $= 0{,}68$, $p < 0{,}05$) und die aufgeklärte Varianz nicht signifkant von null verschieden ist.

Der stärkste Effekt lässt sich bei der Wirkung der Zustimmung zur Argumentkategorie Daten als Evidenz auf die Richtigkeit der Hypothese beobachten ($\beta_{\text{pre}} = 1{,}23$, $p < 0{,}05$). Es handelt sich um einen mittleren positiven Effekt (OR $= 3{,}40$, $p < 0{,}05$). Besonders auffällig ist die aufgeklärte Varianz von $R^2 = 0{,}16$, womit das Modell eine mittlere Varianzaufklärung aufweist. Im Posttest geht die Effektstärke auf einen kleinen positiven Effekt zurück (OR $= 1{,}72$, $p < 0{,}05$). Das Modell weist im Posttest keine von null verschiedene Varianzaufklärung auf.

Das mehrfaktorielle Modell zeigt nur geringe Unterschiede. Signifikante Zusammenhänge lassen sich nur bezüglich der Zustimmung zu den Argumentkategorien Daten als Evidenz und Messunsicherheiten beobachten. Das Modell klärt zum Präzeitpunkt insgesamt 28 % der Varianz auf, was der Summe der einfaktoriellen Modelle entspricht. Es kann also davon ausgegangen werden, dass zum Präzeitpunkt kaum gemeinsame Varianz zwischen der Zustimmung zu den Argumentkategorien bezüglich der Richtigkeit der Hypothese besteht. Im Posttest ist lediglich die Zustimmung zur Argumentkategorie Messunsicherheit ein signifikanter Prädiktor ($\beta_{\text{post}} = 0{,}56$, $p < 0{,}05$). Die Zustimmung zur Argumentkategorie Messunsicherheiten weist einen kleinen positiven Effekt auf die Richtigkeit der Hypothese auf (OR $= 1{,}75$, $p < 0{,}05$). Das Modell klärt einen geringen Anteil der Varianz auf ($R^2 = 0{,}08$).

Tabelle 6.18 zeigt die Modelle, welche den Zusammenhang zwischen der Zustimmung zu den Argumentkategorien und Richtigkeit der Hypothese im Posttest untersuchen. Für die Analysen wurden diejenigen Proband:innen entfernt, welche zum Präzeitpunkt die richtige Hypothese bestimmt haben. Auf diese Weise soll erneut der Fokus auf den Wechsel der Hypothese gelegt werden.

Es zeigt sich, dass im Rahmen der einfaktoriellen Modelle die Zustimmung zu den Argumentkategorien Intuition, Messunsicherheiten und Daten als Evidenz in einem signifikanten Zusammenhang mit der Richtigkeit der Hypothese stehen. Die Zustimmung zur Argumentkategorie Intuition weist dabei einen kleinen negativen Effekt auf die Richtigkeit der Hypothese auf ($\beta_{\text{post}} = -0{,}54, p < 0{,}05$, OR $= 0{,}58$). Die aufgeklärte Varianz von 9 % lässt sich nicht von einem Nulleffekt unterscheiden. Die Varianzaufklärung des Modells kann daher nicht evaluiert werden.

Die Zustimmung zu den Argumentkategorien Messunsicherheiten und Daten als Evidenz weist jeweils positive Effekte auf die Richtigkeit der Hypothese auf.

Die Zustimmung zur Argumentkategorie Messunsicherheiten wirkt mit einem moderaten positiven Effekt ($\beta_{\text{post}} = 1{,}24$, $p < 0{,}5$, OR $= 3{,}44$) auf die Richtigkeit der Hypothese, wohingegen die Zustimmung zur Argumentkategorie Daten als Evidenz einen kleinen positiven Effekt aufweist ($\beta_{\text{post}} = 0{,}78$, $p < 0{,}05$, OR $= 2{,}18$). Hinsichtlich der Varianzaufklärung unterscheiden sich die Modelle ebenfalls. Das Modell zur Zustimmung zur Argumentkategorie Messunsicherheiten klärt mit $R^2 = 0{,}23$ eine mittlere bis hohe Varianz auf. Die Zustimmung zur Argumentkategorie Daten als Evidenz kann im einfaktoriellen Modell keinen signifikanten Anteil der Hypothesenvarianz aufklären.

Das Ergebnis des mehrfaktoriellen Modells zeigt, dass lediglich die Zustimmung zur Argumentkategorie Messunsicherheiten einen signifikanten Zusammenhang zur Richtigkeit der Hypothese aufweist ($\beta_{\text{post}} = 1{,}07; p < 0{,}05$). Dabei liegt ein moderater positiver Effekt vor (OR $= 2{,}92$). Bezüglich der Zustimmung zur Argumentkategorie Daten als Evidenz deutet sich ein kleiner negativer Effekt an (OR $= 0{,}66$, $p > 0{,}05$), der im Rahmen dieser Studie jedoch nicht von einem Nulleffekt unterschieden werden kann. Dieser Effekt wird im weiteren Verlauf daher nicht diskutiert. Das Modell weist insgesamt mit $R^2 = 0{,}24$ eine mittlere bis hohe Varianzaufklärung auf.

Folgerungen

Die vorangehende Analyse sollte die Ergebnisse aus Abschnitt 6.2.5 verallgemeinern und explizit in Bezug zum Wechsel der Hypothese setzen. Es zeigt sich, dass es zwischen Prä- und Postzeitpunkt Unterschiede gibt.

Betrachtet man die Proband:innen, welche eine falsche Eingangshypothese aufweisen, zeigt sich, dass die Zustimmung zu den evidenzbasierten Argumentkategorien Messunsicherheiten und Daten als Evidenz einen moderaten positiven Effekt auf die Richtigkeit der Hypothese aufweist. Während die Ergebnisse für die Argumentkategorie Messunsicherheiten mit vorigen Aussagen übereinstimmen (vKgl. Abschnitt 6.3), zeigt sich ein anderer Zusammenhang für die Wirkung der Zustimmung zur Argumentkategorie Daten als Evidenz. Die Wirkung dieser Argumentkategorie bestätigt nun die Ergebnisse von Ludwig (2017). Das Gesamtmodell klärt 28 % der Varianz auf, wobei es im Vergleich zu den einfaktoriellen Resultaten nur einen geringen Anteil an gemeinsamer Varianz zu geben scheint. Die Summe der aufgeklärten Varianz in den vier einfaktoriellen Modellen entspricht der aufgeklärten Varianz des mehrfaktoriellen Modells. Insgesamt weist das Modell eine hohe Varianzaufklärung auf.

Im Posttest ist derselbe Effekt wie im Posttest in Abschnitt 6.3 zu erkennen, die Wirkung der Zustimmung zu den Argumentkategorien auf die Hypothese verändert sich stark und kann nun nur noch in geringem Maße erklärt werden. Das Modell erklärt nur einen geringen Anteil der Varianz ($R^2 = 0{,}08$). Die Zustimmung zur Argumentkategorie Messunsicherheiten weist einen kleinen positiven Effekt auf die Richtigkeit der Hypothese auf ($OR = 1{,}75$). Die übrigen Faktoren zeigen in Rahmen dieser Studie keinen Effekt auf die Richtigkeit

der Hypothese. Die aufgeklärte Varianz der einfaktoriellen Modelle ist nicht signifikant von null zu unterscheiden, daher ist es nur schwer möglich, über eine geteilte Varianz im mehrfaktoriellen Modell zu sprechen.

Betrachtet und vergleicht man die Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien vor und nach der Intervention, ist hier eine Veränderung zu beobachten. Für die betrachtete Stichprobe erklärt die Datenkompetenz vor der Intervention die Zustimmung zur Argumentkategorie Daten als Evidenz stärker als die Zustimmung zur Argumentkategorie Messunsicherheiten ($\beta = 0{,}49^*, R^2 = 0{,}23^*$ vs. $\beta = 0{,}32^*, R^2 = 0{,}10$). Nach der Intervention verändert sich der Zusammenhang und die Datenkompetenz erklärt nun die Zustimmung zur Argumentkategorie Messunsicherheiten besser als die Zustimmung zur Argumentkategorie Daten als Evidenz ($\beta = 0{,}69^*, R^2 = 0{,}48^*$ vs. $\beta = 0{,}60, R^2 = 0{,}36^*$).

Einerseits ist dieser Effekt nicht verwunderlich, da die durchgeführte Intervention sehr stark die Konzepte von Unsicherheiten und Messergebnissen adressiert. Andererseits erfasst der Test vor und nach der Intervention exakt die gleichen Kompetenzen. Schüler:innen könnten somit die Argumentkategorie Daten als Evidenz bevorzugen, wenn sie keine Datenkompetenz und Konzepte zum Umgang mit Daten und Messunsicherheiten aufweisen. Gestützt wird diese Überlegung von der Tatsache, dass die genannten Argumentkategorien vor und nach der Intervention nicht miteinander korrelieren. Mit der Intervention werden die entsprechenden Konzepte und die Datenkompetenz adressiert, sodass die Schüler:innen nun im vorgegebenen Kontext eher die Argumentkategorie Messunsicherheiten bevorzugen.

6.2.6 Diskussion der Post-hoc-Analyse

Die Ergebnisse der Post-hoc-Analyse haben die offenen Fragen des Abschnitts 6.3 geklärt. Einerseits konnte gezeigt werden, dass die Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien unabhängig von der Hypothesenwahl ist. Sowohl Schüler:innen mit falscher als auch richtiger Hypothese profitieren von den Kompetenzen in gleicher Weise und weisen eine physikdidaktisch adäquatere Zustimmung zu den Argumentkategorien auf.

Die abnehmende Wirkung der Argumentkategorie Messunsicherheiten auf die Richtigkeit der Hypothese konnte ebenso erklärt werden. Durch die Intervention werden die metrologischen Konzepte und die Datenkompetenz zum Vergleich von Messergebnissen derart stark gefördert, dass nach der Intervention ein geringerer Unterschied in der Zustimmung zur Argumentkategorie Messunsicherheiten besteht als zuvor. Dadurch nimmt die Effektstärke der Zustimmung zur Argumentkategorie Messunsicherheiten nach der Intervention ab. Wird die Effektstärke nur auf Grundlage der wechselnden Proband:innen bestimmt, zeigt

sich für die Erhebungszeitpunkte ein konstanter mittlerer positiver Effekt der Zustimmung zur Argumentkategorie Messunsicherheiten auf die Richtigkeit der Hypothese.

Zudem konnte gezeigt werden, dass bei Proband:innen, die ihre Hypothese wechseln, vor und nach der Intervention ein Umdenken bezüglich der Argumentkategorien stattfindet. Die Zustimmung zur Argumentkategorie Messunsicherheiten erfährt einen deutlich größeren Zuwachs als die Argmentkategorie Daten als Evidenz. Neben der Sensibiliserung der Schüler:innen für das Konzept für Messunsicherheiten könnte die beschränkte Skala diesen Effekt erklären. Die Zustimmung zur Argumentkategorie Daten als Evidenz ist bereits im Prätest sehr hoch, sodass ein entsprechender Zuwachs nicht mehr aufgelöst werden kann ($\mu_{\text{pre}} = 2{,}86$, $K\mu_{\text{post}} = 3{,}00$).

Abschließend lässt sich sagen, dass die Erklärung der Hypothese durch die Datenkompetenz und die Zustimmung zu den Argumentkategorien von der betrachteten Situation abhängig sind. Betrachtet man die Ausprägung der Hypothese, so lässt sich die Richtigkeit der Hypothese am besten durch die Datenkompetenz erklären. Betrachtet man den Wechsel der Hypothese, so kann die Hypothese durch die Argumentkategorien gut erklärt werden. Dieser Fall deckt sich mit den Erkenntnissen von Ludwig (2017), der der Zustimmung zur Argumentkategorie Daten als Evidenz insbesondere beim Konzeptwechsel eine besondere Rolle zuweist.

Für die fachdidaktische Praktik bedeutet dies, dass die Förderung einer physikdidaktisch adäquaten Zustimmung zu den Argumentkategorien über die Datenkompetenz insbesondere bei der Erarbeitung von für die Schüler:innen unbekannten Kontexten von großer Bedeutung ist. Ist den Schüler:innen der fachliche Zusammenhang unklar, können durch die Intervention entsprechende Impulse gegeben und eine evidenz- und datenbasierte Begründung gefördert werden, die in einen größeren Lernerfolg mündet.

6.3 Diskussion

Die Ergebnisse aus Abschnitt 6.2 zeigen, dass die vorliegende Arbeit dazu beitragen kann, dass Schüler:innen nach der Intervention weniger der intuitiven Argumentkategorie zustimmen und mehr auf die Argumentkatgeorien Messunsicherheiten und Daten als Evidenz zurückgreifen. Anhand der Ergebnisse lassen sich mehrere Schlussfolgerungen ziehen. Es zeigt sich, dass sich der situative Lernzuwachs der Schüler:innen mit den Argumentkategorien verknüpfen lässt und trotz eines Rückgangs der Zustimmung zur Argumentkategorie Intuition einige Schüler:innen den Erkenntnisgewinn weiterhin mit intuitivem Rechtfer-

tigen assoziieren. In Abschnitt 2.2.3 wurde beschrieben, wie die verwendeten Argumentkategorien im Sinne des ELMs entsprechend ihres kognitiven Charakters in rationalere und eher nicht rationalere Kategorien dichotomisiert werden können. Es lässt sich daher zusammenfassen, dass durch die Intervention die Verarbeitung der vorliegenden Informationen, hier der Messdaten, über zentrale Routen gesteigert und über periphäre Routen verringert wird. Im Hinblick auf das naturwissenschaftliche Argumentieren konnte Ludwig zeigen, dass eine Verarbeitung der Informationen über die zentrale Route einer eher rationalen Verarbeitung entspricht. Entsprechend kann eine Verarbeitung über die periphäre Route als nicht rationale Verarbeitung angesehen werden. Auf Grundlage der vorliegenden Beobachtung und Ergebnisse hinsichtlich der Zustimmung zu den Argumentkategorien vor und nach der Intervention kann somit konstatiert werden, dass der Einsatz der Apps zu einer Förderung einer eher rationalen Verarbeitung von Messdaten beim Argumentieren verwendet werden kann. Vor diesem Hintergrund kann davon ausgegangen werden, dass es sich bei dem Lernerfolg bzw. dem Aufstellen einer richtigen Hypothese um eine permanente Einstellungsänderung und somit um einen zeitlich stabilen Lernerfolg handelt.

6.3.1 Einfluss der Datenkompetenz auf die Zustimmung zu den Argumentkategorien

In Abschnitt 5.2 wurde die empirische Wirksamkeit der Apps zur Förderung von Kompetenzen zum Umgang mit Daten und Unsicherheiten dargelegt. Für die Förderung der Datenkompetenz beim Argumentieren in Studie 3 kann davon ausgegangen werden, dass die Lernenden nach der Intervention im Mittel besser waren als 70 % der Proband:innen vor der Intervention (J. Cohen, 1988).

Es konnte gezeigt werden, dass Kompetenz im Umgang mit Daten und Unsicherheiten negativ mit der Zustimmung zur Argumentkategorie Intuition und positiv mit der Zustimmung zu den Argumentkategorien Messunsicherheiten und Daten als Evidenz zusammenhängt. Diese Effekte werden durch die Förderung der Datenkompetenz verstärkt. Letzteres lässt sich aus den Ergebnissen aus Tabelle 6.8 ableiten: Die dargelegten Effekte lassen sich jeweils lassen sich zum Prä- und Postzeitpunkt beobachten, wobei die Effekte der Datenkompetenz auf die Zustimmung zu den Argumentkategorien Daten als Evidenz und Messunsicherheiten nach der Intervention stärker ausgeprägt sind. Außerdem konnte gezeigt werden, dass der Zuwachs der Datenkompetenz mit den Veränderungen der Zustimmung zu den Argumentkategorien zwischen den Erhebungszeitpunkten zusammenhängt: Je höher der Zuwachs der Ausprägung der Datenkompetenz ausfällt, desto stärker nimmt die Zustimmung zur Argumentkategorie Intuition ab. Außerdem zeigt sich, dass die Zunahme der Zustimmung zur Argumentkategorie Messunsicherheiten stärker ausfällt, je höher die Datenkompetenz der Schüler:innen vor Bearbeitung der App ist.

Vor dem Hintergrund des ELMs ergibt sich für den naturwissenschaftlichen Unterricht daher die Implikation, dass die Datenkompetenz frühzeitig und regelmäßig gefördert werden muss. Auf diese Weise können weniger rationale Verarbeitungen und Begründungen verringert und eher rationale Verarbeitung und Begründungen von Informationen gefördert werden (Ludwig, 2017). Eine frühzeitige Förderung der Datenkompetenz kann eine eher rationale Verarbeitung von Informationen im Unterricht gewährleisten und kann von einer regelmäßigen Adressierung der Datenkompetenz verstärkt werden. Unter Berücksichtigung verschiedener Arbeiten, die die fehlende Datenkompetenz als hemmenden Faktor bezüglich des Argumentierens identifizieren konnten (bspw. Kanari & Millar, 2004; Ludwig et al., 2019; Pols et al., 2021) liegt daher die Vermutung nahe, dass Kompetenz im Umgang mit Daten und Unsicherheiten ebenfalls die Argumentationsfähigkeit beim Experimentieren fördern kann: Die Ergebnisse der vorliegenden Arbeit in Kombination mit den Ergebnissen von Ludwig stellen empirische Evidenz dafür dar, dass die Art der Begründung eines Arguments (Sampson & Clark, 2008) adressiert werden kann und daten- und evidenzbasierte Begründungen innerhalb von Argumenten der Schüler:innen gefördert werden können.

Weiterführende Untersuchungen, die neben der Zustimmung zu den Argumentkategorien die schriftlichen Begründungen und Argumente der Schüler:innen im Fokus der Analyse haben, könnten weitere Hinweise hinsichtlich dieses Zusammenhangs ermitteln. In diesem Zusammenhang wäre insbesondere eine Untersuchung mit mehreren Messwiederholungen sehr interessant: Es könnte einerseits die zeitliche Stabilität des Lernzuwachses betrachtet werden und andererseits, bei wiederholter Förderung der Datenkompetenz, die langfristige Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien sowie auf das Argumentieren untersucht werden. Insbesondere eine detailliertere Analyse, inwiefern die dargestellten verstärkenden Effekte dauerhaft präsent sind, wäre für die Unterrichtsgestaltung zur Förderung der Argumentationsfähigkeit relevant.

6.3.2 Einfluss der Datenkompetenz und der Art der Begründung auf den Lernerfolg

Abschließend bleibt die Wirkung der Zustimmung zur Argumentkategorie zur Richtigkeit der Hypothese zu evaluieren (Forschungsfrage 3). Es konnte gezeigt werden, dass eine höhere Zustimmung zur Argumentkategorie Messunsicherheiten positiv auf die Richtigkeit der Hypothese wirkt. Eine eher rationale Verarbeitung der Informationen führt dazu, dass die Lernenden den funktionalen Zusammenhang zwischen der Pendelmasse und der Periodendauer auf Grundlage der Daten korrekt bestimmen können.

Die Ergebnisse decken sich mit Befunden von Ludwig (2017), der gezeigt hat, dass die Zustimmung zur Argumentkategorie Daten als Evidenz, die ebenfalls der zentralen Route des ELMs zuzuordnen ist, positiv auf den Lernerfolg wirkt (Ludwig, 2017). Es lässt sich somit konstatieren, dass eine physikdidaktisch adäquatere Art der Begründung, die auf Grundlage einer eher rationalen Verarbeitung von Daten entstanden ist, sich positiv auf den Lernerfolg und die zeitliche Stabilität des Lernerfolgs auswirkt.

Ferner zeigt sich, dass die negativen Effekte der Zustimmung zur Argumentkategorie Messunsicherheiten auf die Richtigkeit einer Hypothese (Ludwig et al., 2019; Pols, 2019) umgedreht werden können. Die vorliegenden Ergebnisse zeigen, dass das intuitive Verständnis von Schüler:innen für Messunsicherheiten adressiert und in eine eher rationale Verarbeitung von unsicherheitsbehafteten Daten überführt werden kann.

Neben der Förderung von eher rationalen Verarbeitungen und Begründungen hinsichtlich des naturwissenschaftlichen Argumentierens konnte eine weitere Wirkung der Datenkompetenz gezeigt werden. Nach der Intervention weist die Datenkompetenz einen mittleren positiven Effekt auf die Richtigkeit einer Hypothese auf (vgl. Forschungsfrage 4). Auf Grundlage der in Abschnitt 2.3 dargelegten Aspekte zeigt sich, dass die Datenkompetenz den selbstständigen und datenbasierten Erkenntnisgewinn fördern kann. Eines der zentralen Ziele von naturwissenschaftlichem Unterricht, dass Schüler:innen evidenzbasierte Schlussfolgerungen ziehen (Schiepe-Tiska et al., 2012), kann somit direkt durch die Förderung der Datenkompetenz erreicht werden. Dieser Befund reiht sich in verschiedene Befunde ein (vgl. Abschnitt 2.3.3), dass die Datenkompetenz einen hoch relevanten Faktor für die Ausbildung einer *Scientific Literacy* bei Lernenden darstellt.

Unter Verwendung des mehrfaktoriellen Modells, welches die Zustimmung zu den Argumentkategorien und die Datenkompetenz beinhaltet, zeigt sich, dass die Zustimmung zur Argumentkategorie Messunsicherheiten nach der Intervention keinen signifikanten Einfluss auf die Richtigkeit einer Hypothese aufweist. Außerdem konnte die Post-hoc-Analyse (s. Abschnitt 6.2.5) aufdecken, dass die Zustimmung zur Argumentkategorie Messunsicherheiten weiterhin einen signifikanten Einfluss auf die Richtigkeit einer Hypothese aufweist, wenn *nur* die Proband:innen betrachtet werden, die ihre Hypothese zwischen dem Prä- und Postzeitpunkt ändern. Diese Proband:innen wechselten größtenteils von der falschen Hypothese zur richtigen Hypothese, sodass in dieser Teilstichprobe eher davon auszugehen ist, dass ein kognitiver Konflikt und somit eine strittige Situation als Argumentationsanlass initiiert werden konnte.

Die Ergebnisse der Post-hoc-Analyse haben die offenen Fragen des Abschnitts 6.3 geklärt. Einerseits konnte gezeigt werden, dass die Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien unabhängig von der Hypothesenwahl ist. Sowohl Schüler:innen mit falscher als auch richtiger

Hypothese profitieren von den Kompetenzen in gleicher Weise und weisen eine physikdidaktisch adäquatere Zustimmung zu den Argumentkategorien auf.

Die abnehmende Wirkung der Argumentkategorie Messunsicherheiten auf die Richtigkeit der Hypothese konnte ebenso erklärt werden. Durch die Intervention werden die metrologischen Konzepte und die Datenkompetenz zum Vergleich von Messergebnissen derart stark gefördert, dass nach der Intervention ein geringerer Unterschied in der Zustimmung zur Argumentkategorie Messunsicherheiten besteht als zuvor. Dadurch nimmt die Effektstärke der Zustimmung zur Argumentkategorie Messunsicherheiten nach der Intervention ab. Wird die Effektstärke nur auf Grundlage der wechselnden Proband:innen bestimmt, zeigt sich für die Erhebungszeitpunkte ein konstanter mittlerer positiver Effekt der Zustimmung zur Argumentkategorie Messunsicherheiten auf die Richtigkeit der Hypothese.

Zudem konnte gezeigt werden, dass bei Proband:innen, die ihre Hypothese wechseln, vor und nach der Intervention ein Umdenken bezüglich der Argumentkategorien stattfindet. Die Zustimmung zur Argumentkategorie Messunsicherheiten erfährt einen deutlich größeren Zuwachs als die Argmentkategorie Daten als Evidenz. Neben der Sensibilisierung der Schüler:innen für das Konzept für Messunsicherheiten könnte die beschränkte Skala diesen Effekt erklären. Die Zustimmung zur Argumentkategorie Daten als Evidenz ist bereits im Prätest sehr hoch, sodass ein entsprechender Zuwachs nicht mehr aufgelöst werden kann ($\mu_{pre} = 2{,}86$, $\mu_{post} = 3{,}00$, s. Abbildung 6.3).

Abschließend lässt sich sagen, dass die Erklärung der Hypothese durch die Datenkompetenz und die Zustimmung zu den Argumentkategorien von der betrachteten Situation abhängig ist. Betrachtet man die Ausprägung der Hypothese, so lässt sich die Richtigkeit der Hypothese am besten durch die Datenkompetenz erklären. Wird der Wechsel der Hypothese betrachtet, so kann die Hypothese durch die Argumentkategorien gut erklärt werden. Dieser Fall deckt sich mit den Erkenntnissen von Ludwig (2017), der der Zustimmung zur Argumentkategorie Daten als Evidenz insbesondere beim Konzeptwechsel eine besondere Rolle zuweist.

Für die fachdidaktische Praktik bedeutet dies, dass die Förderung einer physikdidaktisch adäquaten Zustimmung zu den Argumentkategorien über die Datenkompetenz insbesondere bei der Erarbeitung von für Schüler:innen unbekannten Kontexten von großer Bedeutung ist. Ist den Schüler:innen der fachliche Zusammenhang unklar, können durch eine Intervention mit den vorliegenden Apps entsprechende Impulse gegeben und eine evidenz- und datenbasierte Begründung gefördert werden, die in einen größeren Lernerfolg mündet.

Es lassen sich daher drei zentrale Erkenntnisse aus den Ergebnissen ableiten: a) Die Verarbeitung der Informationen und der Lernzuwachs geschehen vor und nach der Intervention über zentrale (Argumentkategorien Daten als Evidenz und Messunsicherheiten) und periphäre (Argumentkategorien Intuition

und Expertenwissen) Routen des ELM. Eine höhere Ausprägung der Datenkompetenz führt dazu, dass Lernende eher die Kategorien Daten als Evidenz und Messunsicherheiten präferieren. Eine Förderung der Datenkompetenz sorgt für eine Abnahme der eher nicht rationalen Verarbeitung von Informationen und des Lernzuwachses. b) Es zeigt sich, dass situationale (Zustimmung zu den Argumentkategorien) und eher personale Faktoren (Ausprägung der Datenkompetenz) einen Einfluss auf die Richtigkeit einer Hypothese haben. Beide Faktoren weisen auf eine zentrale Route der Informationsverarbeitung hin und führen zu einem höheren Lernerfolg. Die Arbeit von Ludwig zeigt zudem, dass ein Lernerfolg, der über die zentrale Route des ELM erreicht wurde, eine höhere zeitliche Stabilität erreichen kann. Es kann festgestellt werden, dass die Förderung der Datenkompetenz eine *Scientific Literacy* der Schüler:innen, eine eher rationale Informationsverarbeitung sowie einen eher daten- und evidenzbasierten Lernerfolg positiv beeinflussen kann. c) Für Proband:innen, bei denen eher ein kognitiver Konflikt initiiert werden konnte, wirkt die Zustimmung zur Argumentkategorie Messunsicherheiten positiv auf die Richtigkeit einer Hypothese. Es zeigt sich daher im Fall eines kognitiven Konflikts die positive Wirkung der Zustimmung zur Argumentkategorie Messunsicherheiten auf die intrapersonelle Argumentation der Lernenden. Vor dem Hintergrund der Arbeiten von Kanari und Millar (2004) sowie Ludwig et al. (2019), die die fehlende Datenkompetenz insbesondere im Umgang mit experimentellen unsicherheitsbehafteten Daten als Hemmfaktor für die Argumentation identifizieren konnten, und beispielsweise Kok (2022), der empirische Evidenz für die positive Wirkung der Datenkompetenz auf die inhaltliche Richtigkeit von Argumenten darlegt, fügen sich die vorliegenden Ergebnisse, die eine positive Wirkung der Datenkompetenz auf die Art der Begründung von Argumenten zeigen, in das Bild ein, dass durch die Förderung der Datenkompetenz die Argumentationsfähigkeit von Schüler:innen gefördert werden kann.

6.3.3 Methodische Limitationen

Im Rahmen der Untersuchung, wie sich Datenkompetenz auf das Argumentieren beim Experimentieren auswirkt (vgl. Abschnitt 6), wurden die Instrumente zur Erfassung der Art der Begründung der Schüler:innen und zur Erfassung der Richtigkeit der Hypothese von Ludwig (2017) verwendet.

Die Darstellung der Zustimmung zur Argumentkategorie Daten als Evidenz zeigt sich zum Prä- und Posttest als linksschief. Es könnte daher sein, dass ein Einfluss dieser Kategorie in der vorliegenden Arbeit nicht gezeigt werden konnte, da ein Deckeneffekt vorliegt (Döring & Bortz, 2016, S. 738). Für die Zustimmung zu den Argumentkategorien Intuition und Messunsicherheiten zeigt sich im Posttest ein Boden- bzw. Deckeneffekt. Letzterer Effekt konnte in der

Post-hoc-Analyse (s. Abschnitt 6.2.5) aufgedeckt werden. Es zeigt sich daher, dass die von Ludwig (2017, S. 175 f.) intendierten mittleren Schwierigkeitsgrade der Skalen im Posttest nicht mehr vorhanden waren.

Eine weitere methodische Limitation stellt die binäre Operationalisierung der Richtigkeit der Hypothese dar. Einerseits ist es dadurch nicht möglich, teilweise Lernfortschritte der Schüler:innen aufzudecken oder aber unter den Schüler:innen mit richtiger Hypothese weiter auszudifferenzieren. Insbesondere ist es somit nicht möglich gewesen, die Wirkung der Datenkompetenz auf den Lernerfolg von Schüler:innen zu untersuchen, welche bereits zum Präzeitpunkt die richtige Hypothese haben und deren Ausprägung der Datenkompetenz durch die Intervention erhöht werden konnte. Andererseits macht die binäre Operationalisierung eine Analyse anhand logistischer Regression notwendig. Für die Evaluation der Modelle wird ein Pseudo-R^2 bestimmt, welches die Verbesserung der Likelihood gegenüber eines Nullmodells beschreibt und daher nicht den Anteil der Varianzaufklärung erklärt, wie beispielsweise das Bestimmtheitsmaß R^2 (Hemmert et al., 2018). Wie in Abschnitt 6.2 beschrieben, eignet sich das Vorgehen trotz der genannten Kritik dennoch, da das in der vorliegenden Arbeit verwendete Pseudo-R^2 von McKelvey und Zavoina (1975) das Bestimmtheitsmaß R^2 gut imitieren kann (Veall & Zimmermann, 1994).

Kapitel 7

Bewertung des Forschungsvorhabens in der Gesamtschau

Ziel der vorliegenden Arbeit ist die Entwicklung von digitalen Lernapps, bei denen das naturwissenschaftliche Experiment als Ausgangspunkt für eine diskursiv-kritische Auseinandersetzung mit naturwissenschaftlichen Informationen dient (vgl. Kapitel 2.5). Vor diesem Hintergrund wurde zunächst die Ausprägung der Kompetenzen im Umgang mit Messdaten gefördert und untersucht, wie sich durch die Lernapps zum Umgang mit Daten und Unsicherheiten das Argumentieren von Schüler:innen im Kontext des naturwissenschaftlichen Experimentierens, hier die Auseinandersetzung mit dem Zusammenhang zwischen der Periodendauer und der Pendelmasse eines Fadenpendels, auswirkt. Im folgenden Abschnitt werden dazu Erkenntnisse aus den Studien 2 und 3 (s. Abschnitt 5.2 und 6.2) mit weiteren Forschungsarbeiten in Zusammenhang gebracht und diskutiert. Gleichzeitig werden Limitationen der Erkenntnisse und Arbeit aufgezeigt sowie einige Kritikpunkte, welche im Rahmen wissenschaftlichen Austauschs im Fokus der Diskussion standen, ausführlich dargelegt und erörtert.

7.1 Übergreifende Diskussion aller drei Studien

Im Rahmen der vorliegenden Arbeit wurden Lernapps zum Umgang mit Daten und Unsicherheiten entwickelt, evaluiert und beim Experimentieren eingesetzt. Die Ergebnisse der Abschnitte 5.2 und 6.2 zeigen die empirische Wirksamkeit der Apps zur Förderung von Datenkompetenz. Der Einsatz der kurzen modularen Interventionen in Form von digitalen Lernapps zeigt sich daher als vielversprechend und deckt sich mit den Ergebnissen anderer vergleichbarer Arbeiten (Kok, 2022). Die Arbeit von Kok (2022) hat den Schwerpunkt auf der Fragestellung, welche Konzepte des Sachstrukturmodells (Hellwig, 2012) Schüler:innen benötigen, um experimentell ermittelte Datensätze miteinander vergleichen zu können. Kok

(2022) konnte zeigen, dass durch Verwendung einer digitalen Lernumgebung zu mehreren Konzepten aus dem Sachstrukturmodell zu Messunsicherheiten (Hellwig, 2012) die Ausprägung der Datenkompetenz zu entsprechenden Konzepten (Schulz, 2022) gefördert werden kann. Die vorliegende Arbeit erweitert diese Erkenntnis dahingehend, dass die Konzepte des Sachstrukturmodells ebenfalls mit kurzen und modularen digitalen Lernapps einzeln adressiert werden können und die dazugehörigen Kompetenzen (Abschnitt 4.1.1 und Schulz, 2022) mit mittleren und großen Effekten gefördert werden können.

In der vorliegenden Arbeit konnte gezeigt werden, dass es einen positiven Zusammenhang zwischen der Förderung der Datenkompetenz zum Umgang mit Daten und Messunsicherheiten und dem Lernerfolg gibt (Forschungsfrage 4). Dieser Befund reiht sich in eine Reihe von Arbeiten ein, die in verschiedenen Kontexten eine ähnliche Wirkung eines adäquaten Umgangs mit Daten und Unsicherheiten auf den Lernerfolg zeigen konnten (Allie & Buffler, 1998; Kok, 2022; Kung, 2005). Die vorliegende Arbeit verwendet Beispiele aus dem Alltag, um die Förderung der Datenkompetenz im Umgang mit Daten und Messunsicherheiten zu motivieren, und einen physikalischen Kontext, um den Lernerfolg der Schüler:innen zu evaluieren. Eine Besonderheit stellt dabei die Kürze der Apps dar, die erlaubt, die Apps in Form von 20- bis 25-minütigen Interventionen in den Unterricht zu integrieren. Vor diesem Hintergrund ist es bemerkenswert, dass der Einsatz der Apps zu physikdidaktisch adäquateren Argumenten von Schüler:innen führen kann. Betrachtet man die zuvor genannten Arbeiten sowie die kontextunabhängige Förderung der Datenkompetenz der vorliegenden Arbeit, könnte die dargelegte Wirkung der Datenkompetenz unabhängig vom Kontext bestehen.

Gleichzeitig zeigt sich in Abschnitt 6.3, dass die Förderung der Datenkompetenz über die Lernapps zulässt, die Art und Weise, wie Schüler:innen Informationen verarbeiten, zu beeinflussen. Insbesondere die Förderung der Ausprägung der Datenkompetenz und die damit einhergehende eher rationalere Informationsverarbeitung bei den Lernenden können als Indiz für eine gesteigerte Argumentationsfähigkeit der Schüler:innen betrachtet werden (s. Abschnitte 6.3 und 2.2.3).

Die Ergebnisse dieser Arbeit zeigen daher eine direkte Wirkung der Ausprägung der Datenkompetenz auf das naturwissenschaftliche Argumentieren beim Experimentieren und bestätigen und bestärken damit frühere Forschungsarbeiten, die beispielsweise das korrekte Auswerten von experimentellen Messdaten als eine Voraussetzung des Argumentierens beim Experimentieren (Kanari & Millar, 2004) identifiziert oder die Förderung von Kompetenz für das Heranziehen von Daten unter Berücksichtigung der Messunsicherheit als Methode zum „Erlernen evidenzbasierten Argumentierens“ (Ludwig, 2017, S. 139) postuliert haben.

Vor diesem Hintergrund stellt die gewählte Methode zur Förderung des Argumentierens beim Experimentieren eine weitere Besonderheit der vorlie-

genden Arbeit dar: Es wurden kurze Unterrichtssequenzen entwickelt, die sich verschiedener Formate zur Förderung des Argumentierens bedienen. Einerseits handelt es sich bei den Lernapps um Lernumgebungen und kurze Interventionen, die von den Lehrkräften unmittelbar in den Unterricht integriert werden können (Berland & Reiser, 2009; Clark & Sampson, 2008; Osborne et al., 2004; Zohar & Nemet, 2002). Andererseits wurden die Lernapps in einem digitalen Format entwickelt (Bell & Linn, 2000; Clark & Sampson, 2008; de Vries et al., 2002), um Lernende während des Lernprozesses selbstreguliert unterstützen zu können (s. Abschnitt 4.1.3). Die Wirkung der vorliegenden Arbeit deckt sich mit den exemplarisch genannten Arbeiten und Arbeiten, die ebenfalls auf digitale Interventionen setzen (bspw. Iordanou & Constantinou, 2015): Es konnte gezeigt werden, dass ein Einsatz kurzer Interventionen im Unterricht a) das Argumentieren begünstigen, b) zu mehr evidenzbasierten Aussagen von Schüler:innen führen und c) die Argumentationsqualität steigern kann.

Zusammenfassend zeigt sich, dass aus den hier durchgeführten Studien relativ starke Evidenz für die Förderung der Datenkompetenz durch die Apps und die Wirkweise der Apps auf das Argumentieren beim Experimentieren konstruiert werden kann. Die Lernapps fördern den adäquaten Umgang mit Daten und Messunsicherheiten, adressieren das prozedurale und deklarative Wissen über Evidenz und tragen somit dazu bei, dass Lernende über Wissenschaft lernen, Messdaten und ihre Unsicherheit als Evidenz verwenden, weniger eher intuitiven und mehr eher rationalen Begründungskategorien zustimmen und somit Informationen eher rational verarbeiten und begründen. Zentrale Ziele einer naturwissenschaftlichen Bildung wie *Scientific Literacy* und *Practices* können somit durch kurze Interventionen in den naturwissenschaftlichen Unterricht integriert und bei den Schüler:innen gefördert werden.

7.2 Limitationen

Die vorliegende Arbeit unterliegt, bedingt durch beispielsweise methodische oder inhaltliche Entscheidungen, verschiedenen Limitationen, die nachfolgend in der Gesamtschau aller drei Teilstudien ausführlich dargelegt und diskutiert werden. Ebenso werden die Durchführung der Studie als Online-Studie sowie die externe Validität der Ergebnisse in der Gesamtschau diskutiert.

situative und kontextuelle Spezifizität der Ergebnisse

Die vorliegende Arbeit orientiert sich methodisch stark an der Vorgehensweise von Ludwig (2017). Damit gehen einige inhalts- und situationsspezifische Limitationen einher, welche nachfolgend kurz dargestellt werden (s. Ludwig, 2017, S.

135 f. für eine ausführliche Diskussion). Anschließend wird der Fokus auf die Intervention durch die digitalen Lernapps gelegt.

Durch die Einnahme einer stark situativen Forschungsperspektive (Ludwig, 2017, S. 135 f.) sind die Ergebnisse der vorliegenden Arbeit stark an die untersuchte Situation, die Untersuchung des Zusammenhangs zwischen der Periodendauer und der Pendelmasse eines Fadenpendels, gebunden. Es ist beispielsweise nicht klar, wie Schüler:innen in Experimenten argumentieren, zu dem Schüler:innen in deutlich geringerem Maß fachlich defizitäre Präkonzepte besitzen oder deren Daten mit geringeren Unsicherheiten behaftet sind. Außerdem erwähnt Ludwig (2017), dass vor dem Hintergrund der situativen Spezifizität unterschiedliche epistemische Zielsetzungen der Proband:innen zu Verzerrungen bei der Stärke der Zustimmung zu den Argumentkategorien führen könnten und sich dadurch die Vorhersagekraft des Modells verringern könnte.

Trotz einer Einleitungs- bzw. Motivationsphase zu Beginn der Studien kann die durchgeführte Intervention nicht mit Schulunterricht gleichgesetzt werden. Einerseits entsteht hieraus das Desiderat, die Wirkung der Lernapps auf das Argumentieren bei einer stärkeren Einbindung in den Schulunterricht zu untersuchen. Andererseits muss vorerst bei der Übertragung dieser Ergebnisse auf den Unterricht auf die Besonderheiten der vorliegenden Arbeit zurückgegriffen werden, die nachfolgend kurz erörtert werden.

Im Rahmen der Arbeit wurde ein Experiment zum Fadenpendel als physikalischer Kontext gewählt (bspw. Kanari & Millar, 2004; Ludwig, 2017). Die Arbeit verwendet somit ein Experiment, zu welchem Schüler:innen stark fachlich defizitäre Präkonzepte haben und welches den Schüler:innen zumeist unbekannt ist. Eine Übertragung auf andere physikalische Experimente, in denen ähnliche Voraussetzungen für die Schüler:innen gegeben sind, erscheint nachvollziehbar. Eine Untersuchung hinsichtlich der Wirkung der digitalen Lernapps auf das Argumentieren im Kontext von physikalischen Experimenten, welche nicht den oben genannten Bedingungen unterstehen, steht daher noch aus und könnte in weiterführenden Forschungsarbeiten ebenfalls adressiert werden.

Die intrapersonelle Argumentation der Schüler:innen sollte in der vorliegenden Arbeit durch den zuvor genannten physikalischen Kontext über einen kognitiven Konflikt initiiert werden. Es ist naheliegend, dass die Ergebnisse der Arbeit in vergleichbaren Situationen, beispielsweise in anderen physikalischen Domänen oder bei Experimenten zu anderen physikalischen Domänen, reproduziert werden können. Grundsätzlich können Argumentationen im naturwissenschaftlichen Unterricht jedoch vielfältig initiiert werden (s. Abschnitt 2.1). Es bleibt daher offen, ob die Wirkung der digitalen Lernapps in anderen Situationen, die zur Initiierung von Argumentationen von Schüler:innen verwendet werden, anders bzw. verschieden ausfällt.

Zusammenfassend zeigt sich, dass die Ergebnisse hochsituativ sind und Deutungen und Übertragungen der Ergebnisse *cum grano salis* erfolgen müssen.

Weiterführende Forschungsarbeiten können die situative und kontextuelle Spezifizität adressieren, um perspektivisch Evidenz für die Generalisierbarkeit der Ergebnisse abzuleiten (vgl. Abschnitt 6.3).

Stichprobenakquise

Die Studie wurde bedingt durch die Corona-Pandemie vollständig online durchgeführt. Bei der Stichprobenakquise wurde eine Vielzahl von Schulen angeschrieben, welche als MINT-freundliche Schule zertifiziert sind (vgl. Abschnitte 5.1.7 und 6.1.8). Daraus ergeben sich für die vorliegende Arbeit zwei Charakteristika: Einerseits ist die Interaktion mit den Lernenden vergleichsweise gering ausgefallen. Die Schüler:innen arbeiteten selbstständig unter der Aufsicht der Lehrkraft und konnten bei Fragen stets den Studiendurchführenden kontaktieren. Eine persönliche Interaktion oder ein persönlicher Austausch mit den Lernenden war nicht möglich. Andererseits handelt es sich bei MINT-freundlichen Schulen um Einrichtungen, welche eine naturwissenschaftliche Profilierung aufweisen. Es ist daher davon auszugehen, dass die Klassen, die teilgenommen haben, ein höheres Interesse und ein höheres Leistungsniveau als vergleichbare Schüler:innen von Schulen, welche nicht MINT-freundlich profiliert sind, aufweisen. Hinsichtlich der externen Validität bedeutet dies, dass eine vergleichbare Untersuchung mit Schulen durchgeführt werden müsste, um die Generalisierbarkeit der Ergebnisse mit Sicherheit festzustellen.

7.3 Häufig geäusserte Kritik

Während der Projektlaufzeit wurden zum Beispiel auf Kongressen verschiedene Kritikpunkte geäußert. Nachfolgend sollen die Kritikpunkte kurz dargestellt und ihnen anschließend vor dem Hintergrund der Arbeit begegnet werden.

7.3.1 Programmierte Instruktion

Die Vermittlung der Kompetenz zu Daten und Messunsicherheiten durch digitale Lernapps unter Rückgriff auf direkte Instruktionen (Kirschner et al., 2006) hat dazu geführt, dass die in Abschnitt 4.1 dargestellten und entwickelten Apps dahingehend kritisiert wurden, dass sie eine Art „Programmierter Instruktion“ in einem neuen Gewand seien. Programmierte Instruktionen stellten ein in den 1960er Jahren sehr beliebtes Format dar, um Lernende selbstständig Inhalte erarbeiten zu lassen (Geis, 1966; Hendershot, 1971; Laube, 1966). Die Namensge-

bung der programmierten Instruktionen bezieht sich primär auf die Funktion als das verwendete Medium: So gibt es viele programmierte Instruktionen in Form von Printmedien, in denen Lernende in Abhängigkeit der gewählten Antwort zu unterschiedliche Seitenzahlen geleitet werden (Laube, 1966; Stolurow & Davis, 1963). Das Verfahren steht der Kritik gegenüber, dass den Lernenden bestimmte Verhaltensweisen „einprogrammiert“ werden. Auf die Frage, ob mit programmierten Instruktionen kreative und phantasievolle Verhaltensweisen gefördert werden können, entgegnet Geis (1966) den Gedanken: „If you can specify for me, or with me, what behaviors you regard as creative and imaginative, we can together attempt to write a program which will produce those behaviors (Geis, 1966, S. 112).“ Nichtsdestrotz betont Geis (1966) ebenfalls die Bedeutung der programmierten Instruktionen bei der Konzeptvermittlung und grenzt sie von der Entwicklung von Routinen ab.

Vor diesem Hintergrund soll nun diskutiert werden, inwiefern die entwickelten Lernapps als programmierte Instruktionen zu betrachten sind. Die Schüler:innen werden zu Beginn der App, der Einführung von interaktiven Elementen und bei der Bearbeitung von Aufgaben sehr stark geführt. Die Bedienung der App arbeitet daher mit direkten Instruktionen, die den Lernenden die Funktionsweise der Bedienelemente näherbringt. Formulierungen der Art „Drücke nun auf weiter, um in den nächsten Abschnitt zu gelangen“ finden sich in den ersten Bereichen der entsprechenden Lernapps. In diesem Fall wird den Schüler:innen zwar eine Verhaltensweise hinsichtlich der Bedienung vermittelt, jedoch stellt dies nicht den zentralen Inhalt der Lernapp dar und spiegelt auch nicht das Ziel der Lernapp-Verwendung wider. Dieser Aspekt kann daher trotz seiner funktionalen Nähe zu den programmierten Instruktionen auf Grundlage der unterschiedlichen Zielsetzung von ebendiesen Instruktionen abgegrenzt werden.

Nach der dargestellten Einführung sind die Schüler:innen selbstständig in der Bearbeitung und können die Lernapp und insbesondere die enthaltenen digitalen und interaktiven Elemente explorativ bearbeiten. Die Lernapp weist als zentrale Gestaltungselemente das individuelle Bearbeitungstempo sowie die selbstregulierte und -ständige Bearbeitung aus. In diesem Bereich steht weniger der Fokus auf eine bestimmte Verhaltensweise, sondern auf der individuellen Erarbeitung der durch die Lernapp adressierten Konzepte. Die offene und individuelle Bearbeitung ist daher ein weiteres Merkmal, das die Lernapp grundlegend von programmierten Instruktionen unterscheidet.

Im letzten Teil der Lernapps steht die Bearbeitung von Aufgaben im Fokus, die eine Möglichkeit bieten, das jeweilige adressierte Konzept anzuwenden und zu vertiefen. Bei der Entwicklung der Aufgaben wurde darauf geachtet, übliche Präkonzepte der Lernenden in den Distraktoren zu berücksichtigen und unmittelbar nach Abgabe der Lösung rückzumelden. Durch die Eindeutigkeit der Lösung werden alle Schüler:innen in dieselbe Richtung gelenkt. Es könnte somit vor dem Hintergrund der programmierten Instruktionen von einer Art Verhalten

gesprochen werden, welches von den Schüler:innen zur korrekten Bearbeitung der Aufgaben erwartet wird.

Anhand der vorangehenden Darstellung von Eigenschaften der Lernapps und der Kontrastrierung mit Charakteristiken von programmierten Instruktionen lässt sich die Frage, ob es sich bei den Lernapps um digitale programmierte Instruktionen handelt, sehr deutlich verneinen: Die im Rahmen der Bedienung und Aufgabenbearbeitung aufgezeigten Ähnlichkeiten beziehen sich nicht auf die wesentliche Eigenschaft der Lernapps: die Konzeptvemittlung. Die enge Führung der Bedienung ist notwendig, um eine erfolgreiche Bearbeitung der App zu ermöglichen. Das *richtige* Verhalten bei den Aufgaben wird nicht erzwungen, da der Test beendet werden kann, ohne alle Aufgaben korrekt gelöst zu haben. Die Bearbeitung der Aufgaben durch die Schüler:innen soll eine Auseindersetzung mit dem neuen Konzept und den Präkonzepten ermöglichen.

Es zeigt sich daher, dass die Kritik der Programmierten Instruktion sich rein auf die Oberflächenstruktur der Apps bezieht. Aspekten wie der engen Führung und den Aufgaben mit einer richtigen Lösung stehen Explorationsphasen, Simulationen und individuelle sowie selbstregulierte Bearbeitungswege der Lernapps gegenüber. Die Tiefenstruktur der Apps adressiert den Aufbau komplexer fachlicher Konzepte und orientiert sich am Basismodell Konzeptaufbau nach Oser und Patry (1994) und steht damit im Widerspruch zu der Kritik der programmierten Instruktion. Für den Konzeptaufbau in den Apps wurde bewusst auf direkte Instruktionen und nicht auf entdeckendes Lernen (bspw. Van Joolingen & De Jong, 1997) zurückgegriffen, da direkte Instruktionen sich als sehr wirksam erwiesen haben (Kirschner et al., 2006).

7.3.2 konkurrierende Modelle in der Auswertung

Ein weiterer Kritikpunkt bezieht sich auf die Frage, ob die Strukturgleichungsmodelle `MoArgHyp` und `MoKompArgHyp` (s. Abschnitt 6.2) gleichzeitig betrachtet werden können oder ob eines davon abgelehnt werden müsste. Die Modelle unterscheiden sich darin, dass das Modell `MoKompArgHyp` neben dem Einfluss der Zustimmung zu den Argumentkategorien zusätzlich den Einfluss der Datenkompetenz auf die Richtigkeit einer Hypothese untersucht. In beiden Modellen ist die Wirkung der Datenkompetenz auf die Zustimmung zu den Argumentkategorien modelliert. Hintergrund der Frage ist, dass die beiden Modelle als konkurrierend aufgefasst werden können. Der Einfluss des Erfassungszeitpunkts auf die Zustimmung zu den Argumentkategorien und die Datenkompetenz der Lernenden ist differenziert zu betrachten: Während die beobachtete Wirkung der Datenkompetenz durchaus eine Abhängigkeit zur Situation aufweisen kann, ist die Ausprägung der Datenkompetenz an sich weniger situationsspezifisch.

In Anbetracht der Tatsache, dass sich die Merkmale in der zeitlichen Stabili-

tät unterscheiden, liegt der Fokus der Forschungsfragen 3 und 4 auf unterschiedlichen Bereichen: Die Analyse unter Verwendung des Modells `MoArgHyp` analysiert den Einfluss der Zustimmung zu den Argumentkategorien, welche in diesem Rahmen ein kontextspezifisches und situatives Merkmal darstellt (Ludwig, 2017). Der Schwerpunkt der Analyse liegt auf der Evaluation des situativen Faktors und dessen Wirkung auf die Richtigkeit der fachlichen Hypothese, wohingegen beim Modell `MoKompArgHyp` der zeitlich stabilere und nicht situative Faktor Datenkompetenz hinzugenommen wird, um dessen Wirkung auf die Richtigkeit der Hypothese zu untersuchen. Die Ausführungen in den Abschnitten 6.3 und 6.2.5 zeigen, dass beide Faktoren einen Einfluss auf die Richtigkeit der Hypothese besitzen und in unterschiedlichen Bedingungen verwendet werden können, um die Wahl der Hypothese zu erklären.

In weiterführenden Untersuchungen könnte der Zusammenhang der beiden Faktoren detaillierter untersucht werden. Die bisherigen Ergebnisse zeigen, dass die Datenkompetenz sowohl die Zustimmung zu den Argumentkategorien als auch die Wahl der Hypothese beeinflusst. Eine weitere Untersuchung mit einem hinsichtlich einer durchzuführenden Mediationanalyse optimierten Design und einer größeren Stichprobe könnte in diesem Fall Aufschluss darüber geben, in welchem Zusammenhang die Wirkungen miteinander stehen und inwiefern ein indirekter Effekt der Datenkompetenz durch die Zustimmung zu den Argumentkategorien mediiert werden kann (Eid et al., 2010, S. 952).

Es kann konkludiert werden, dass eine Konkurrenz der Modelle `MoArgHyp` und `MoKompArgHyp` nicht unmittelbar vorliegt. Die vorangehend dargelegten Aspekte verdeutlichen jedoch, dass weiterführende Untersuchungen der Beziehung zwischen den Faktoren Datenkompetenz und Zustimmung zu den Argumentkategorien vielversprechend sein können. Die Hinzunahme einer qualitativen Argumentanalyse kann ebenfalls dabei helfen, das Verständnis in dieser Hinsicht weiter zu verbessern (z. B. Kok, 2022).

Kapitel 8

Fazit und Ausblick

Ausgangspunkt der vorliegenden Arbeit war eine empirische Untersuchung von Ludwig (2017) zu der Frage, wie Schüler:innen mit quantitativen Messdaten für bzw. gegen eine physikalische Hypothese argumentieren. Die Ergebnisse deuteten darauf hin, dass Kompetenzen im Umgang mit Daten und Messunsicherheiten eine Stellschraube für das Argumentieren beim Experimentieren sein können. In den abschließenden Gedanken seiner Arbeit formuliert Ludwig (2017) das Postulat, dass „der Umgang mit experimentellen Beobachtungen, Messdaten und Unsicherheiten“ explizit gefördert werden solle (Ludwig, 2017, S. 139). Diesem Desiderat ist die vorliegende Arbeit nachgegangen, indem Apps entwickelt worden sind, welche die Datenkompetenz mit mittleren und großen Effekten fördern. Es liegen somit zehn hinsichtlich der Wirksamkeit elaborierte, moderne und didaktisch aufbereitete digitale Lernapps vor, welche das Themenfeld Messunsicherheiten in der Sekundarstufe I adressieren. Die digitalen Lernapps wurden längsschnittlich und in schulüblichen Settings des Physikunterrichts eingesetzt und empirisch evaluiert. Weiterhin konnte gezeigt werden, dass der Erwerb von Datenkompetenz durch die Lernapps dazu beitragen kann, das Argumentieren beim Experimentieren zu verändern: Der Einsatz der Apps führte zu weniger intuitiven oder auf Heuristiken basierenden und zu eher physikdidaktisch wünschenswerten Argumenten von Schüler:innen beim Experimentieren.

Die Förderung von Datenkompetenz wirkt sich positiv auf das Argumentieren mit Messdaten als Evidenz aus.

Ferner existieren nun zehn Apps, um Kompetenzen im Umgang mit Messdaten und Messunsicherheiten in den Unterricht zu integrieren, um so wiederum Schüler:innen beim Argumentieren während des Experimentierens zu unterstützen und damit ein zentrales Ziel naturwissenschaftlichen Unterrichts, Schüler:innen dazu zu befähigen, evidenzbasierte Aussagen zu tätigen, zu adressieren (Schiepe-Tiska et al., 2012).

Durch den Fokus auf einzelne Teilkonzepte aus dem Sachstrukturmodell zu Messunsicherheiten (Hellwig, 2012) konnten kurze Interventionen entwickelt werden, welche in ca. K25 Minuten selbstreguliert durch die Schüler:innen bearbeitet werden können. Zudem besitzen die Lehrkräfte dadurch größere

didaktische Freiheiten bei der Gestaltung der Unterrichtsreihe zum Umgang mit Daten und Messunsicherheiten. Denn die Apps können sequenziell bearbeitet und zu größeren Einheiten zusammengesetzt oder nach Bedarf in einzelnen Unterrichtsphasen eingebunden werden. Ein weiterer Vorteil der Modularität besteht darin, dass die Lehrkräfte besser auf die Voraussetzungen der Lerngruppe eingehen können.

Die Ergebnisse der vorliegenden Arbeit unterstreichen damit die Relevanz eines adäquaten Umgangs mit Daten und Unsicherheiten im Physikunterricht. Eine Förderung der Datenkompetenz kann dazu führen, dass Schüler:innen sich beim Experimentieren eher rational mit experimentellen Daten auseinandersetzen, dass Schüler:innen eher physikdidaktisch wünschenswerte Argumente formulieren und dass der Lernerfolg eine höhere zeitliche Stabilität aufweist. Als Teil der Dissemination der Forschungsprojekts wurden die Apps bereits in Zeitschriften, Workshops und Fortbildungen für Lehrkräfte integriert, um so die Verwendung und Verbreitung der Apps zu begünstigen.

Die vorliegenden Ergebnisse sind vielversprechend, bieten jedoch eine Vielzahl an möglichen Anknüpfungspunkten für weitere Untersuchungsschwerpunkte: Einerseits deuten die Ergebnisse aus Kapitel 6 darauf hin, dass Schüler:innen mit einer höheren Ausprägung von Datenkompetenz stärker von der Apps profitieren. Dieser Aspekt könnte durch einen längsschnittlichen Einsatz der digitalen Lernapps untersucht und mit der Wirksamkeit einer einzelnen Lernapp (s. Kapitel 5) verglichen werden. Andererseits ist noch offen, welche zeitliche Stabilität der Lernerfolg der Schüler:innen aufweist. In diesem Zusammenhang konnten verschiedene Arbeiten zeigen, dass eine positive Wirkung der geförderten Datenkompetenz auf den Lernerfolg zeitlich stabil sein kann (Allie & Buffler, 1998; Pollard et al., 2020). Während diese Arbeiten exemplarisch aufzeigen, dass die positive Wirkung von Datenkompetenz auf den Lernerfolg erhalten bleiben kann, bleibt offen, wie stabil die erreichte Ausprägung der Kompetenzen im Umgang mit Daten und Unsicherheiten durch die Lernapps ist. Insbesondere ist zu beachten, dass es sich bei der vorliegenden Stichprobe um eine andere Zielgruppe als in den genannten Arbeiten handelt. Es ist davon auszugehen, dass Lernende der Sekundarstufe I weniger Experimentiererfahrungen haben und nicht über die mathematischen Fähigkeiten von Studierenden verfügen. Ein adäquater Umgang mit Daten und Messunsicherheiten stellt ähnlich wie Energie oder Kräfte ein hochrelevantes Konzept im Physikunterricht dar. Vor diesem Hintergrund ist es ein Desiderat, die Datenkompetenz beim Lernen von Naturwissenschaften in der Schule zeitlich im Sinne einer Learning Progression zu modellieren. Eine derartige Untersuchung könnte Erkenntnis darüber geben, welchen Kompetenzstand Schüler:innen in Abhängigkeit der Klasse haben und wie die Schüler:innen besser gefördert werden können (K. Neumann et al., 2013).

Die vorliegenden Apps erheben nicht den Anspruch, das Themengebiet der Messunsicherheiten vollständig abzudecken. Die vorliegende Arbeit hatte

den Schwerpunkt auf Inhalten des Physikunterrichts der Sekundarstufe I. Die vorliegende didaktische Struktur könnte dementsprechend um weitere Inhalte ergänzt werden, um Kompetenzen im Umgang mit Daten und Unsicherheiten zu integrieren, welche für den naturwissenschaftlichen Unterricht der Sekundarstufe II relevant sind. Bei der Erstellung der didaktischen Strukturierung wurde Wert darauf gelegt, dass die didaktisch reduzierten Konzepte erweiterbar und anschlussfähig sind. Ziel könnte es daher sein, die didaktische Struktur sukszessive zu erweitern und ein Spiralcurriculum für den Physikunterricht aufzubauen (bspw. Kircher et al., 2020; Stern, 2017).

In weiteren Arbeiten könnte die Argumentqualität der Schüler:innen unter Verwendung weiterer Instrumente differenzierter betrachtet werden: Das Instrument von Kok (2022) ermöglicht es, die inhaltliche Richtigkeit von Argumenten zu evaluieren. In Kombination mit dem Instrument von Ludwig et al. (2019), welches die Art der Begründung evaluiert, wäre es somit möglich, schriftliche Begründungen und Argumente von Lernenden aus physikdidaktischer Perspektive noch präziser zu analysieren.

Darüber hinaus gibt es für die Gestaltung der digitalen Lernapps weiteres Potenzial. Die bisherigen Aufgabenformate könnten insbesondere um offene Antwort- und Begründungsfelder ergänzt werden. Unter Verwendung von Methoden Künstlicher Intelligenz, wie z. B. *Machine Learning*, wäre es denkbar, dass im Sinne eines formativen und individualisierten Feedbacks instantan Rückmeldungen an Schüler:innen gegeben werden können, welche weit über das üblicherweise mögliche Maß an Lehrer:in-Schüler:in-Interaktionen hinausgehen. Die Validität solcher Modelle ist an vielen Einzelfällen getestet worden und es ist davon auszugehen, dass das beschriebene Rückmeldesystem realisiert werden kann. Die Wirksamkeit solcher Methoden in Lernsettings muss jedoch geprüft werden (Zhai et al., 2020). Die Apps könnten somit durch Implementation von formativem Feedback hinsichtlich ihrer Wirkung auf den Lernerfolg weiter verbessert werden.

Wie eingangs skizziert, hängt die Datenkompetenz auch mit den modernen Ansprüchen an eine naturwissenschaftliche Bildung von Schüler:innen zusammen, die unter Konzepten wie *Critical Thinking, Data Literacy, Statistical Literacy* oder *Information Literacy* subsumiert werden können. Die Förderung der Datenkompetenz adressiert unmittelbar eine *Data* und *Statistical Literacy* von Lernenden. Der adäquate Umgang mit Daten und Unsicherheiten kann gleichzeitig dazu verwendet werden, um Schüler:innen Wissen über Evidenz zu vermitteln. Aus dem vorliegenden Beispiel der Arbeit kann von den Schüler:innen ebenfalls gelernt werden, wie in ihrer Umwelt aus Messdaten Evidenz generiert werden kann. Auf diese Weise kann auch eine *Information Literacy* bei den Schüler:innen gefördert werden. Eine adäquate Kommunikation von Evidenz sowie daten- und evidenzbasierten Begründungen von Vermutungen und Hypothesen stellt einen Kernaspekt des Argumentierens dar.

Somit zeigt sich, dass die Förderung der Datenkompetenz und Argumentationsfähigkeit der Schüler:innen einen unmittelbaren Einfluss auf die Herausbildung der *21st century skills* (OECD, 2018) aufweisen kann, deren Bedeutung im Hinblick auf die zuvor aufgezeigten Probleme stetig zunimmt (Marsh & Yang, 2017). Vor dem Hintergrund der Ergebnisse kann konstatiert werden, dass die Förderung von Datenkompetenzen nicht nur die Argumentationsfähigkeit der Schüler:innen beim Experimentieren und den Lernerfolg fördern kann, sondern auch Schüler:innen auf den Umgang mit gesellschaftlichen Herausforderungen vorbereiten kann.

Anhang

Anhang A

Didaktische Ausarbeitungen zu den Inhalten und Konzepten der Lernapps

A.1 A Direktes Messen

Die allgemeine Entwicklung und Gestaltung der Apps wird an dieser Stelle nicht dargestellt. Diesbezüglich findet sich eine Darstellung in der Arbeit des Autors (Kardas, 2019). Ebenso finden sich dort die didaktischen Überlegungen, Kompetenzen und Konzepte zu den Apps A1 und D1 (Kardas, 2019, S. 25, 40).

A.1.1 A2 Bestimmung von Unsicherheiten

Konzept

Es gibt unterschiedliche Methoden, um die Unsicherheit einer Messung zu bestimmen bzw. zu erfassen. Die Wahl der Methoden bzw. die Bestimmungsart der Unsicherheit ist abhängig von den während des Experiments verwendeten Messgeräten.

Struktur

Ziel

Diese Lernumgebung hat zum Ziel, dass die Schüler:innen die Abhängigkeit der Unsicherheitsbestimmung vom eigentlichen Messprozess erkennen, indem sie bei unterschiedlichen Experimenten ebendiese bestimmen sollen. Dabei sollen die Lernenden die Unsicherheitsbestimmung durch Typ A bzw. B durchführen, um diese Eigenschaft bzw. Abhängigkeit durch das Messergebnis und die dazugehörige Unsicherheit quantitativ wiederzugeben. Das zentrale Element

dieser Lernumgebung ist das Konzept, dass es unterschiedliche Methoden zur Bestimmung von Messunsicherheiten gibt und dass die Wahl der Methode vom Experiment und den dazugehörigen Experimentierutensilien abhängig ist.

Bezug zum Sachstrukturmodell

Ein wichtiger Bestandteil dieses Konzepts ist die Dimension Ergebnis der Messung. Das Ergebnis einer Messung ist nicht nur eine Zahl, sondern vielmehr eine Zusammenfassung vieler Informationen, die eine Aussage über das Experiment und den Messprozess ermöglicht. Die Schüler:innen sollen in diesem Bereich insbesondere für den Aspekt Messergebnis als Zusammenfassung aller Informationen sensibilisiert werden. Sowohl die Bestimmung der Unsicherheit als auch ihre Quantität ist von der Messung und den dazugehörigen Messgeräten abhängig. In diesem Zusammenhang bestimmen die Schüler:innen zu zwei gegebenen Experimenten bzw. mithilfe zweier Simulationen die jeweiligen Unsicherheiten und stellen diese dar. Dadurch wird der Aspekt Messergebnis als Zusammenfassung aller Informationen durch Schüler:innen in der Lernumgebung wahrgenommen. In der Lernumgebung werden den Lernenden zwei unterschiedliche Experimente zur Bestimmung der gleichen Größe vorgestellt. Die Schüler:innen sollen kennenlernen, dass die Angabe der Unsicherheit im Messergebnis nicht nur ein quantitativer Wert ist, sondern auch durch das Experiment bzw. die Messung und den Messprozess beeinflusst wird.

In diesem Zusammenhang wird Bezug auf die Teildimension Erfassung einer Unsicherheitskomponente bei direkter Messung genommen. Den Schüler:innen ist eventuell eine statistische Methode zur Bestimmung von Unsicherheiten bekannt, die sich im vorliegenden Kontext jedoch nicht auf beide Experimente anwenden lässt. Auf diese Weise soll vermittelt werden, dass die Methode zur Bestimmung der Unsicherheit auf Grundlage des Experiments ausgewählt werden muss. Den Lernenden wird eine exemplarische Methode zur Bestimmung dieser Unsicherheit beigebracht: Den Schüler:innen wird im Auswertungsteil der Lernumgebung vorgegeben, dass das verwendete Gerät auf $\pm 5\,\mathrm{m}$ genau arbeitet. Es wird verallgemeinert, dass es in Fällen, in denen eine Wiederholung der Messung zu gleichen Messwerten führt, die Unsicherheitsbestimmung über eine Recherche bzw. die Nachschlage-Methode durchgeführt werden kann. Die Eigenschaft des Geräts fließt somit unmittelbar in die daraus resultierende Unsicherheit der Messung und des Messergebnisses ein.

Im Rahmen der beiden Experimente lernen die Schüler:innen unterschiedliche Ursachen für die Messunsicherheiten kennen. In diesem Kontext wird die Dimension Grundsätzliche Existenz von Messunsicherheiten und insbesondere die Teildimension Einflussgrößen adressiert. Einerseits wird die Unvollkommenheit der Messgeräte hervorgehoben als auch die Umwelteinflüsse während der Messung berücksichtigt. Das erste Experiment, welches die Strecke mittels einer

Kartennavigation ermittelt, thematisiert die Unvollkommenheit der Geräte, indem die Länge bzw. der Standort stets nur auf 5 m genau (Vorsicht!) angegeben werden kann. Im zweiten Experiment wird derselbe Aspekt aufgegriffen, indem das dort verwendete Messgerät (Smartphone mit App zum Schrittzählen) konstruktionsbedingt schwanken kann. Der Umwelteinfluss wird nur im zweiten Experiment aufgegriffen, wo ein Teil der variierenden Messwerte auf wechselnde meteorologische Bedingungen zurückgeführt wird.

Didaktische Reduktion

Das Messergebnis, bestehend aus dem Messwert und der Messunsicherheit, soll es ermöglichen, das Experiment sowohl qualitativ als auch quantitativ zu beschreiben. Aus diesem Grund werden beim Aufstellen der Unsicherheit stets alle relevanten Unsicherheitsquellen berücksichtigt. In der Lernumgebung wird in einem der Experimente bzw. in einer der Simulationen ein Spezialfall aufgegriffen, indem zufällige Messunsicherheiten ausgeschlossen werden, damit der Fokus auf der neuen Bestimmungsmethode liegt. An dieser Stelle ist darauf zu achten, dass den Schüler:innen nicht das Bild vermittelt wird, dass es immer nur eine richtige Bestimmungsmethode für die Unsicherheit gibt, sondern es wird Wert daraufgelegt, zu betonen, dass die Bestimmungsmethode der Unsicherheit zum Experiment passen muss. Auf diese Weise soll verhindert werden, dass die Schüler:innen denken, dass es jeweils nur eine richtige Methode zur Unsicherheitsbestimmung gibt. Außerdem kann das Konzept zu einem späteren Zeitpunkt (höhere Klassenstufe) noch erweitert werden, indem erwähnt werden kann, dass sich die Unsicherheiten ergänzen. Dieser Zusammenhang kann exemplarisch mit der pythagoreischen Summe eingeführt und verdeutlicht werden.

Konzeptbildung nach Oser

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Bewusstmachung des Vorwissens	Start bzw. Einleitung	Der Start (in Form eines Concept-Cartons o. Ä.) wird verwendet, um das Vorwissen der Schüler:innen zu aktivieren. Dies geschieht in einem Kontext, der den Schüler:innen aus einem außerschulischen Umfeld bekannt ist. Den Schüler:innen wird ins Gedächtnis gerufen, was die grundlegenden Eigenschaften eines Messergebnisses sind.

. . . Tabelle geht auf der nächsten Seite weiter

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Durcharbeiten eines Prototyps	Vermessung/ Erklärung	Die Schüler:innen führen ein Experiment zur Bestimmung der Schulweglänge mit zwei unterschiedlichen Methoden durch (bzw. ist es eher angemessen, von einer Art Simulation zu sprechen). Während bei der ersten Methode keine Streuung in den Messwerten zu erkennen ist, streuen die Messwerte der zweiten Versuchsanordnung. Es kommt die Frage auf, wie nun das Messergebnis bestimmt werden kann. Dazu wird den Schüler:innen für die ersten Messwerte bzw. den ersten Messwert gezeigt, dass in diesem Fall eine externe Recherche für die Bestimmung der Unsicherheit nötig ist. Das zweite Messergebnis kann mittels statistischer Methoden bestimmt werden.
Beschreibung der wichtigen Merkmale des neuen Konzepts	Erklärung	Die Schüler:innen bekommen ein exemplarisches Beispiel für die Bestimmung der Unsicherheit, falls bei wiederholter Messung stets derselbe Messwert erfasst wird. Es wird verdeutlicht, dass es neben den statistischen Methoden weitere Methoden zur Bestimmung der Unsicherheit gibt.
Aktiver Umgang mit dem neuen Konzept	Aufgaben	Die Schüler:innen müssen selbstständig Aufgaben lösen und das neue Konzept anwenden. Beispielsweise muss entschieden werden, welche Methode zur Bestimmung von Unsicherheiten bei einem vorgegebenen Experiment sinnvoller ist. Entsprechend gibt es Variationen.
Anwenden des neuen Konzepts in anderen Kontexten	Aufgaben	Die Aufgaben werden nun geöffnet. Es werden weitere physikalische Kontexte (bspw. Massebestimmung bei der Bestimmung der Erdbeschleunigung) und Variationen der Aufgaben und Methoden bearbeitet.

Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene

Phase	Konzeptebene	Kontextebene
Allgemein (Ziele)	Messunsicherheiten können nicht nur auf Grundlage von statistischen Methoden ermittelt werden, sonderna uch mit anderen Informationsquellen, bspw. Recherchen, Geräteeigenschaften und anderen Einflussfaktoren wie bspw. die Reaktionszeit, ermittelt werden. Schüler:innen sollen verstehen, dass nicht die physikalische Größe, sondern das Experiment und die verwendeten Geräte die Unsicherheit bestimmen. Dazu gehört insbesondere das Verständnis, dass die Messunsicherheit das Experiment adäquat wiedergibt.	Schüler:innen vermessen einen virtuellen Schulweg und verwenden dabei einmal die Kartennavigation und ein anderes Mal eine App, um Schritte zu zählen. Aus beiden Experimenten sollen Messergebnisse für die Länge des Schulwegs ermittelt werden.
Einleitung	Wiederholung der Inhalte, die sie bereits kennen (Ergebnis besteht aus Messwert bzw. Mittelwert und Unsicherheit, die mit Maximalfehler abgeschätzt wird) Ziel der Lernumgebung: neue Methode zur Bestimmung von Unsicherheiten	Nennung des Kontexts: Bestimmung der Schulweglänge
Vermessung	Schüler:innen führen zwei Versuche durch: Messung 1 erfolgt mit einer Methode, welche bei wiederholter Messung den gleichen Messwert liefert. Messung 2 erfolgt mit einer Methode, welche bei wiederholtem Messen eine Streuung aufweist. Beobachtung bzw. zusammenfassende Bemerkung um dieses Kapitel zu beenden und den nächsten Abschnitt einzuführen.	Schüler:innen vermessen den Schulweg, wobei die Messungen virtuell durchgeführt werden . Zunächst messen sie in einem Kartenausschnitt den vorgegebenen Weg aus. Selbst bei wiederholter Messung erfolgt stet der gleiche Messwert. Danach wird die Schulweglänge mittels einer App bestimmt, die die Schritte zählt. Die Länge des Schulwegs wird an fünf Schultagen gemessen und man erhält fünf unterschiedliche Messwerte für die Länge des Schulwegs.

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptebene	Kontextebene
Erklärung	Messergebnisse werden den Schüler:innen präsentiert	Die beiden ermittelten Messwerte werden gegenübergestellt.
	Unsicherheit zu Methode 1 wird motiviert: Bisheriges Wissen funktioniert nicht → Schüler:innen erhalten Informationen zum Experiment und können anschließend die Unsicherheit benennen.	Methode 1: Auf die Eigenschaft der Kartennavigation zu sprechen kommen. Die Angaben stimmen bis auf 10 m. Da die Werte an sich nicht streuen, kann in diesem Fall die Eigenschaft der Kartennavigation als Unsicherheit verwendet werden.
	Unsicherheit zu Methode 2 wird motiviert: Die einzelnen Messwerte streuen → Die Unsicherheit kann mit dem Wissen aus App A1 (Verwendung von Mittelwert und Maximalfehler) bestimmt werden. Warum streuen sie? Alleine gelaufen, mit Freunden, es hat geregnet, rutschig, Schnee etc.	Methode 2: Die Werte streuen. Zur Bestimmung der Unsicherheit kann der Maximalfehler verwendet werden, der bereits bekannt ist. Dazu müssen Mittelwert und der größte Unterschied vom Mittelwert bestimmt werden.
	Die Messergebnisse der beiden unterschiedlichen Methoden werden gegenübergestellt: Mit beiden Experimenten war es möglich, die Länge des Schulwegs zu ermitteln. Die Messergebnisse erlauben Rückschlüsse auf die Messung.	Beide Experimente und Wege führen somit zu einem Messergebnis. Man erhält mit beiden Wegen einen ähnlichen Wert. Aufgrund der unterschiedlichen Methoden ergeben sich aber verschiedene Unsicherheiten. Die Unsicherheit repräsentiert also auch den Messvorgang und die verwendeten Geräte.
Aufgaben		
Zusammenfassung	Es gibt mehrere Methoden, Messunsicherheiten zu bestimmen. Insbesondere bei Messungen, die keine Streuung besitzen, kann nicht auf statistische Methoden zurückgegriffen werden.	Der Schulweg lässt sich mit unterschiedlichen Experimenten bestimmen. Bei Experimenten, die bei einer Wiederholung der Messung stets denselben Messwert erhalten, sollte die Unsicherheit mit der Nachschlagemethode ermittelt werden. Diese bietet vielfältige Möglichkeiten, um die Unsicherheit im Rahmen des Experiments abzuschätzen.

A.1.2 A3 Bestimmung von Unsicherheiten

Konzept

Die Messunsicherheit ist ein Parameter, der die Eigenschaften der Messung und das Vertrauen in die Messung beschreibt, während die Messabweichung aussagt, wie groß die Differenz zwischen dem Messergebnis und dem Referenzwert (wahren Wert) ist.

Struktur

Ziel

Schüler:innen lernen in dieser Lernumgebung, dass die Messabweichung ein Fachbegriff beim Experimentieren ist. Sie lernen die Definition dieses Begriffs kennen und können zwischen Messunsicherheit und Messabweichung in unterschiedlichen Szenarien voneinander unterscheiden.

Bezug zum Sachstrukturmodell

Einige grundlegende Begriffe der Metrologie sind bei Schüler:innen mit Präkonzepten versehen. Insbesondere die Begriffe Unsicherheit, Abweichung und Fehler werden beim Experimentieren in einigen Fällen synonym gebraucht (Hamacher et al., 2015; Nagel, 2021). Dabei betrachtet jeder Begriff einen anderen Aspekt der Messung und aus diesem Grund muss auf die fachgerechte Verwendung geachtet werden.

Die Messunsicherheit ist ein quantitativer Parameter, der den Messprozess beschreibt und ein Index für die Streuung der Messwerte ist. Auf Grundlage der Messunsicherheit wird anhand des Messwertes ein Intervall konstruiert, in dem sich der Messwert mit einer hohen Wahrscheinlichkeit befindet. Anhand der Messunsicherheit kann im Nachhinein die Präzision der Messung auf qualitativer und quantitativer Basis beurteilt werden. Je kleiner die Messunsicherheit ist, desto präzisier ist die Messung. Dabei sollte nicht versucht werden, die Messunsicherheit beliebig zu verkleinern, sondern anhand des Ziels des Experiments und der verwendeten Geräte eine sinnvolle Messunsicherheit für die Messung zu bestimmen. Die Messabweichung bezeichnet den Unterschied des Messwerts bzw. Messergebnisses vom wahren Wert der Messgröße. Der wahre Wert ist hierbei ein idealisiertes Konzept (Hellwig, 2012, S. 173), welches in der Realität nahezu nie betrachtet werden kann. Aus diesem Grund wird in diesen Fällen ein anderer Wert zum Vergleich herangezogen, wobei es sich um angegebene Referenzwerte, Literaturwerte oder Messwerte bzw. Messergebnisse handeln kann. Die Unterscheidung zwischen Messunsicherheit und Messabweichung findet sich

im Sachstrukturmodell von Hellwig (2012) in der Dimension Grundsätzliche Existenz von Messunsicherheiten im Unterpunkt Unterscheidung zwischen Messunsicherheit und Messabweichung. Die Messabweichung wird in die Bereiche systematische und zufällige Messabweichung unterteilt. Unter systematischen Messabweichungen versteht man Auswirkungen von wiederkehrenden Einflussgrößen, die die Messung stets auf dieselbe Weise beeinflussen. Systematische Abweichungen können nachträglich reduziert bzw. in einigen Fällen sogar behoben werden. Im Allgemeinen beeinflusst die systematische Messabweichung nicht die Messunsicherheit der Messung, da die Einflussgröße stets dieselbe Auswirkung auf die Messung besitzt. Bei zufälligen Messabweichungen handelt es sich um das Resultat von Einflussgrößen, welche keine systematische Wirkung besitzen. Die Auswirkungen dieser Einflussgrößen sind unvorhersehbar bzw. stochastisch verteilt. Zufällige Messabweichungen können bis zu einem gewissen Maß während des Messprozesses reduziert werden, jedoch ist eine nachträgliche Reduzierung nicht möglich. Die zufälligen Messabweichungen haben einen unmittelbaren Einfluss auf die Messunsicherheit einer Messung.

Schließlich können die Messabweichung und -unsicherheit eines Ergebnisses verwendet werden, um die metrologische Verträglichkeit eines Messergebnisses mit der Messgröße zu prüfen. Das folgende Konzept entstammt dem Sachstrukturmodell (Hellwig, 2012) aus der Dimension Aussagekraft von Messunsicherheiten und dem dazugehörigen Unterkonzept Vergleich von Messwerten. Ist die Messunsicherheit des Messergebnisses größer als die Messabweichung, sind die beiden Werte nicht mehr unterscheidbar. Ist die Messabweichung größer als die Messunsicherheit, so sind die Werte im Allgemeinen nicht miteinander verträglich. Anhand der Messabweichung kann die Genauigkeit einer Messung beurteilt werden. Je näher der Messwert bzw. das Messergebnis an der eigentlichen Messgröße liegt, desto genauer ist die Messung. Hierbei ist zu berücksichtigen, dass die Attribute Genauigkeit und Präzision unabhängig voneinander ausgeprägt sein können. Messungen können demnach unpräzise sowie genau oder präzise und ungenau sein. Es ist daher von großer Bedeutung, die Parameter Messunsicherheit und Messabweichung im Rahmen des Experimentierens voneinander zu trennen.

Didaktische Reduktion

Der zentrale Aspekt dieser Lernumgebung ist die Unterscheidung zwischen Messunsicherheit und Messabweichung. Aus diesem Grund wird Wert darauf gelegt, dass die Schüler:innen eine grundlegende Vorstellung des Konzepts Messabweichung erlangen. Die Messabweichung wird als Differenz zwischen Messergebnis und einem Referenz- oder Literaturwert eingeführt. Eine Unterscheidung zwischen systematischer und zufälliger Messabweichung findet in dieser Lernumgebung nicht statt. Zunächst soll stattdessen die grundlegende Idee der Messabweichung als Differenz zu einem Referenz- oder Literaturwert etabliert werden, bevor das Konzept der Abweichung bzw. korrekter Differenz

zur Messgröße betrachtet wird, die bereits während des Messens auftreten kann. Letzteres stellt ein erweitertes und vollständiges Konzept von Messabweichungen dar, welches in einer weiteren Lernumgebung thematisiert werden kann. So soll vermieden werden, dass die Schüler:innen die Konzepte Messunsicherheit und Messabweichung durcheinanderbringen, insbesondere unter dem Aspekt, dass die zufällige Messabweichung die Messunsicherheit beeinflusst. Die Messabweichung und Messunsicherheit des Messergebnisses, welche im Kontext dieser Lernumgebung ermittelt werden, werden grafisch gegenübergestellt, um die Unterschiede zu verdeutlichen. Es wird jedoch darauf verzichtet, die Verträglichkeit des Messergebnisses mit dem Literaturwert explizit zu thematisieren und auf die quantitativen Verhältnisse von Messabweichung und -unsicherheit zurückzuführen. Die Unterscheidbarkeit bzw. die Verträglichkeit von Messwerten bzw. Messergebnissen stellt ein separates Konzept dar, welches in einer selbstständigen Lernumgebung realisiert wird. Auf diese Weise wird die Modularität der Lernumgebungen beibehalten und ein eindeutiger Schwerpunkt für diese Lernumgebung gesetzt.

Konzeptbildung nach Oser

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Bewusstmachung des Vorwissens	Start bzw. Einführung	In diesen Phasen werden unterschiedliche Ziele bezüglich des Kontextes und des Konzepts umgesetzt: Während der Startphase sollen die Schüler:innen mit dem Kontext der Lernumgebung vertraut werden und einen strittigen Punkt kennenlernen. Dadurch sollen die Schüler:innen für einen kritischen Umgang mit Messungen sensibilisiert und für diesen Themenbereich aktiviert werden. Während der Einleitung werden die Voraussetzungen für diese Lernumgebung genannt und die Ziele formuliert. Diese Phase hat somit eine große Bedeutung auf der Ebene des Konzepts.

. . . Tabelle geht auf der nächsten Seite weiter

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Durcharbeiten eines Prototyps	Bestimmung des Messergebnisses/Erklärung	Die Schüler:innen bestimmen zunächst ein Messergebnis auf Grundlage eines gegebenen Datensatzes und werden anschließend dazu aufgefordert die Differenz zu einem Referenzwert zu bestimmen. Auf diese Weise haben die Schüler:innen die Messunsicherheit und -abweichung bestimmt und können.
Beschreibung der wichtigen Merkmale des neuen Konzepts	Erklärung/Eigenschaft	Die Phase Erklärung/Eigenschaft wird verwendet, um die vorangehend bestimmte Messabweichung zu charakterisieren. Der Begriff Messabweichung wird eingeführt und die Schüler:innen lernen die grundlegenden Eigenschaften der Messabweichung kennen. Insbesondere von großer Bedeutung ist hier, dass Konzept Messabweichung derart auszugestalten, dass die Schüler:innen nicht Gefahr laufen, dieses mit der Vorstellung von Messunsicherheiten zu verwechseln.
Aktiver Umgang mit dem neuen Konzept	Aufgaben	Die Schüler:innen sollen im selben Kontext das neue Konzept anwenden. Unter anderem sollen sie zu gegebenen Referenzwerten und Messergebnis die Messabweichung bestimmen, zu gegebenen Szenarien entscheiden, ob es sich um Messunsicherheiten oder Messabweichungen handelt etc. Durch die Variation der Aufgaben soll die Internalisierung des neuen Konzepts gewährleistet werden.

. . . Tabelle geht auf der nächsten Seite weiter

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Anwenden des neuen Konzepts in anderen Kontexten	Aufgaben	Abschließend sollen die Schüler:innen das Konzept auf neue Kontexte anwenden. Durch den Transfer soll sichergestellt werden, dass die Schüler:innen das Konzept nicht nur einem Kontext zuweisen, sondern auf unterschiedliche Kontexte und Situationen anwenden können. Diese Aspekt ist insbesondere vor dem Hintergrund des Experimentierens von Bedeutung, da die Experimente in der Schule oft zu unterschiedlichen Kontexten erfolgen.

Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene

Phase	Konzeptebene	Kontextebene
Allgemein (Ziele)	Schüler:innen lernen die Messabweichung als Differenz des Messergebnisses zu einem Referenzwert kennen und können diesen Begriff von der Messunsicherheit einer Messung trennen.	Fahrt mit einem E-Scooter und die Frage, inwiefern die angegebene Höchstgeschwindigkeit stimmt bzw. mit den gesetzlichen Vorgaben von 20 km/h übereinstimmt.
Einleitung	Wissen der Schüler:innen aktivieren: Vollständiges Messergebnis besteht aus Messwert und Messunsicherheit. Ziel der LU: Die Differenz zwischen Messergebnis und einem Vergleichswert nennt man Messabweichung. Sie ist von der Messunsicherheit zu unterscheiden.	Lee und seine Freunde fahren in ihrer Freizeit mit dem E-Scooter. Nun hat Lee erfahren, dass das Gesetz vorgibt, dass die E-Scooter nur 20 km/h fahren dürfen. Seine Freunde und er diskutieren, ob diese Vorgabe bei dem Modell, welches sie verwenden, erfüllt ist. Dazu beschleunigt jeder seine Freunde auf die Höchstgeschwindigkeit und fährt dann eine abgesteckte Strecke ab. Die Zeit wird von den anderen Freunden gemessen.

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptebene	Kontextebene
Bestimmung des Messergebnisses	Schüler:innen werden kurz angewiesen, das Messergebnis zum vorliegenden Datensatz zu bestimmen. Die Schüler:innen sollen die vorliegenden Daten in die entsprechenden Felder eintragen und das Messergebnis nach und nach erhalten.	Die Freunde haben die folgenden Messwerte in km/h für jeden Durchlauf erhoben: 19,3 19,5 20,0 20,2 19,1 18,3 Sie möchten nun ein Messergebnis bestimmen. Hilf ihnen dabei, die fehlenden Größen zu bestimmen.
Erklärung / Eigenschaften	Schüler:innen werden Messergebnis und Refenzwert grafisch eingeblendet. Die bereits bekannte Messunsicherheit und die neue Messabweichung werden visualisiert → die Unterschiede werden wie in der App D1 grafisch hervorgehoben.	Gemeinsam habt ihr das Messergebnis 19,4(11) km/h ermittelt. Betrachtet nun den Unterschied zur gesetzlichen Vorgabe von 20 km/h und tragt ihn in das entsprechende Feld ein.
Aufgaben		
Zusammenfassung	Beim Messen muss zwischen den Begriffen Messunsicherheit und Messabweichung unterschieden werden. Die Messabweichung gibt an, wie weit das Messergebnis vom Vergleichswert entfernt ist. Die Messunsicherheit beschreibt die Messung und die dazugehörige Streuung der Werte.	

A.2 B Indirektes Messen

A.2.1 B1 Addition von absoluten Unsicherheiten

Konzept

Werden Messwerte bzw. -ergebnisse mit Unsicherheiten addiert oder subtrahiert, ergibt sich die Unsicherheit des Ergebnisses, indem die absoluten Unsicherheiten der Messwerte addiert werden.

Struktur

Ziel

Die Schüler:innen sollen in dieser App erarbeiten, wie sich die Unsicherheit beim indirekten Messen von physikalischen Größen im Fall der Addition und Subtraktion fortpflanzt. Den Schüler:innen wird das Konzept vermittelt, dass sich im FallK der Addition bzw. Subtraktion die absoluten Unsicherheiten der gemessenen Größen addieren.

Theoretischer Hintergrund

Einige physikalische Größen lassen sich nicht direkt oder mit großem Aufwand direkt messen. In diesen Fällen kann die eigentliche Messung über eine indirekte Messung durchgeführt werden. Aus den Messwerten kann die gesuchte physikalische Größe durch eine mathematische Modellierung bestimmt werden. Abhängig von der zugrunde liegenden mathematischen Modellierung muss die Unsicherheit des Ergebnisses bestimmt werden. Die entstehende Unsicherheit wird durch die Unsicherheiten der Eingangsgrößen beeinflusst, daher wird in diesem Fall von einer Unsicherheitsfortpflanzung gesprochen. Für eine beliebige mathematische Modellierung mit den unkorrelierten Eingangsgrößen $x_1, \ldots, x_n$ mit den dazugehörigen Unsicherheiten $\delta_1, \ldots, \delta_n$ und der Modellierung $x = f(x_1, \ldots, x_n)$ kann die Unsicherheit über

$$\delta_x = \sqrt{\Sigma_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \cdot \delta_i \right)^2}$$

abgeschätzt werden (Möhrke & Runge, 2020, S. 89).

Diese Art der Unsicherheitsbestimmung entspricht einem ganzheitlichen Ansatz und wird im Sachstrukturmodell von Hellwig in der Dimension Zusammensetzung der Messunsicherheit aus mehreren Komponenten im Unteraspekt Fortpflanzung der Messunsicherheit adressiert. Sie setzt voraus, dass

die gemessenen Werte und die dazugehörigen Unsicherheiten nicht miteinander korreliert sind.

Der Schwerpunkt dieser Lernumgebung liegt nicht darin, den oben genannten allgemeinen ganzheitlichen Ansatz zur Unsicherheitsfortpflanzung zu vermitteln, sondern eine schrittweise Bestimmung der Unsicherheit für den Spezialfall der Addition bzw. Subtraktion von zwei gemessenen Werten darzustellen. Daraus ergibt sich der Spezialfall $\delta = \sqrt{\delta_1^2 + \delta_2^2}$. Eine allgemeinere Betrachtung des Falls Addition bzw. Subtraktion kann der Teildimension Schrittweise Bestimmung der Gesamtunsicherheit unter dem Punkt Summen/Differenzen gemessener Größen entnommen werden. Neben der Fortpflanzung der Unsicherheiten hat die Lernumgebung die mathematische Modellierung als Inhalt. Die mathematische Modellierung ist theoriebasiert und stellt somit den physikalischen Zusammenhang zwischen der zu ermittelnden Größe und den gemessenen Größen dar. Die Schüler:innen sollen erkennen, welche Grundlage die mathematische Modellierung besitzt und dass sie die Unsicherheitsfortpflanzung beeinflusst.

Didaktische Reduktion

Die Schüler:innen werden in dieser App nicht mit dem Spezialfall der Unsicherheitsfortpflanzung für die Addition konfrontiert. Die phyagoreische Summe ist schwer zu motivieren und setzt eine stochastische Unabhängigkeit voraus, die im Rahmen dieser App nicht behandelt wird. Es wird auf das Supremum für die Abschätzung der Unsicherheit zurückgegriffen. Die Schüler:innen lernen kennen, dass die Unsicherheit der Summe von zwei Messwerten x_a und x_b, die mit Unsicherheiten δ_a und δ_b behaftet sind, durch die Summe der absoluten Unsicherheiten bestimmt werden kann. Für $x = x_a + x_b$ ergibt sich dann $\delta_x = \delta_a + \delta_b$. Diese Abschätzung der Unsicherheit wird aus zwei Gründen bevorzugt: Einerseits kann diese Abschätzung intuitiv auf grafischem Wege hergeleitet werden. Andererseits werden keine stochastischen Betrachtungen benötigt, die den Umfang der App inhaltlich sprengen würden. Diese Form der Unsicherheit kann jedoch jederzeit zur allgemeinen Form $\delta_x = \sqrt{\delta_a^2 + \delta_b}$ erweitert werden, indem die Unabhängigkeit der Unsicherheiten sowie die Standardabweichung als Unsicherheit betrachtet werden. Die App zum Bereich Direktes Messen führt als statistische Unsicherheit den Maximalfehler ein, aus dem intuitiv die Supremums-Abschätzung für die zusammengesetzte Unsicherheit motiviert werden kann.

Die Erarbeitung der Supremums-Abschätzung für die Addition wird ausführlich und interaktiv mit den Schüler:innen durchgeführt. Die minimalen und maximalen Fehler werden grafisch konstruiert und anschließend zur Unsicherheit der Addition überführt. Von einer analogen ausführlichen Betrachtung für die Subtraktion wird abgesehen. Den Schüler:innen wird übersichtlich anhand einer Erklärung aufgezeigt, dass die Subtraktion sich in diesem Fall genauso verhält. Auf eine erneute Erarbeitung wird verzichtet, da eine gleichartige Herleitung

der Subtraktion keinen didaktischen Mehrwert hat. Sofern die Lernenden das Prinzip beim Vorgang der Addition verinnerlicht haben, werden sie die Analogie der Subtraktion erkennen. Auf diese Weise kann der Fokus auf der einmaligen Erarbeitung beibehalten werden und die Schüler:innen können nach der Verallgemeinerung und der Zusammenfassung mit der Anwendung des neuen Konzepts beginnen.

Die Erarbeitung und Erklärung des Konzepts geschieht für den Fall von zwei Summanden. Diese sind ausreichend, um das Verhalten der Unsicherheiten im Fall einer Summe nachvollziehbar darzustellen. Das Erfassen der minimalen und maximalen Länge wird mit zwei Summanden durchgeführt und im Nachhinein für eine beliebige Anzahl von Summanden verallgemeinert. Auf diese Weise können sich die Schüler:innen auf eine Stelle konzentrieren, an der sich die Messwerte mit ihren Unsicherheiten berühren bzw. schneiden.

Konzeptaufbau nach Oser

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Bewusstmachung des Vorwissens	Start/Einleitung	Diese Phase der Lernumgebung dient dazu, den Schüler:innen den Kontext der Lernumgebung (100 m Sprint bzw. Staffel) zu nennen und das benötigte Vorwissen zu aktivieren (Bezug zur App A1).
Durcharbeiten eines Prototyps	Berechnung	Die Schüler:innen arbeiten kontextbasiert einen Fall durch, in dem zwei unsicherheitsbehaftete Messwerte addiert werden. Durch grafische Konstruktion eines minimalen und maximalen Werts im Rahmen der Messunsicherheit wird ein Bereich erzeugt, in dem sich die Summe der Messwerte befinden kann.

. . . Tabelle geht auf der nächsten Seite weiter

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Beschreibung der wichtigen Merkmale des neuen Konzepts	Erklärung	Beide Messwerte sind unsicherheitsbehaftet. Diese Eigenschaft muss auch bei der Berechnung der Summe berücksichtigt werden. Im vorigen Abschnitt wurden ein minimaler und ein maximaler Wert der Summe konstruiert. Dies tritt ein, wenn die Summanden sich an den Intervallgrenzen befinden, welche durch Messwert $\pm$ Unsicherheit dargestellt werden können. Wird also der minimale Wert für die Summe bestimmt, werden beide Unsicherheiten von dem Summenwert abgezogen. Im Fall des maximalen Werts werden beide Unsicherheiten addiert. Daraus ergibt sich, dass die Unsicherheit der Summe bestimmt werden kann, indem die Unsicherheiten der Summanden addiert werden.
Aktiver Umgang mit dem neuen Konzept	Aufgaben	Nach Verallgemeinerung des Spezialfalls werden nun kontextbasierte Aufgaben bearbeitet. Dabei werden zunächst Routinen abgefragt. Durch Variation der Aufgaben wird sichergestellt, dass das Konzept internalisiert wurde.
Anwenden des neuen Konzepts in anderen Kontexten	Aufgaben	Es folgen eine Variation der Aufgaben und eine neue Kontextorientierung für die restlichen Aufgaben dieser App. In diesem Fall können nicht nur alltagsnahe Kontexte, sondern auch fachphysikalische Inhalte und Kontexte in die App integriert werden.

Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene

Phase	Konzeptebene	Kontextebene
Allgemein (Ziele)	Schüler:innen wissen, dass Messwerte mit Unsicherheiten behaftet sind. Schüler:innen lernen, dass bei der Summe von unsicherheitsbehafteten Messwerten die absoluten Unsicherheiten addiert werden, um die Unsicherheit der Summe zu bestimmen.	100-Meter-Staffel von zwei Schüler:innen Die gemessenen Einzelzeiten sollen zu einer Gesamtzeit zusammengesetzt werden.
Einleitung	Vorwissen der Schüler:innen soll wiederholt und aktiviert werden. Ziel der App wird genannt: Bestimmung der Unsicherheit, wenn zwei unsicherheitsbehaftete Messwerte addiert werden.	Zwei Schüler:innen sind in einer Staffel 100 m gelaufen. Zeit wurde nacheinander von einem Mitschüler gemessen. Es wurden die Zeiten $(11{,}3 \pm 0{,}2)$ s und $(12{,}1 \pm 0{,}2)$ s ermittelt. Bestimme die Gesamtzeit und die dazugehörige Unsicherheit. (Ziel dieser Lernumgebung auf Kontextebene)

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptebene	Kontextebene
Berechnung	Darstellung der unsicherheitsbehafteten Messwerte	Darstellung der beiden Zeiten in einem Diagramm
	Bestimmung der Summe der Messwerte (ohne Unsicherheit), des minimalen und maximalen Werts der Summe (mit Unsicherheit) Daraus ergibt sich eine Darstellung mit dem kleinsten und größten Wert der Summe der Messwerte → Intervall, in dem sich die Summe der Messwerte befinden kann. Dieses kann nun verwendet werden, um die Unsicherheitsfortpflanzung im kommenden Abschnitt zu motivieren.	Die SuS sollen nun mit einem Schieberegler und mehreren verfügbaren Knöpfen die Summe der beiden Zeiten (die Messwerte→ ohne Berücksichtigung der Unsicherheiten), den minimalen (Best-Case-Szenario) und maximalen Wert (Worst-Case-Szenario) der Summe unter Berücksichtigung der Unsicherheiten bestimmen. Das soll erreicht werden, indem eine der Zeiten statisch ist und nicht verschoben werden kann, während die zweite angezeigte Zeit mit dem Schieberegler verschoben werden kann, bis die geforderten Werte bestimmt worden sind. Es wurde nun also ein Bereich erarbeitet, in dem sich die tatsächliche Staffelzeit der beiden Schüler:innen befindet.

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptebene	Kontextebene
Erklärung	Aus der vorigen Darstellung wird die Addition der absoluten Fehler im Fall der Summe motiviert und erläutert. Die Differenz wird als Strichrechnung kategorisiert und die Regel mit einem kurzen Beispiel, ohne Aktivität der Schüler:innen, ebenfalls mit der Addition der absoluten Fehler abgedeckt. Die vorangehend genannten Aspekte werden nicht in Form von Text dargestellt, sondern in der Abbildung grafisch hervorgehoben. Somit kann weiterhin mit der Darstellung gearbeitet werden, die die Schüler:innen erarbeitet haben, und die Lernumgebung ist nicht so textlastig.	Betrachte nun, wie sich jeweils die minimale und maximale Staffelzeit der Schüler:innen ergeben. Die minimale Zeit kommt zustande, wenn wir von den minimalen bzw. schnellsten möglichen Zeiten ausgehen (die US wird von den Messwerten abgezogen). Maximale Zeit entsteht, wenn von den maximalen bzw. langsamsten möglichen Zeiten ausgegangen wird (US werden zum Messwert addiert). Aus den Unsicherheiten der einzelnen Messwerte kann somit ein Bereich konstruiert werden, in dem sich alle möglichen Staffelzeiten befinden. Die Unsicherheiten der Messwerte bleiben also erhalten und ergeben die Unsicherheit der Staffelzeit. Werden unsicherheitsbehaftete Messwerte addiert, ist die Summe auch unsicherheitsbehaftet. Die Unsicherheit erhält man, indem man alle Unsicherheiten der Summanden zusammenfügt. → Überleitung zur Subtraktion bzw. Strichrechnung
Aufgaben		
Zusammenfassung	Wiederholung der wichtigsten Aspekte dieses Konzepts: Werden unsicherheitsbehaftete Messwerte addiert bzw. subtrahiert, ergibt sich die Unsicherheit als Summe der absoluten Unsicherheiten.	

A.2.2 B2 Addition von relativen Unsicherheiten

Konzept

Werden Messwerte mit Unsicherheiten multipliziert oder dividiert, ergibt sich die Unsicherheit des Ergebnisses, indem die relativen Unsicherheiten der Messwerte addiert werden.

Struktur

Ziel
Das Ziel dieser App ist das Konzept, dass die relativen Unsicherheiten addiert werden, um die Unsicherheit eines Produkts von zwei unsicherheitsbelasteten Messwerten zu ermitteln

Theoretischer Hintergrund
Einige physikalische Größen lassen sich nicht direkt oder nur mit großem Aufwand direkt messen. In diesen Fällen kann die eigentliche Messung über eine indirekte Messung durchgeführt werden. Aus den Messwerten kann die gesuchte physikalische Größe durch eine mathematische Modellierung bestimmt werden. Abhängig von der zugrunde liegenden mathematischen Modellierung muss die Unsicherheit des Ergebnisses bestimmt werden. Die entstehende Unsicherheit wird durch die Unsicherheiten der Eingangsgrößen beeinflusst, daher wird in diesem Fall von einer Unsicherheitsfortpflanzung gesprochen. Für eine beliebige mathematische Modellierung mit den unkorrelierten Eingangsgrößen $x_1, \ldots, x_n$ mit den dazugehörigen Unsicherheiten $\delta_1, \ldots, \delta_n$ und der mathematischen Modellierung $x = f(x_1, \ldots, x_n)$ kann die Unsicherheit über

$$\delta_x = \sqrt{\Sigma_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \delta_i \right)^2}$$

abgeschätzt werden (Möhrke & Runge, 2020, S. 89).

Diese Art der Unsicherheitsbestimmung entspricht einem ganzheitlichen Ansatz und wird im Sachstrukturmodell von Hellwig in der Dimension Zusammensetzung der Messunsicherheit aus mehreren Komponenten im Unteraspekt Fortpflanzung der Messunsicherheit adressiert. Sie setzt voraus, dass die gemessenen Werte und die dazugehörigen Unsicherheiten nicht miteinander korreliert sind.

Der Schwerpunkt dieser Lernumgebung liegt nicht darin, den oben genannten allgemeinen ganzheitlichen Ansatz zur Unsicherheitsfortpflanzung zu vermitteln, sondern eine schrittweise Bestimmung der Unsicherheit für den Spezialfall der Multiplikation bzw. Division von zwei gemessenen Werten darzustellen. Daraus

ergibt sich der Spezialfall $\delta_x = \sqrt{(\delta_1 \cdot x_2)^2 + (\delta_2 \cdot x_1)^2}$ für die Multiplikation ($x = x_1 \cdot x_2$) und der Spezialfall $\delta_x = \sqrt{\left(\frac{\delta_1}{x_2}\right)^2 + \left(\delta_2 \cdot \frac{x_1}{x_2}\right)^2}$ für die Division ($x = \frac{x_1}{x_2}$). Eine allgemeinere Betrachtung des Falls Multiplikation bzw. Division kann der Teildimension Schrittweise Bestimmung der Gesamtunsicherheit unter dem Punkt Produkte/Quotienten gemessener Größen entnommen werden.

Neben der Fortpflanzung der Unsicherheiten hat die Lernumgebung die mathematische Modellierung als Inhalt. Die mathematische Modellierung ist theoriebasiert und stellt somit den physikalischen Zusammenhang zwischen der zu ermittelnden Größe und den gemessenen Größen dar. Die Schüler:innen sollen erkennen, welche Grundlage die mathematische Modellierung besitzt und dass sie die Unsicherheitsfortpflanzung beeinflusst.

Didaktische Reduktion

Die Schüler:innen werden in dieser Lernumgebung nicht mit dem Spezialfall der Unsicherheitsfortpflanzung für die Division konfrontiert. Die pythagoreische Summe ist schwer zu motivieren und setzt eine stochastische Unabhängigkeit voraus, die im Rahmen dieser

Lernumgebung nicht behandelt wird. Es wird auf das Supremum für die Abschätzung der Unsicherheit zurückgegriffen. Die Schüler:innen lernen kennen, dass die Unsicherheit des Quotienten von zwei Messwerten x_a und x_b, die mit Unsicherheiten δ_a und δ_b behaftet sind, durch die Summe der relativen Unsicherheiten bestimmt werden kann. Für $x = \frac{x_a}{x_b}$ ergibt sich dann

$$\frac{\delta_x}{x} \approx \frac{\delta_a}{a} + \frac{\delta_b}{b}.$$

Diese Abschätzung der Unsicherheit wird aus den Gründen bevorzugt, dass in diesem Fall keine stochastischen Betrachtungen nötig sind und die Regel für die Unsicherheitsfortfplanzung über eine Einheitenbetrachtung motiviert werden kann. Diese Vorstellung der Unsicherheit kann jedoch jederzeit zur allgemeinen Form

$$\frac{\delta_x}{x} = \sqrt{\left(\frac{\delta_a}{a}\right)^2 + \left(\frac{\delta_b}{b}\right)^2}$$

erweitert werden, indem die Unabhängigkeit der Unsicherheiten sowie die Standardabweichung als Unsicherheit betrachtet werden. Die Lernumgebungen zum Bereich Direkten Messen führen als statistische Unsicherheit den Maximalfehler ein, aus der intuitiv die Supremums-Abschätzung für die Unsicherheit motiviert werden kann.

Die Erarbeitung der Regel wird mit den Schüler:innen gemeinsam durchgeführt und erfordert das Mitdenken dieser. Die Vorgehensweise und die neue Regel werden den Schüler:innen sukzessive präsentiert. In Form von Anweisungen müssen die Schüler:innen diesen Input direkt im gegebenen Kontext umsetzen. Die

Resultate der Schüler:innen werden kontrolliert und entsprechendes Feedback wird aufgezeigt. Die nächsten Abschnitte werden erst freigeschaltet, wenn die Schüler:innen den aktuellen Abschnitt vollständig korrekt bearbeitet haben. Dabei müssen die Schüler:innen keine selbstständigen Rechnungen durchführen, sondern die gegebenen Werte in die richtigen Stellen der vorgegebenen Rechnung einsetzen. Somit wird verhindert, dass durch etwaige Rechen- oder Rundungsfehler Abweichungen zustande kommen.

Die Erarbeitung und Erklärung des Konzepts geschieht für den Fall eines Quotienten. Dieser wird im Kontext einer Geschwindigkeitsmessung für den Schulweg realisiert. Der einfache Quotient $v = \frac{s}{t}$ ist ausreichend, um das Verhalten der Unsicherheiten im Fall eines Quotienten nachvollziehbar darzustellen. Das Verfahren wird im Nachhinein für eine beliebige Anzahl von Faktoren und Divisoren verallgemeinert.

Konzeptaufbau nach Oser

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Bewusstmachung des Vorwissens	Start/Einleitung	Diese Phase der Lernumgebung dient dazu, den Schüler:innen den Kontext der Lernumgebung (Geschwindigkeitsmessung für den Schulweg) zu nennen und das benötigte Vorwissen zu aktivieren (Bezug zu App A1).
Durcharbeiten eines Prototyps/ Beschreibung der wichtigen Merkmale des neuen Konzepts	Berechnung/ Erklärung	Die Schüler:innen arbeiten kontextbasiert einen Fall durch, in dem sie den Quotienten von zwei Messergebnissen bilden, welche mit Unsicherheiten behaftet sind (Strecke und Zeit). Dabei finden die Erarbeitung des Prototyps und die Beschreibung der Merkmale des neuen Konzepts parallel statt. Die Schüler:innen werden schrittweise durch die Unsicherheitsbestimmung des Quotienten geführt und müssen die neuen Informationen in der Prototypbearbeitung umsetzen. Nach und nach wird so die Unsicherheit des Quotienten generiert und die Schüler:innen erkennen die Bedeutung der relativen Unsicherheit für die Unsicherheitsfortpflanzung bei der Bestimmung von Quotienten.

. . . Tabelle geht auf der nächsten Seite weiter

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Aktiver Umgang mit dem neuen Konzept	Aufgaben	Nach Verallgemeinerung des Spezialfalls werden nun kontextbasierte Aufgaben bearbeitet. Dabei werden zunächst Routinen abgefragt. Durch Variation der Aufgaben wird sichergestellt, dass das Konzept internalisiert wurde.
Anwenden des neuen Konzepts in anderen Kontexten	Aufgaben	Es folgen eine Variation der Aufgaben und eine neue Kontextorientierung für die restlichen Aufgaben dieser Lernumgebung. In diesem Fall können nicht nur alltagsnahe Kontexte, sondern auch fachphysikalische Inhalte und Kontexte in die Lernumgebung integriert werden.

Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene

Phase	Konzeptebene	Kontextebene
Allgemein (Ziele)	Schüler:innen wissen, dass Messwerte mit Unsicherheiten versehen sind. Schüler:innen lernen, wie unsicherheitsbehaftete Messwerte miteinander dividiert (später multipliziert) werden.	Der Schulweg einer Schüler:in wurde mit Strecke und Zeit ausgemessen. Die Geschwindigkeit dieser Fahrt soll bestimmt werden. David möchte letzten Endes wissen, welchen Einfluss die unsicherheitsbehafteten Messwerte auf die Geschwindigkeit haben.
Start/ Einleitung	Vorwissen der Schüler:innen soll wiederholt und aktiviert werden. Ziel der Lernumgebung wird genannt: Bestimmung der Unsicherheit, wenn zwei unsicherheitsbehaftete Messwerte dividiert werden.	David stellt sich die Frage, wie schnell er auf seinem Schulweg mit dem Fahrrad ist. Für die Auswertung liegen ihm nun zwei unsicherheitsbehaftete Messwerte vor (Strecke und Zeit), die er zur Geschwindigkeit verrechnen möchte ($s = 2100(50)\,\mathrm{m}$ und $t = 330(30)\,\mathrm{s}$).
Berechnung/ Erklärung	Einführung von relativen Unsicherheiten und der Regel für die Unsicherheitsfortpflanzung der Division. Diese Phase wird in drei Unterphasen realisiert, wobei jede mit Aktivität der Schüler:innen verbunden ist.	

. . . Tabelle geht auf nächster Seite weiter

Phase	Konzeptebene	Kontextebene
	1. Berechnung der Geschwindigkeit ohne Berücksichtigung der Unsicherheiten	Zunächst berechnet David die Geschwindigkeit ohne Berücksichtigung der Unsicherheit. Schüler:innen helfen David dabei, indem sie die Messwerte in die vorbereiteten Felder eintragen. Dadurch müssen die Schüler:innen während des Prozesses mitdenken und sich konzentrieren. Entsprechend den Eingaben gibt es Feedback. Bei richtiger Bearbeitung wird der nächste Abschnitt freigeschaltet, ansonsten muss das Feld neu bearbeitet werden.
	2. Einführung und Motivation von relativen Unsicherheiten Vorschrift für die Bestimmung von relativen Unsicherheiten wird gegeben Für jeden der Messwerte wird einzeln die relative Unsicherheit bestimmt	David überlegt sich nun, wie die Unsicherheiten der Messwerte für die Geschwindigkeit berücksichtigt werden kann. Da die Unsicherheiten aufgrund der unterschiedlichen Einheiten nicht zusammengefügt werden können, muss ein anderer Weg betrachtet werden→ Einführung der relativen Unsicherheiten. Die relative Unsicherheit kann bestimmt werden, indem die Unsicherheiten jeweils durch die Messwerte geteilt werden. SuS werden ähnlich wie im vorigen Bereich zu einer Mitarbeit aufgefordert und müssen die entsprechenden Felder ausfüllen.
	3. Unsicherheit der Geschwindigkeit Gesamte relative Unsicherheit wird bestimmt Absolute Unsicherheit des Produkts wird auf dieser Grundlage bestimmt	Nun möchte David wissen, wie diese relativen Unsicherheiten zur Unsicherheit der Geschwindigkeit zusammengefügt werden können. Zunächst werden die Schüler:innen aufgefordert, die relativen Unsicherheit zu addieren. Anschließend wird dieser Wert mit dem Ergebnis der Geschwindigkeit aus der ersten Unterphase multipliziert→ Auch hier erfolgt die Eingabe für die Berechnung durch die SuS und ein Fortfahren innerhalb der App ist erst möglich, wenn einzelne Teilschritte richtig bearbeitet sind.

. . . Tabelle geht auf nächster Seite weiter

Phase	Konzeptebene	Kontextebene
	Abschließende Betrachtung zum Vorgehen und Verallgemeinerung der Regel auf beliebige Operationen der Punktrechnung	David nennt das Ergebnis für seine ermittelte Geschwindigkeit. Verallgemeinerung für Schüler:innen: Diese Vorgehensweise kannst du immer dann verwenden, wenn du unsicherheitsbehaftete Messwerte multiplizieren oder dividieren möchtest. Denn in diesen Fällen haben die Unsicherheiten stets unterschiedliche Einheiten und müssen zunächst verallgemeinert werden.
Aufgaben		
Zusammenfassung	Wiederholung der Eigenschaften der relativen Unsicherheit und Verallgemeinerung der Regel für die Unsicherheitsfortpflanzung bei der Quotientenbildung	

A.3 C Grafische Auswertung

A.3.1 C1 Unsicherheitsflächen

Konzept

Die Schüler:innen lernen, dass die Darstellung von unsicherheitsbelasteten Messwerten bzw. (Mess-)Wertepaaren in einem Koordinatensystem mit Unsicherheitsbalken in x- und y-Richtung verbunden ist. Darauf aufbauend erweitern die Lernenden die Unsicherheitsbalken zu Unsicherheitsflächen und verstehen die zugrunde liegende Idee der Unsicherheitsfläche.

Struktur

Ziel
Das Ziel dieser Lernumgebung ist es, den Schüler:innen zu vermitteln, dass die Darstellung eines oder mehrerer unsicherheitsbelasteter Messwertepaare zu Unsicherheitsflächen führt. Diese Unsicherheitsflächen repräsentieren den

Bereich, in dem sich das Messwertepaar unter Berücksichtigung der jeweiligen Unsicherheit befinden kann.

Theoretischer Hintergrund

Die Auswertung von experimentell ermittelten Messwerten zur Untersuchung von funktionalen Zusammenhängen zwischen physikalischen Größen kann unter anderem grafisch durchgeführt werden. Dazu werden die beiden betrachteten Größen in ein Koordinatensystem eingetragen, wobei jeweils eine der physikalischen Größen als unabhängige fungiert. Betrachtet man die physikalischen Größen a und b, so bedeutet das, dass beispielsweise Größe a auf der x-Achse aufgetragen wird (Abszisse) und auf der y-Achse die abhängige Größe b, hier modelliert als $f(a)$, (Ordinate). Für eine angemessene grafische Auswertung muss die Unsicherheit der beiden physikalischen Größen ebenfalls berücksichtigt werden. Das Konzept, dass die Unsicherheiten der beiden physikalischen Größen gemeinsam ein Unsicherheitsrechteck für jeden einzelnen Messwert erzeugen, lässt sich mit unterschiedlichen Aspekten des Sachstrukturmodells von Hellwig (2012) vereinen.

Zunächst einmal findet sich dieser Aspekt in der Dimension Grafische Regression einer linearen Regression wieder. Diese Dimension beinhaltet die Eigenschaften einer Ausgleichsgeraden eines linearen bzw. linearisierten funktionalen Zusammenhangs. Dabei wird erwähnt, dass eine korrekte Ausgleichsgerade durch alle Unsicherheitsrechtecke der Messwerte bzw. des Messwertepaars verläuft. Somit ist eine Voraussetzung für die grafische Regression, dass die Messwerte im Koordinatensystem mit Unsicherheitsrechtecken dargestellt werden. Den Schüler:innen wird in dieser Lernumgebung dieser grundlegende Bestandteil der Regression vermittelt, sodass sie in späteren Lernabschnitten in der Lage sind, Ausgleichsgeraden anzufertigen. Der Schwerpunkt in dieser Lernumgebung liegt darin, dass die Schüler:innen erkennen, wie aus den einzelnen Unsicherheitsintervallen der jeweiligen physikalischen Größe ein Unsicherheitsrechteck entsteht. Dazu wird eine interaktive Darstellung verwendet, die den Schüler:innen ermöglicht, den Messwert im Rahmen der Unsicherheit zu verschieben. Auf diese Weise kann entdeckt werden, dass aus den Unsicherheitsintervallen Unsicherheitsflächen resultieren.

Um die Unsicherheitsrechtecke adäquat anfertigen zu können, müssen die Unsicherheiten der Messwerte bestimmt und berücksichtigt werden. Dieser Vorgang kann als Aufstellen einer Unsicherheitsbilanz angesehen werden, bei der die Unsicherheiten aller Eingangsgrößen notiert werden. Dieser Aspekt findet sich im Konzept Zusammensetzung der Messunsicherheit aus mehreren Komponenten wieder. Die Schüler:innen müssen zunächst alle relevanten Unsicherheiten bestimmen, um die Unsicherheitsflächen zu erlangen.

Die resultierenden Unsicherheitsflächen bilden in anderen Lernumgebungen die Grundlage für die lineare Regression und den funktionalen Zusammenhang,

der sich daraus ergibt. In diesem Fall kann die Betrachtung der Unsicherheitsrechtecke ebenso als Unsicherheitsfortpflanzung auf grafischem Wege angesehen werden. Die Lernenden zeichnen ein Wertepaar ein, welches mit Unsicherheiten versehen ist, und konstruieren damit einen Bereich, in dem die Ausgleichsgerade bzw. in allgemeineren Fällen die Ausgleichskurve verläuft. Die Unsicherheitsflächen bestimmen somit den Bereich, in dem die grafische Regression durchgeführt werden kann.

Darauf aufbauend kann die Dimension Verlässlichkeit der Messung und ihres Ergebnisses mit dem Unteraspekt Grad des Vertrauens adressiert werden. Die Schüler:innen wissen aus anderen Lernumgebungen bereits, dass ein Unsicherheitsintervall einen Bereich angibt, in dem sich alle Messwerte befinden. Somit kann man mit einer gewissen Sicherheit sagen, dass sich in dem Intervall, welches vom Messwert und der dazugehörigen Unsicherheit aufgespannt wird, die Messwerte mit einer gewissen Sicherheit befinden. Das ist der erste Schritt, den Aspekt Grad des Vertrauens zu adressieren. In dieser Lernumgebung liegt der Fokus darauf, dass die Schüler:innen ein grundlegendes Verständnis der Unsicherheitsflächen und ihrer Bedeutung im Hinblick auf den Grad des Vertrauens erwerben.

Didaktische Reduktion

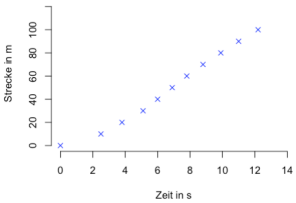
Die Lernumgebung legt den Fokus auf die Existenz der Unsicherheitsflächen und die Form ebendieser. Aus diesem Grund werden die Unsicherheiten vorgegeben, um eine eventuelle Hürde durch die Bestimmung der Unsicherheiten zu vermeiden. Außerdem handelt es sich in diesem Fall um absolute Unsicherheiten, die für jeden Messwert gleich groß sind. Daraus resultiert ein Spezialfall, nämlich dass die Unsicherheitsintervalle und die entstehenden Unsicherheitsflächen stets gleich groß sind.

Die entstehenden Flächen werden den Lernenden als Bereiche erklärt, in denen sich der Messwert befinden kann. Im Rahmen der Lernumgebung wird dementsprechend keine Wahrscheinlichkeitsdichtefunktion oder Ähnliches modelliert. Die Betrachtung der Unsicherheitsflächen und des Komplements ist binär, das bedeutet, dass sich außerhalb der Unsicherheitsflächen keine Messwerte befinden können. Dieser Ansatz wurde in den Lernumgebungen zum Konzept A Direktes Messen ebenfalls gewählt und wird erst im Konzept D Signifikanz im Konzept D2 Boxplots überarbeitet, wenn das Maß der Überlagerung zweier Messwerte betrachtet wird.

Konzeptaufbau nach Oser

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Bewusstmachung des Vorwissens	Start bzw. Einleitung	Der Start wird verwendet, um die Schüler:innen zu aktivieren. In diesem Fall wird der alltagsnahe Kontext des 100-Meter-Laufs eingeführt. Zudem werden die Voraussetzungen zu dieser App (die App zum Konzept A1 Streuung von Messwerten) wiederholt.
Durcharbeiten eines Prototyps	Darstellung der Messwerte	In dieser Phase wird den Lernenden ein Messwertepaar gegeben, welches in beiden Koordinaten mit einer Unsicherheit versehen ist, da beide physikalische Größen unsicherheitsbehaftet ermittelt worden sind. Der Messwert wird in einem Koordinatensystem dargestellt, in dem schrittweise die Unsicherheiten eingeblendet werden können. Im Rahmen der angezeigten Unsicherheiten kann nun der Messwert bewegt werden. Es entsteht eine Spur der Bewegung, die die Existenz der Unsicherheitsfläche suggeriert.
Beschreibung der wichtigen Merkmale des neuen Konzepts	Erklärung	Das Resultat aus dem vorigen Abschnitt wird verallgemeinert. Es wird erklärt, wie die Unsicherheitsfläche zustande kommt und dass jedes Messwertepaar mit einer entsprechenden Unsicherheitsfläche versehen werden muss.
Aktiver Umgang mit dem neuen Konzept	Erklärung/ Aufgaben	Es werden Aufgaben zum gleichen Kontext gestellt. Die Aufgabentypen werden variiert, um unterschiedliche Aspekte des neuen Konzepts zu betonen. Durch diese Abwechslung sollen das Verständnis und der Umgang mit dem Konzept verbessert werden.
Anwendung des neuen Konzepts in anderen Kontexten	Aufgaben	Der Inhalt der Aufgaben kann nun geändert werden. Sowohl Kontexte aus dem Alltag als auch fachwissenschaftliche Kontexte können nun Inhalt der Aufgaben sein.

Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene

Phase	Konzeptebene	Kontextebene
Allgemein (Ziele)	Schüler:innen wissen bereits, dass Messwerte mit Unsicherheiten behaftet sind (App A1) Schüler:innen sollen erkennen, dass die Darstellung von Messwerten mit Unsicherheiten im Koordinatensystem in Unsicherheitsflächen resultiert.	100-Meter-Lauf
Einleitung	Wiederholung des bereits erarbeiteten Wissens: Jede Messung ist mit einer Unsicherheit versehen. Ziel der App: Darstellung von Messwerten bzw. Messwertepaaren, die beide mit Unsicherheiten versehen sind.	Während des 100-Meter-Laufs wurde alle 10 m die Zeit erfasst. Es ergibt sich das folgende Diagramm.  Unsicherheit der Zeit → 0,3 s Unsicherheit der Strecke → 1 m Wie stellt man das dar?

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptebene	Kontextebene
Darstellung	Fokus auf ein Messwertepaar Variation des Messwertepaars bezüglich der vorhandenen Messunsicherheiten Beobachtung bzw. zusammenfassende Bemerkung, um zum nächsten Abschnitt überzuleiten.	Zunächst werden die Eingangsmesswerte gezeigt, danach wird der Fokus auf ein Messwertepaar gelegt (bspw. (2,5 s|10 m)). Unsicherheit der Zeit wird grafisch im Koordinatensystem angezeigt. Der Messwert kann über einen Schieberegler entlang des Intervalls variiert werden. Unsicherheit der Strecke wird grafisch im Koordinatensystem angezeigt. Der Messwert kann über einen Schieberegler entlang des Intervalls variiert werden. Es ist möglich, beide Schieberegler gleichzeitig zu verwenden. Besonderheit: Während der Messwert entlang der Unsicherheitsintervalle in x- und y-Richtung variiert wird, wird eine Spur entlang des Pfads generiert. Auf diese Weise kann die Unsicherheitsfläche grafisch motiviert werden. Die Schüler:innen können erkennen, dass die Unsicherheitsfläche ein Rechteck ergibt. Durch einen Knopf kann die Erkenntnis der Unsicherheitsflächen auf die restlichen Messwertepaare der Messreihe übertragen werden. Die Unsicherheitsflächen sind eine Eigenschaft von unsicherheitsbehafteten Messwerten und nicht von einem betrachteten Einzelfall.
Aufgaben		

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptebene	Kontextebene
Zusammenfassung	Werden Messwerte mit Unsicherheiten im Koordinatensystem dargestellt, so ergeben die vorhandenen Unsicherheitsintervalle eine Unsicherheitsfläche. Der gemessene Wert kann sich irgendwo in der Unsicherheitsfläche befinden.	Sowohl die Unsicherheit der Zeit als auch der Strecke lassen sich grafische darstellen. Dabei werden die beiden Unsicherheitsintervalle zu einer Unsicherheitsfläche erweitert, in denen der Messwert sich beliebig befinden kann. Unsicherheitsflächen müssen immer dann berücksichtigt werden, wenn unsicherheitsbehaftete Messwerte in das Koordinatensystem eingezeichnet werden.

A.3.2 C2 Ausgleichsgerade

Konzept

Die Schüler:innen lernen, was eine Ausgleichsgerade ist und welche Eigenschaft diese hat: Die Ausgleichsgerade muss alle Unsicherheitsflächen berühren bzw. schneiden. Liegt dieser Fall vor, so handelt es sich um einen linearen Zusammenhang. In anderen Fällen ist der vermutete lineare Zusammenhang zu verwerfen.

Struktur

Ziel

Schüler:innen lernen das Konzept einer Ausgleichsgeraden kennen. Sie fertigen Ausgleichsgeraden zu verschiedenen Datensätzen an und entscheiden, ob eine Ausgleichsgerade existiert bzw. von einem linearen Zusammenhang ausgegangen werden kann. Als Entscheidungskriterium wird das Schneiden aller Unsicherheitsflächen durch die Ausgleichsgerade verwendet.

Theoretischer Hintergrund

Liegt ein linearer funktionaler Zusammenhang zwischen zwei betrachteten physikalischen Größen vor, so kann eine Ausgleichsgerade für die grafische Regression verwendet werden. Die Messwertepaare müssen mit entsprechenden Unsicherheitsrechtecken im Koordinatensystem dargestellt sein. Schneidet eine Ausgleichsgerade alle Unsicherheitsrechtecke im Koordinatensystem, so kann von einem linearen bzw. linearisierten Zusammenhang der physikalischen Größen ausgegangen werden. Diese Aspekte werden im Folgenden auf Dimensionen und Konzepte des Sachstrukturmodells von Hellwig (2012) zurückgeführt.

Der Fokus der Lernumgebung liegt in der Vermittlung der Eigenschaften einer Ausgleichsgeraden. Diese finden sich im Sachstrukturmodell unter der Dimension Grafische Regression einer linearen Regression wieder. Die Schüler:innen sollen lernen, dass die Ausgleichsgerade alle Unsicherheitsflächen der Messwerte schneiden muss. Nur wenn diese Bedingung erfüllt ist, kann von einem linearen funktionalen Zusammenhang der beiden dargestellten physikalischen Größen ausgegangen werden. Außerdem sollen die Schüler:innen vermittelt bekommen, wie das Ergebnis zu interpretieren ist, wenn die Eigenschaften einer Ausgleichsgeraden nicht erfüllt sind. Hierzu befindet sich in der gleichen Dimension des Sachstrukturmodells die Antwort: Schneidet eine Ausgleichsgeraden nicht alle Unsicherheitsflächen, so liegt entweder kein funktionaler Zusammenhang vor oder die Unsicherheiten der Messwerte wurden nicht adäquat abgeschätzt.

Die Unsicherheitsflächen geben somit einen Bereich vor, in dem sich der tatsächliche Messwert (hier der ermittelte funktionale Zusammenhang) befindet. Anhand der Größe der Unsicherheitsflächen können somit Rückschlüsse auf die Güte der Ausgleichsgerade gezogen werden. Die Lernenden sollen erkennen, dass durch die Unsicherheitsflächen ein Intervall für die Schätzung des funktionalen Zusammenhangs anhand der Ausgleichsgerade aufgestellt wird. Je nach Größe und Anordnung der Messwerte sind ebenso mehrere Ausgleichsgeraden möglich. Dieser Aspekt findet sich in der Dimension Verlässlichkeit der Messung und ihres Ergebnisses unter dem Aspekt Genauigkeit des Schätzwerts wieder. Damit einher geht der Grad des Vertrauens für den funktionalen Zusammenhang. Die Unsicherheitsflächen werden anhand der Unsicherheitsintervalle der Eingangsgrößen bestimmt und repräsentieren den Bereich, in dem sich der Messwert mit einer gewissen Sicherheit befinden kann. Die Ausgleichsgerade muss alle Unsicherheitsflächen schneiden, um den funktionalen Zusammenhang auf Grundlage des Grads des Vertrauens zu ermitteln. Der Grad des Vertrauens findet sich im Sachstrukturmodell in der Dimension Verlässlichkeit der Messung und ihres Ergebnisses wieder.

Didaktische Reduktion

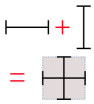
Die grafische Regression wird für die Schüler:innen wie folgt vereinfacht. Sobald die Ausgleichsgerade alle Unsicherheitsflächen schneidet, kann der lineare Zusammenhang zwischen den Größen bestätigt werden. Es wird nicht unterschieden, in welchem Maß die Ausgleichsgerade die Unsicherheitsflächen schneidet. Eine Ausgleichsgerade, die die Unsicherheitsflächen nahezu mittig schneidet, und eine weitere Ausgleichsgerade, die die Unsicherheitsflächen größtenteils nur tangiert, werden als gleichwertig angesehen bzw. vermittelt.

Der Schwerpunkt der Lernumgebung liegt auf dem Anfertigen der Ausgleichsgeraden und dem Verständnis des Konzepts, dass diese alle Unsicherheitsflächen schneiden muss. Bei der Darstellung und Auswertung der Messwerte wird daher auf lineare Zusammenhänge zurückgegriffen. Theoretisch könnten ebenso

linearisierte Zusammenhänge dargestellt und ausgewertet werden. In diesen Fällen wäre es jedoch nötig, den physikalischen Zusammenhang zunächst zu linearisieren, um die entsprechende Darstellung zu erläutern. Dieser Schritt ist durchaus adäquat für Schüler darstellbar, der Fokus der Lernumgebung würde in diesem Fall jedoch etwas verschoben werden und für die Schüler:innen wäre der Inhalt bzw. das Konzept, welches vermittelt werden soll, nicht mehr eindeutig.

Diese Lernumgebung ist der zweite Teil eines dreiteiligen Bearbeitungsstrangs, der die grafische Auswertung von Daten zum Ziel hat. Wie bereits in der vorangehenden Lernumgebung wird der Spezialfall betrachtet, dass alle Unsicherheitsflächen gleich groß sind. Im erklärenden Teil der Lernumgebung wird nur auf die Eigenschaften der Ausgleichsgeraden eingegangen. Der Umgang mit einem Datensatz, falls das Anfertigen einer Ausgleichsgeraden nicht möglich ist, wird nicht besprochen bzw. aufgegriffen. Das Vorliegen eines Ausreißers oder zu kleiner Unsicherheitsintervalle bzw. -flächen wird nicht berücksichtigt und nicht besprochen.

Konzeptaufbau nach Oser

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Bewusstmachung des Vorwissens 	Start bzw. Einleitung	Schüler:innen aktivieren und ihnen bewusst machen, wie Messwerte bzw. Paare im Koordinatensystem dargestellt werden. Dazu wird den Schüler:innen eine Abbildung gezeigt, die die Entstehung von Unsicherheitsflächen darstellt. Es ist notwendig, dass die Schüler:innen wissen, dass Unsicherheitsflächen entstehen.
Durcharbeiten eines Prototyps	Ausgleichsgerade	Die Schüler:innen sollen eine Gerade anfertigen. Die Lernenden untersuchen, ob sie auf dem Schulweg stets dieselbe Geschwindigkeit haben bzw. den gleichen Weg pro Zeit zurücklegen.

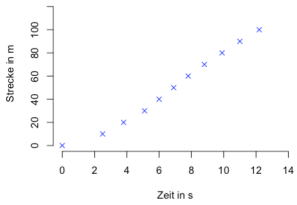
. . . Tabelle geht auf der nächsten Seite weiter

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Beschreibung der wichtigen Merkmale des neuen Konzepts	Erklärung	Es erfolgt eine Definition der Ausgleichsgeraden als Gerade, die sämtliche Unsicherheitsflächen schneidet. Gibt es eine Gerade mit dieser Eigenschaft, so können die Messwerte als auf einer Gerade liegend betrachtet werden. Dabei ist an dieser Stelle nicht relevant, ob es die „beste“ Ausgleichsgerade ist. Das ist Ziel der nachfolgenden App C3 Drei-Geraden-Methode.
Aktiver Umgang mit dem neuen Konzept	Erklärung bzw. Aufgaben	Es werden Aufgaben zum gleichen Kontext gestellt. Die Aufgabentypen werden variiert, um unterschiedliche Aspekte des neuen Konzepts zu betonen. Durch diese Abwechslung sollen das Verständnis und der Umgang mit dem Konzept verbessert werden.
Anwenden des neuen Konzepts in anderen Kontexten	Aufgaben	Der Inhalt der Aufgaben kann nun geändert werden. Sowohl Kontexte aus dem Alltag als auch fachwissenschaftliche Kontexte können nun Inhalt der Aufgaben sein.

Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene

Phase	Konzeptuelle Ebene	Kontextuelle Ebene
Allgemein (Ziel)	Die Schüler:innen wissen bereits, dass Wertepaare, die mit Unsicherheiten versehen sind, im Koordinatensystem mit Unsicherheitsflächen dargestellt werden (Konzepte der App C1). Schüler:innen sollen lernen, dass die Unsicherheitsflächen ein Kriterium für die Legitimation einer Ausgleichsgeraden darstellen.	Bestimmung der Schulweglänge

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptuelle Ebene	Kontextuelle Ebene
Einleitung	Wiederholung, dass Messwerte, die mit Unsicherheiten versehen sind, im Koordinatensystem mit Unsicherheitsflächen dargestellt werden. Ziel dieser Lernumgebung: Solche Daten mit einer Ausgleichsgeraden auswerten	Schüler:innen haben bei der Messung der Schulweglänge die zurückgelegte Strecke sowie die dafür benötigte Zeit aufgezeichnet. Es soll unter anderem untersucht werden, ob die Geschwindigkeit v konstant ist. Die dazugehörigen Werte sind in dem folgenden s-t-Diagramm dargestellt.  Untersuche, ob die Messwerte auf einer Geraden liegen.

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptuelle Ebene	Kontextuelle Ebene
Vermessung / Darstellung	Darstellung der Messwerte ohne Unsicherheitsflächen Erste Anfertigung einer Ausgleichsgeraden: Schüler:innen haben kein Kriterium, wann die Ausgleichsgerade gültig ist und wann die Bestimmung der Geraden beendet ist.	Das Eingangsdiagramm wird den Schüler:innen erneut gezeigt. Zu sehen sind Messwerte, die aufgenommen wurden, als die Schüler:innen für Aufgaben der letzten Woche die Länge des Schulwegs bestimmt haben (App B2). Im ersten Fall werden die Unsicherheiten der Messwerte nicht angezeigt.
	Darstellung der Messwerte mit Unsicherheitsflächen Zweite Anfertigung der Ausgleichsgeraden: SuS erkennen, dass die Ausgleichsgerade die US-Flächen in einigen Fällen schneidet, und variieren die Ausgleichsgerade derart, dass sie durch möglichst viele bzw. alle Unsicherheitsflächen geht.	Schüler:innen bekommen den Arbeitsauftrag, eine gültige Gerade durch den Datensatz zu zeichnen, um zu überprüfen, ob die aufgezeichneten Werte auf einer Linie liegen. Auf die damit einhergehende lineare Funktionalität wird später eingegangen. Die Werte streuen und liegen nicht exakt auf der Geraden. Die Schüler:innen werden keine Gerade anfertigen können, auf der alle vorhandenen Messwerte liegen.
	Zusammenfassung	
		Anschließend werden die Unsicherheiten der Strecke $\delta_s = 43\,\mathrm{m}$ und der benötigten Zeit $\delta_t = 2\,\mathrm{min}$ dargestellt und zu Unsicherheitsflächen erweitert.
		Die Schüler:innen haben nun Gelegenheit, eine neue Gerade zu zeichnen. Die vorige Gerade wird weiterhin angezeigt. So können die Schüler:innen vergleichen, wo die Unterschiede liegen. Durch die Variation der Geraden können die Schüler:innen erkennen, dass sie so verschoben werden kann, dass sie alle Unsicherheitsflächen zumindest berührt.
		Die Gerade kann derart konstruiert werden, dass alle Unsicherheiten geschnitten werden.

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptuelle Ebene	Kontextuelle Ebene
Erklärung	Schwerpunkt: Die Güte der Ausgleichsgeraden kann mittels der Unsicherheitsflächen bestimmt werden: Werden alle Unsicherheitsflächen von der Ausgleichsgeraden geschnitten, so liegen die Werte auf der Geraden und es liegt ein linearer Zusammenhang vor	Kann man eine Gerade so durch die Daten zeichnen, dass alle Unsicherheiten geschnitten werden, so liegen die Werte im Rahmen der Messunsicherheiten auf einer Geraden. Eine Gerade, die diese Eigenschaft hat, nennt man Ausgleichsgerade. In diesem Fall bedeutet es, dass die Schüler:in, die die Messwerte aufgezeichnet hat, pro Minute die gleiche Strecke zurückgelegt hat → ihre Geschwindigkeit war also konstant.
	Verallgemeinerung: Die Methode ist immer anwendbar, wenn überprüft werden soll, ob die Messwerte auf einer Geraden liegen oder nicht. zZ überprüfendes Kriterium bleibt weiterhin, ob alle Unsicherheitsflächen geschnitten werden.	Diese Methode kann man immer verwenden, wenn man überprüfen möchte, ob aufgezeichnete Messwerte auf einer Geraden liegen.
	Liegen die Messwerte auf einer Geraden, so spricht man von einem linearen Zusammenhang.	Liegen die Messwerte auf einer Geraden (man kann eine Ausgleichsgerade konstruieren), so liegt ein linearer Zusammenhang zwischen den beiden Größen vor.
Aufgaben		
Zusammenfassung	Der Zusammenhang zwischen zwei Größen kann mittels einer Ausgleichsgeraden untersucht werden. Liegen die Messwerte auf der sogenannten Ausgleichsgeraden, besteht ein linearer funktionaler Zusammenhang zwischen den Größen. Damit eine Gerade als Ausgleichsgerade bezeichnet werden kann, muss diese alle Unsicherheitsflächen schneiden.	Eine Ausgleichsgerade ist eine Gerade, die durch alle Unsicherheitsflächen der gegebenen Messwerte verläuft. Auf diese Weise kann man überprüfen, ob die Messwerte auf einer Geraden liegen oder nicht.

A.3.3 C3 Drei-Geraden-Methode

Konzept

Die Schüler:innen erkennen, dass das Anfertigen einer Ausgleichsgeraden bei einer linearen Regression nicht eindeutig ist. Die sinnvollste Ausgleichsgerade ergibt sich aus der Drei-Geraden-Methode.

Inhaltlicher Schwerpunkt

Ziel

In dieser Lernumgebung sollen die Schüler:innen erkennen, dass die Ausgleichsgerade zu einem gegebenen Kontext nicht eindeutig ist, indem sie mehrere Ausgleichsgeraden zum selben Datensatz anfertigen. Insbesondere soll darauf aufbauend die Drei-Geraden-Methode eingeführt werden, die auf grafischem Wege die sinnvollste Ausgleichsgerade konstruiert (Hellwig, 2012, S. 80 f.).

Theoretischer Hintergrund

Ein linearer funktionaler Zusammenhang zwischen zwei betrachteten physikalischen Größen kann mittels einer Ausgleichsgeraden ermittelt werden. Die grafische Auswertung von unsicherheitsbehafteten Daten wird in drei Lernumgebungen aufgearbeitet. In den vorangehenden Lernumgebungen des Strangs Grafische Auswertung haben die Schüler:innen bereits kennengelernt, dass die Unsicherheitsintervalle der Eingangsgrößen zu Unsicherheitsrechtecken erweitert werden und dass eine Ausgleichsgerade sämtliche Unsicherheitsrechtecke schneiden muss, damit von einem linearen funktionalen Zusammenhang ausgegangen werden kann. Beim virtuellen Anfertigen der Ausgleichsgeraden in der vorangehenden Lernumgebung könnten Schüler:innen bereits bemerkt haben, dass die Ausgleichsgerade nicht eindeutig ist. Sie ist im Rahmen der Unsicherheitsflächen oft etwas variabel. Während die optimale Ausgleichsgerade beispielsweise mit der Methode der kleinsten Quadrate bestimmt werden kann, ist der Fokus dieser Lernumgebung auf einer grafischen Methode, die den Schüler:innen eine sinnvolle Ausgleichsgerade ermöglicht.

Die Drei-Geraden-Methode findet sich in der Dimension Regression im Konzept Grafische Regression einer linearen Regression im Sachstrukturmodell von Hellwig (2012). Zunächst werden zwei Ausgleichsgeraden angefertigt, welche mit maximaler (Ausgleichsgerade A) bzw. minimaler Steigung (Ausgleichsgerade B) sämtliche Unsicherheitsflächen schneiden. Diese Ausgleichsgeraden spannen einen Schlauch bzw. Bereich auf, in dem sich alle möglichen Ausgleichsgeraden zum gegebenen Datensatz befinden. Für die Ermittlung der sinnvollen Ausgleichsgeraden wird die Winkelhalbierende der Ausgleichsgeraden A und B ermittelt. Diese Ausgleichsgerade schneidet konstruktionsbedingt sämtliche

Unsicherheitsflächen und bildet eine Art Durchschnitt aller möglichen Ausgleichsgeraden (angenommen, alle möglichen Ausgleichsgeraden innerhalb des Schlauchs bzw. Bereichs sind gleichverteilt, ergibt sich die Winkelhalbierende als Erwartungswert dieser Verteilung). Diese beiden Eigenschaften charakterisieren sie als sinnvollste Ausgleichsgerade.

Didaktische Reduktion

Für die Gestaltung der Lernumgebung wurden einige Aspekte aus didaktischer Perspektive vereinfacht, um den Fokus auf die Drei-Geraden-Methode zu richten. In der Lernumgebung werden stets lineare Zusammenhänge dargestellt und keine Zusammenhänge, die erst durch eine geeignete Darstellung linear erscheinen. Einerseits werden in den Bildungs- und Rahmenlehrplänen der unterschiedlichen Bundesländer in der Sekundarstufe kaum andere funktionale Zusammenhänge betrachtet. Andererseits müsste bei einer linearisierten Darstellung zunächst erläutert werden, wie diese zustande kommt. Wenngleich dies für Schüler:innen nachvollziehbar und überzeugend dargestellt werden kann, würde sich dadurch der inhaltliche Schwerpunkt der Lernumgebung verschieben.

Die Methode zur Bestimmung der sinnvollsten Ausgleichsgeraden wird nur auf grafischem Wege vermittelt. Eine mathematische Motivation über den Durchschnitt oder den Erwartungswert der Steigungen im Bereich, der von den Ausgleichsgeraden A und B aufgespannt wird, wird den Schüler:innen nicht mitgegeben. Die Winkelhalbierende wird aufgrund ihrer Lage motiviert und den Schüler:innen als eine Art Durchschnitt aller möglichen Ausgleichsgeraden in dem Bereich vermittelt.

Die Grenzen der Methode werden in der Lernumgebung nicht thematisiert. Die ideale Lösung, beispielsweise über die Methode der kleinsten Quadrate ermittelt, stimmt im Allgemeinen nicht mit der Ausgleichsgeraden überein, die unter Verwendung der Drei-Geraden-Methode ermittelt wird. Aus diesem Grund wird in der Umsetzung und Darstellung der Methode darauf geachtet, dass die ermittelte Lösung nicht als beste, sondern grafisch am sinnvollsten zu ermittelnde bezeichnet wird. Ebenso wird vermieden, andere Ausgleichsgeraden als schlechter oder falsch zu bezeichnen. Die sinnvolle Lösung soll eine Möglichkeit für die Lernenden darstellen, sich für eine Ausgleichsgerade zu entscheiden.

Die verwendeten Datensätze der Lernumgebung sind so ausgewählt, dass es keine Ausreißer gibt und dass stets Ausgleichsgeraden für die grafische Regression möglich sind. Diese Tatsache liegt in der didaktischen Reduktion der vorangehenden Lernumgebung begründet.

Konzeptbildung nach Oser

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Bewusstmachung des Vorwissens	Start bzw. Einleitung	Das Vorwissen der Lernenden wird aktiviert, indem ein Streitgespräch zwischen zwei Schüler:innen bezüglich Ausgleichsgeraden initiiert wird. So werden sie an die Eigenschaften einer Ausgleichsgeraden erinnert. Dieser Schritt ist essenziell und notwendig für das Verständnis und die erfolgreiche Bearbeitung dieser Lernapp.
Durcharbeiten eines Prototyps	Erarbeitung bzw. Erklärung	Die Schüler:innen fertigen zunächst eine Ausgleichsgerade mit minimaler und maximaler Steigung an. Dadurch wird der Bereich definiert, in dem sich die Ausgleichsgeraden befinden dürfen. Nun können sie sich überlegen, wie auf dieser Grundlage eine sinnvolle Lösung des Problems gefunden werden kann. Den Schüler:innen wird die Winkelhalbierende als Lösungsvorschlag präsentiert, die durch Betätigen eines Schalters eingeblendet werden kann.
Beschreibung der wichtigen Merkmale des neuen Konzepts	Erklärung	Die sinnvollste Lösung in dieser Lernapp stellt die Winkelhalbierende dar. Sie halbiert den Bereich, in dem sich alle Ausgleichsgeraden befinden können, und ist eine Art Durchschnitt der Ausgleichsgeraden. Die Winkelhalbierende kann auf grafischem Wege einfach ermittelt werden und stellt eine sinnvolle Ausgleichsgerade für eine grafische Regression einer linearen Regression dar.
Aktiver Umgang mit dem neuen Konzept	Aufgaben	Die Schüler:innen müssen Aufgaben im noch zu bestimmenden Kontext lösen. Um ein tiefes Verständnis zu fördern, werden die Aufgaben variiert.

. . . Tabelle geht auf der nächsten Seite weiter

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Anwenden des neuen Konzepts in anderen Kontexten	Aufgaben	Die vorangehende Kontextorientierung wird aufgelöst, die Schüler:innen werden nun mit weiteren alltagsnahen oder fachphysikalischen und kontextorientierten Aufgaben konfrontiert, die für eine erfolgreiche Bewältigung eine Anwendung des neuen Konzepts benötigen.

Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene

Phase	Konzeptuelle Ebene	Kontextuelle Ebene
Allgemein (Ziel)	Schüler:innen kennen bereits das Konzept der Ausgleichsgeraden. Nach Bearbeitung dieser App können die Schüler:innen die Drei-Geraden-Methode anwenden.	Fahrt mit dem E-Scooter Schüler:innen haben eine Fahrt mit E-Scooter gemacht und dabei für die zurückgelegte Strecke die Zeit aufgenommen.
Einleitung	Vorwissen der Schüler:innen und Voraussetzung für diese App werden genannt: Schüler:innen können Ausgleichsgeraden anfertigen (App C2). Damit einhergehend müssen Schüler:innen wissen, was Unsicherheitsflächen sind (App C1). Ziel: Schüler:innen lernen eine Methode, um eine sinnvolle Ausgleichsgerade zu zeichnen.	Gespräch zwischen Schüler:innen Schüler:innen unterhalten sich darüber, dass sie zum gegebenen s-t-Diagramm ihrer E-Scooter-Fahrt unterschiedliche Ausgleichsgeraden angefertigt haben. Es wird herausgearbeitet, dass beide Ausgleichsgeraden legitim sind, d. h. sie schneiden sämtliche Unsicherheitsflächen. In diesem Kontext bedeutet eine höhere Steigung des Graphen eine höhere Geschwindigkeit. Die Schüler:innen diskutieren also über unterschiedliche Geschwindigkeiten der Fahrt und müssen nun überlegen, ob und wie man sich diesbezüglich einigen kann.

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptuelle Ebene	Kontextuelle Ebene
Erarbeitung	Zunächst werden zwei Ausgleichsgeraden von den Schüler:innen angefertigt: Geraden mit maximaler und minimaler Steigung Es ergibt sich ein Bereich, in dem sich alle weiteren Ausgleichsgeraden befinden Der Lösungsweg wird den Schüler:innen in Form von direkten Instruktionen vermittelt: Der Lernprozess kann auf diese Weise gesteuert werden. Das Entdecken der Winkelhalbierenden und ihrer Eigenschaft als sinnvolle Lösung ist über forschende und entdeckende Methoden schwer zeiteffizient umzusetzen.	Schüler:innen werden angewiesen, die Gerade einzuzeichnen, die die geringste Steigung aufweist. Dies entspricht der langsamsten Fahrt. Der vorige Prozess wird wiederholt, wobei in diesem Fall die maximale Steigung und somit die höchste Geschwindigkeit eingezeichnet werden soll. Den Schüler:innen wird nun der Bereich visualisiert, in dem sich alle weiteren Ausgleichsgeraden befinden. Damit ist auch ein Bereich für alle Geschwindigkeiten festgelegt. Durch Knopfdruck kann nun die Winkelhalbierende eingezeichnet werden. Diese geht mittig durch den erarbeiteten Bereich und wird anschließend als sinnvollste Lösung motiviert.

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptuelle Ebene	Kontextuelle Ebene
Erklärung	Motivation und Einblendung der Winkelhalbierenden Die Schüler:innen konstruieren die Winkelhalbierende nicht selbstständig, sondern können diese auf Grundlage der bestimmten Ausgleichsgeraden mit minimaler und maximaler Steigung einblenden. Einerseits soll so die repetitive Tätigkeit des Konstruierens möglichst gering gehalten werden und andererseits ist es schwer, die exakte Winkelhalbierende grafisch am Computer durch Variation der Parameter mittels Schiebereglern zu ermitteln. Die Winkelhalbierende als Ausgleichsgerade ist sinnvoll, da sie in der Mitte des Bereichs liegt, der alle Möglichkeiten abdeckt und somit eine Art Durchschnitt darstellt.	Die Winkelhalbierende halbiert den Bereich aller möglichen Ausgleichsgeraden. Die Hälfte der Geraden verläuft steiler und die Hälfte flacher. Daher ist diese Gerade gut geeignet, um die Messwerte auszuwerten. Die Winkelhalbiere ist also eine Art Durchschnitt. Sie kann verwendet werden, falls es mehrere Möglichkeiten gibt, Ausgleichsgeraden anzufertigen. In diesem Fall entspricht die Steigung dieser Ausgleichsgeraden einer Art von Durchschnittsgeschwindigkeit. Die Drei-Geraden-Methode bestimmt also auf grafischem Wege eine sinnvolle Ausgleichsgerade. Achtung: Das bedeutet nicht, dass alle anderen Ausgleichsgeraden nicht korrekt sind. Solange alle Kriterien für die Ausgleichsgeraden erfüllt sind, sind diese stets richtig. Die Drei-Geraden-Methode soll eine Möglichkeit darstellen, sich für eine Ausgleichsgerade zu entscheiden.
Aufgaben		
Zusammenfassung	Die wichtigsten Merkmale der Drei-Geraden-Methode werden genannt. Es wird hervorgehoben, dass es nicht die einzig richtige Ausgleichsgerade ist, sie jedoch oftmals die sinnvollste ist. Einheitliches Design mit Start-Abschnitt der Lernumgebung: Realisierung des Zusammenfassungs-Abschnitts über ein Concept-Cartoon	Dialog zwischen den Schüler:innen, der die Richtigkeit der unterschiedlichen Ausgleichsgeraden aufgreift. Um eine sinnvolle Gerade zu bestimmen, kann die Drei-Geraden-Methode verwendet werden. Dabei ermittelt man die Ausgleichsgeraden mit minimaler und maximaler Steigung. Bezug auf den Kontext: Von allen möglichen Geschwindigkeiten, die man auf Grundlage dieser Messwerte ermitteln kann, kann ein Durchschnittswert für alle Möglichkeiten bestimmt werden.

A.4 D Signifikanz

A.4.1 D2 Boxplots

Konzept

Schüler:innen können Messwerte durch einen Boxplot darstellen und diese Darstellung verwenden, um unterschiedliche Messergebnisse miteinander zu vergleichen.

Inhaltlicher Schwerpunkt

Ziel

Das Ziel dieser Lernumgebung ist es, dass die Lernenden die Messwerte mit einem Boxplot darstellen und auswerten können. Das bedeutet insbesondere, dass sie die Bedeutung der Quartile verstanden haben. Den Lernenden soll bewusst gemacht werden, dass die Darstellung des Boxplots die Messwerte in Viertel aufteilt und somit mehr Informationen darstellt als die reine grafische Visualisierung mit Messwert und Messunsicherheit.

Theoretischer Hintergrund

Der Vergleich von Messergebnissen untereinander und mit Referenzwerten ist im Rahmen der Auswertung ein wesentlicher Bestandteil eines Experiments. Gegeben seien die Messergebnisse $x_1 = w_1 \pm u_1$ und $x_2 = w_2 \pm u_2$. In der ersten App zur Reihe Signifikanz (D1) wurde den Schüler:innen vermittelt, dass zwei Messergebnisse miteinander verträglich sind, wenn $|w_1 - w_2| \leq u_1 + u_2$ gilt. Diese Idee ist zwar im metrologischen Sinne nicht falsch, jedoch lässt sich die Verträglichkeit mit dem Normalized Error Ratio etwas systematischer bewerten. Das Normalized Error Ratio ist ein Maß für die Verträglichkeit der Messergebnisse. Im grafischen Sinne bedeutet das, dass das Verhältnis näher an 0 ist, desto stärker die Messergebnisse überlappen. Das Normalized Error Ratio ist über $E_N = |w_1 - w_2|/\sqrt{u_1^2 + u_2^2}$ definiert (Steele & Douglas, 2006, S. 235). Bezüglich der Interpretation des E_N-Werts gibt es verschiedene Interpretationen (Steele & Douglas, 2006, S. 235). Die vorliegende Ausarbeitung orientiert sich an den Grenzwerten, welche von Hellwig (2012, S. 78) verwendet werden. Zwei Messergebnisse sind demnach im metrologischen Sinne verträglich und nicht voneinander unterscheidbar, wenn $E_N \leq 0{,}5$ gilt. Bei einem E_N–Wert von größer als 1 liegen die Messergebnisse so weit auseinander, dass es sich um unterschiedliche Ergebnisse handeln muss.

Diese Kategorisierung unter Zuhilfenahme des Normalized Error Ratio führt nicht immer zu denselben Resultaten bezüglich der Verträglichkeit wie mit dem Konzept, welches die Schüler:innen aus der Lernumgebung zum Vergleich von Mittelwerten kennen. Sie müssen somit entsprechend darauf vorbereitet und hingeführt werden. Als gute Möglichkeit bietet sich in diesem Fall der Boxplot an. Mithilfe der Quantile lässt sich einerseits die Verteilung der Messwerte und des Messergebnisses zügig darstellen und andererseits können Aussagen darüber getroffen werden, wie stark die Verträglichkeit bzw. die den Schüler:innen bekannte Überlappung ausgeprägt ist. Auf diese Weise wird ein Bewusstsein dafür geschafft, dass es eine Güte von Überlappung gibt und Messergebnisse stärker miteinander übereinstimmen können als mit anderen, obwohl alle drei Messergebnisse sich auf grafische Weise in der Darstellung mit Unsicherheitsintervallen schneiden. Nicht zu vergessen ist an dieser Stelle, dass die Schüler:innen nicht mit der Standardabweichung, sondern mit der größten Differenz zum Mittelwert agieren. Im Zusammenhang mit dem Boxplot und der daraus resultierenden Darstellung ist es durchaus möglich, den Aspekt Grad des Vertrauens aus dem Sachstrukturmodell zu adressieren. Die Schüler:innen sehen nun, wie 50 % der Werte bzw. 100 % der Werte vom Mittelwert aus verteilt sind. Somit kann auf simple Weise bereits von einem 50 % Vertrauensgrad bzw. 100 % Vertrauensgrad gesprochen werden.

Zusammenfassend lässt sich somit festhalten, dass die Boxplots dazu genutzt werden, auf grafischem und rechnerischem Wege den Vergleich von Messergebnissen differenzierter zu betrachten und den Schüler:innen ein grobes Konzept von Grad des Vertrauens zu vermitteln, indem die Verteilung der Messwerte durch die Quartile und der daraus resultierende 50 % Vertrauensgrad hervorgehoben werden.

Durch die Darstellung des Medians und der Quartile im Rahmen des Boxplots erhalten die Lernenden einen Einblick in die Verteilung und Streuung des Datensatzes bzw. der Messreihe. Somit können Schüler:innen auch verschiedene Messergebnisse bezüglich dieser Parameter vergleichen. Auf dieser Ebene wird der Boxplot bereits als eine Art Wahrscheinlichkeitsdichtefunktion interpretiert. Daher ist die Dimension Aufstellen einer Wahrscheinlichkeitsdichtefunktion aus dem Sachstrukturmodell ebenfalls Teil des Konzepts D2.

Didaktische Reduktion

Im Rahmen der Gestaltung der App wurden einige didaktische Entscheidungen getroffen, um den Fokus auf die Güte der Überlappung zu richten. Zunächst weisen die in der App dargestellten Messreihen keinerlei Ausreißer auf. Dies deckt sich mit der Entscheidung aus den Apps C1-C3. Somit kann während der App und den Aufgaben stets ein Boxplot erstellt werden, der den gesamten Datensatz repräsentiert.

Bei der Erstellung wurde zudem darauf geachtet, keinen Spezialfall bezüglich der Quartile zu konstruieren. Die resultierenden Boxplots weisen keine Symmetrie auf. Zudem wird in den Aufgaben ein Spezialfall betrachtet, in dem der obere Whisker und das Maximum zusammenfallen. Dadurch können die Schüler:innen während der Bearbeitung der Aufgabe bereits für diesen besonderen Fall bei der prozentualen Übereinstimmung zwischen den Datensätzen sensibilisiert werden.

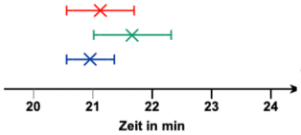
In der App wird davon abgesehen, den Schüler:innen Faustregeln für die Interpretation der verschiedenen Datensätze zur Verfügung zu stellen. Im Laufe der Bearbeitung sollen die Schüler:innen verstehen, dass es zu verschiedenen Graden an Überlappungen kommen kann und man somit quantitativ bewerten kann, dass zwei Datensätze stärker miteinander überlappen. Zu einem späteren Zeitpunkt kann dieser Gedanke im Rahmen des Spiralprinzipis aufgegriffen und zum Signifikanzbegriff erweitert werden. In diesem Zuge würden die Schüler:innen zudem lernen, dass neben der prozentualen Überlappung auch die explizite Verteilung und Streuung der Datensätze von Bedeutung sind. Zusätzlich können die Schüler:innen nun auch mehrere Datensätze miteinander vergleichen und diese auch bezüglich der Verträglichkeit quantitativ bewerten. Während nach den Apps A1 und D1 nur zwischen der Verträglichkeit und Unterscheidbarkeit differenziert werden kann, kann durch die App D2 der Grad der Verträglichkeit bewertet werden.

Von einer Darstellung des Normalized Error Ratios wird abgesehen, da dieses eine abstrakte Größe für die Schüler:innen darstellt, die Schüler:innen im Rahmen der Apps A1 und A2 den Maximalfehler als Unsicherheitsabschätzung gelernt haben und eine Expertenbefragung von Hellwig (2012, S. 150) gezeigt hat, dass dieser Wert in der Mittelstufe nicht benötigt wird. Daher wurde bei der Gestaltung der App vom E_N-Wert abgesehen und die Güte der Überlappung stattdessen auf einer prozentualen Ebene durchgeführt. Zudem wird darauf geachtet, dass das Konzept der App D2 mit den Konzepten A1 und D1 kompatibel ist.

Konzeptaufbau nach Oser

Handlungskettenschritt	Phase der App	Bezug zum Handlungskettenschritt
Bewusstmachung des Vorwissens	Start bzw. Einleitung	Eine Aktivierung der Lernenden findet durch ein Streitgespräch statt. Die Lernenden vergleichen drei Messergebnisse und streiten über die verschieden starke Überschneidung der Messergebnisse. Dabei wird wiederholt, dass Messergebnisse nicht voneinander unterscheidbar sind, wenn sie sich im Rahmen ihrer Unsicherheiten überschneiden.
Durcharbeiten eines Prototyps	Boxplot	Der Boxplot wird gemeinsam mit den Lernenden erstellt. Nach Darstellung der Messwerte werden der Median, die Quantile sowie das Minimum und Maximum eingeführt, gegebenenfalls erklärt und in einer Abbildung dargestellt. Auf diese Weise wird der Boxplot nach und nach generiert. Abschließend wird die Bedeutung der Quantile wiederholt.
Beschreibung der wichtigen Merkmale des neuen Konzepts	Erklärung	Den Schüler:innen wird grafisch und textbasiert vermittelt, wie ein Vergleich zwischen zwei Datenreihen auf Grundlage eines Boxplots durchgeführt werden kann. Dabei wird den Schüler:innen verdeutlicht, dass jedes Quartil mindestens 25 % der Messwerte beinhaltet.
Aktiver Umgang mit dem neuen Konzept	Aufgaben	Nach dem Kennenlernen eines Vergleichs, der auf der Betrachtung der Boxplots basiert, bearbeiten die Schüler:innen nun selbstständig Aufgaben und erstellen Boxplots und vergleichen Messreihen miteinander.
Anwenden des neuen Konzepts in anderen Kontexten	Aufgaben	Abschließend bearbeiten die Lernenden Aufgaben mit neuen Kontexten und bewerten die entsprechenden Messreihen hinsichtlich des Grads der Verträglichkeit.

Ausarbeitung der einzelnen Phasen auf konzeptueller und kontextueller Ebene

Phase	Konzeptionelle Ebene	Kontextuelle Ebene
Allgemein		Dauer des Schulwegs
Einleitung	Wiederholung der Voraussetzungen für diese Lernumgebung (Konzepte A1 und D1) Ziel: Das vorherige Konzept D1 zur Verträglichkeit von Messergebnissen wird erweitert und mit der Darstellung durch Boxplots um einen quantitativen Aspekt ergänzt.	Streitgespräch zwischen drei Schüler:innen bezüglich der Verträglichkeit der dargestellten Messwerte  Den Schüler:innen wird verdeutlicht, dass sie eine neue Methode zum Vergleichen von Ergebnissen lernen.
Boxplot	Sukszessive Einführung des Boxplots Nachdem die Lernenden die Messwerte eingeblendet haben, werden der Median, die Quantile und Minimum sowie Maximum in der App dargestellt. Nachdem der Boxplot vervollständigt wurde, wird die Bedeutung der Quantile den Schüler:innen nähergebracht. Jedes Quantil enthält mindestens 25 % der Messwerte. Jeder Zwischenschritt bei der Erstellung des Boxplots ist mit einer Hilfefunktion abgesichert, falls die Begrifflichkeiten den Schüler:innen nicht bekannt sein sollten.	Den Schüler:innen werden Messdaten in einem Koordiantensystem dargestellt. Die Bestandteile des Boxplots werden anhand des Datensatzes erarbeitet und visualisiert. Nach Bedarf der Schüler:innen können Hilfestellungen und Berechnungen eingeblendet werden. Zunächst wird der Median bestimmt und eingezeichnet. Anschließend werden die Quantile eingeführt und in das Diagramm eingetragen. Zum Abschluss werden den Schüler:innen das Maximum und Minimum eingeführt und der Boxplot vervollständigt. Abschließend wird den Schüler:innen gezeigt, wie die Messwerte innerhalb des Boxplots angeordnet sind und dass jedes Quantil mindestens 25 % der Messwerte beinhaltet.

. . . Tabelle geht auf der nächsten Seite weiter

Phase	Konzeptionelle Ebene	Kontextuelle Ebene
Erklärung	Auf Grundlage der Quantile wird den Schüler:innen vermittelt, wie die Verträglichkeit von Datensätzen zusätzlich untersucht werden kann. Anhand von drei Abbildungen wird visualisiert, wie die prozentuale Übereinstimmung von Messdatensätzen über Boxplots bestimmt werden kann.	Den Schüler:innen wird die Eigenschaft von Boxplots genannt: Eine übersichtliche Darstellung von statistischen Kenngrößen wie Median, Quartile usw. Den Schüler:innen wird erläutert, dass in jedem Quartil 25 % der Messwerte enthalten sind und dass man dadurch schnell Messdaten miteinander vergleichen kann. Die Abbildung demonstriert, wie prozentuelle Übereinstimmungen von 25 %, über 25 % und 50 % abgelesen bzw. bestimmt werden können.
Aufgaben		
Zusammenfassung	Anhand von Boxplots kann die prozentuale Überlappung zwischen Datensätzen beurteilt werden. Damit können verschiedenr Grade an metrologischer Verträglichkeit evaluiert werden. Dabei bedeutet eine geringere prozentuale Übereinstimmung nicht, dass die Daten nicht miteinander verträglich sind. Eine höhere prozentuale Übereinstimmung suggeriert eine stärkere Verträglichkeit der Daten.	Den Schüler:innen wird die Diskussion aus der Einleitung gezeigt. Es wird erkannt, dass alle drei Datenreihen miteinander metrologisch verträglich sind, aber Unterschiede in der prozentualen Übereinstimmung existieren. Abschließend wird quantitativ festgehalten, dass der grüne und blaue Datensatz sich zu mindestens 50 % überschneiden und dass das für den roten und grünen Datensatz nicht gilt.

Anhang B

Beschreibung der verwendeten Items zur Evaluation der Wirksamkeit der Apps

B.1 Itemselektion

App A1

Die App A1 adressiert die folgenden Kompetenzen von Schulz (2022, S. 41 ff.):

> Die Schüler:innen ...
>
> > ... finden in beschriebenen Experimentalsituationen verschiedene Ursachen von Messunsicherheiten [...].
> >
> > ... können die Definition von Messunsicherheit und Messabweichung angeben und anwenden.
> >
> > ... geben das Messergebnis als Zusammensetzung aus Gesamtunsicherheit und Ergebniswert an.

Da die App A1 nicht evaluiert werden soll, wird jeweils nur ein Item in den Itemkatalog des entsprechenden Tests integriert. Dieses Item soll eine Art Kontrolle darstellen, ob die Schüler:innen die in der App adressierten Kompetenzen und Konzepte gelernt haben.

App A2

Die App A2, welche im Rahmen dieser Studie hinsichtlich ihrer Wirksamkeit untersucht wird, adressiert die folgende Kompetenz von Schulz (2022, S. 43):

> Die Schüler:innen können die Definition von Messunsicherheit und Messabweichung angeben und anwenden.

Bezüglich der von Schulz (2022) formulierten Kompetenzen kommt es somit zu einer Überschneidung bei den Apps A1 und A2. Vor dem Hintergrund der

entwickelten didaktischen Struktur (vgl. Abschnitt 5.2) handelt sich dabei jedoch um unterschiedliche Konzepte (A1 & A2). Im Rahmen der Kompetenzen von Schulz (2022) lässt sich die explizite Thematisierung der Unsicherheitsbestimmung nach Typ A und B (JCGM, 2008) als Schwerpunkt der App A2 feststellen.

Vor diesem Hintergrund haben Lauer und Lenz (2021) neue Items entwickelt, welche sich diesem Konzept widmen. Die Entwicklung der Items orientierte sich streng an den übrigen Items des Testinstruments von Schulz (2022). Einerseits wurden die zuvor dargestellten Frageformate beibehalten und andererseits die gleichen Aspekte wie bei der Entwicklung der Items von Schulz berücksichtigt (Lauer und Lenz, 2021, S. 55 ff. Schulz, 2022, S. 33 ff.).

Indirektes Messen

Die App B1 vermittelt den Schüler:innen, wie Messwerte im Rahmen von Unsicherheiten miteinander addiert bzw. subtrahiert werden. Diese Konzept findet sich in der Kompetenzformulierung von Schulz (2022, S. 46, 56) wieder:

> Die Lernenden ...
>
> > ... geben das Messergebnis als Zusammensetzung aus Gesamtunsicherheit und Ergebniswert an.
> >
> > ... bestimmen aus mehreren unsicherheitsbehafteten Größen die Gestamunsicherheit in Summen, Differenzen, Produkten und Quotienten.

Die genannten Kompetenzen werden von den Items 801, 804, 805, 806, 808, 811 und 815 adressiert. Bei den Items 801, 804, 805, 806 und 815 handelt es sich um Items im Multiple-Choice-single-select-Format. Die übrigen Items bestehen aus mehreren Teilfragen und sind daher tendenziell als schwerer zu erachten.

Grafische Auswertung

Die Wirksamkeit der Apps für den Kompetenzbereich Grafische Auswertung wird stellvertetend durch die App C1 evaluiert. Diese App adressiert das Konzept, dass die Unsicherheitsbalken in x- und y-Richtung zu einer Unsicherheitsfläche erweitert werden und dass eine Ausgleichsgerade sämtliche Unsicherheitsflächen der dargestellten Messwerte im Koordinatensystem schneiden muss. Eine Adressierung von Ausreißern findet an dieser Stelle nicht statt. Schulz (2022, S. 48ff.) formuliert zu den zuvor beschriebenen Konzepten die Kompetenzen:

> Die Lernenden...
>
> > ... stellen eine Unsicherheitsbilanz auf und analysieren sowie bewerten diese hinsichtlich ihrer Vollständigkeit und der Stärke der Einflussfaktoren.
> >
> > ... überprüfen Kriterien für eine lineare Regression und kennen Eigenschaften einer Ausgleichsgeraden.

> ... entscheiden in gegebenen Beispielen, ob eine lineare Regression korrekt durchgeführt wurde.

In dem von Schulz (2022) zur Verfügung gestellten Testinstrument adressieren die Items 402, 404, 405, 407, 409, 412 und 414 diese Kompetenzen. Zusätzlich wurde ein weiteres Item ergänzt, welches explizit das Verständnis einer Unsicherheitsfläche adressiert (Lauer & Lenz, 2021, S. 47). Dabei weisen die Items 402, 404, 407 und 409 das Multiple-Choice-Single-Select-Format, wohingegen die übrigen Items jeweils aus mehreren Teilfragen zusammengesetzt sind.

Signifikanz

Der vierte Kompetenzbereich der didaktischen Strukturierung aus Abschnitt 5.2 adressiert den Vergleich von Messergebnissen. Die App D1 geht dabei auf die Überlappung der Unsicherheitsintervalle als Kriterium für die metrologische Verträglichkeit ein. Schulz (2022, S. 50) adressiert dieses Konzept mit den folgenden Kompetenzen:

> Die Lernenden ...
>
> > ... prüfen Messwerte auf ihre Verträglichkeit mit einem Referenzwert, indem sie den Referenzwert in Relation zum Messergebnis und dessen Messunsicherheit setzen.
> >
> > ... prüfen zwei oder mehrere Messwerte auf ihre Verträglichkeit, indem sie die Überlappung der Messunsicherheiten der Messergebnisse untersuchen.

Für den Einsatz im Testinstrument eignen sich somit die Items 301, 303, 304, 305, 307, 309, 314 und 315. Die Items 301, 304, 314 und 315 weisen das Multiple-Choice-Single-Select-Format auf. Entsprechend müssen bei den Items 303, 305, 307 und 309 Teilaussagen mit wahr oder falsch bewertet werden.

B.2 Reliabilität der Testinstrumente

Tab. B.1: Übersicht über die ermittelten Reliabilitäten im Rahmen der Pilotierung.

App	Cronbachs α	
	Prä	Post
	[95 % CI]	[95 % CI]
A2	-0,28	0,32
	[-1,03;0,42]	[-0,48;1,12]
B1	-0,20	0,56
	[-1,44;1,04]	[0,15;0,97]
C1	0,09	0,64
	[-0,7;0,87]	[0,34;0,93]
D1	-0,09	-0,07
	[-0,36;0,54]	[-0,35;0,21]

Anhang C

Fragebogen zur Evaluation der Wirksamkeit der Lernumgebung

C.1 Direktes Messen

***Kraftmesser**

Nora zieht beim Experimentieren im Physikunterricht einen Wagen über den Tisch. Sie misst dabei die Kraft, die sie aufbringen muss. Sie erhält folgende Messwerte:

Messung	Kraft in Newton
1	6,7
2	6,0
3	7,2
4	7,1
5	5,9
6	7,2

Ihr bester Freund Paul hat nur eine Messung durchgeführt und erhält 6,7 Newton.

Entscheide, wie Nora vorgehen kann, um das Messergebnis zu bestimmen.

Bitte wähle eine der folgenden Antworten:

- ○ Bei der Messung ist die Messunsicherheit **0 Newton**, da der erste Messwert mit dem Wert von Paul übereinstimmt. Die weiteren Werte sollte Nora vernachlässigen.
- ○ Nora sollte den Wert 7,2, der **zweimal** vorkommt, als Ergebniswert verwenden. Als Messunsicherheit sollte sie die **Differenz** zu Pauls Wert angeben
- ○ Sie sollte die drei Werte nehmen, die Pauls Wert **am nächsten liegen**. Aus diesen Werten sollte sie den **Mittelwert** bilden. Die Messunsicherheit ist dann der Abstand des kleinsten Werts zum Mittelwert.
- ○ Sie sollte den Mittelwert aus **allen sechs** Werten bilden und den **Maximalfehler** als Messunsicherheit verwenden.

***Einfache Aufnahme von Messwerten**

Welche Aussagen treffen zu, wenn eine Messung nur **einmal** durchgeführt werden kann?

	trifft zu	trifft nicht zu
Ich kann keine Messunsicherheit berechnen, also gebe ich das Messergebnis ohne Unsicherheit an.	○	○
Ich recherchiere nach der Unsicherheit des Messgeräts und gebe diesen Wert im Messergebnis an.	○	○
Ich verwerfe den Messwert, weil der Versuch nicht aussagekräftig ist.	○	○
Ich rate die Messunsicherheit.	○	○

*Mehrfache Aufnahme von Messwerten

Welche Aussagen treffen zu, wenn bei der bei der **fünffachen** Durchführung eines Versuchs verschiedene Messwerte aufgenommen werden?

	trifft zu	trifft nicht zu
Ich nutze die Wiederholungsmethode um das Messergebnis zu bestimmen.	○	○
Ich nehme den dritten aufgenommen Wert, weil er in der Mitte liegt. Anschließend schätze ich die Unsicherheit ab.	○	○
Ich suche im Handbuch die Unsicherheit des Messgeräts.	○	○
Ich bilde den Mittelwert der Messwerte und berechne die Unsicherheit.	○	○

*Methodenwahl

Entscheide, ob die Unsicherheit mit der Nachschlage- oder Wiederholmethode bestimmt werden sollte.

	trifft zu	trifft nicht zu
Paul verwendet die Bedienungsanleitung zur Bestimmt der Unsicherheit.	○	○
Petra misst mit einer Uhr eine Woche lang jeden Tag die Zeit, die sie für ihre Hausaufgaben benötigt und berechnet daraus den Mittelwert.	○	○
Im Fußballtraining wird mit einem Radargerät die Schussgeschwindigkeit gemessen. Jeder Schütze erhält einen Versuch.	○	○
Beim Start einer Silvesterrakete wird mit einer Lichtschranke die Zeit für die ersten 5 Meter gemessen.	○	○

*Chemieversuch

Anton hat in der Schule gelernt, dass Messungen immer wiederholt durchgeführt werden sollen. Deshalb wiegt er das Magnesium für seinen Chmieversuch dreimal ab. Er ist überrascht, dass die Waage immer den gleichen Wert anzeigt.

Entscheide, ob die Aussagen zutreffen oder nicht.

	trifft zu	trifft nicht zu
Die Unsicherheit der Messung beträgt 0.	○	○
Um die Messunsicherheit der Messung zu bestimmen, sollte Anton die die Herstellerangabe der Waage recherchieren.	○	○
Anton soll die Messung solange wiederholen, bis ein anderer Wert angezeigt wird.	○	○
Anton sollte für jede Messung eine andere Waage benutzen.	○	○

*Strom- und Spannung

Im Physikunterricht untersucht Kim einen aufgebauten Stromkreis. Für die Spannung misst sie U = 4,5 V und die Unsicherheit beträgt 4%. Bei der Stromstärke erhält sie I = 1,8 A und eine Unsicherheit von 0,2 Ampere.

Entscheide, welche der angegebenen Unsicherheiten richtig sind.

ℹ Bitte wähle eine der folgenden Antworten:

○ ΔU = ± 0,18 V und ΔI = ± 0,36 A

○ ΔU = ± 4 V und ΔI = ± 0,20 A

○ ΔU = ± 0,18 V und ΔI = ± 0,20 A

○ ΔU = ± 0,25 V und ΔI = ± 0,36 A

*Außentemperatur

Mark hat in der Schule gelernt, dass jeder Messwert eine Unsicherheit besitzt. Zuhause liest er die Außentemperatur an einem Thermometer ab. Das Thermometer zeigt eine Temperatur von 16,7 °C.

Entscheide, welche der Aussagen zutreffen.

	trifft zu	trifft nicht zu
Mark liest innerhalb kürzester Zeit mehrmals die Temperatur ab und bestimmt über den Mittelwert die Unsicherheit.	○	○
Mark nimmt zwei weitere identische Thermometer und misst die Temperatur jeweils einmal und bestimmt den Mittelwert und dann die Unsicherheit.	○	○
Mark kann keine Unsicherheit bestimmen, weil das Thermometer sie nicht anzeigt.	○	○
Mark holt sich das Handbuch des Thermometers und recherchiert die Unsicherheit.	○	○

*Durchschnittsgeschwindigkeit

Yen möchte die durchschnitte Geschwindigkeit v einer U-Bahn zwischen zwei Stationen bestimmen. Ihre Rechnungt sieht wie folgt aus:

Formel für die Geschwindigkeit v=s/t v Geschwindigkeit, s Strecke, t Zeit

Abstand zwischen den Stationen: (mit Google Maps bestimmt) s = 833m ± 5m

gemessene Fahrtzeiten

Messung	1	2	3	4	5	6	7
Zeit in s	58	63	64	61	63	59	60

Mittelwert für die Zeit: t = 61s
Unsicherheit für die Zeit: Δt = 2s

Berechnung der Geschwindigkeit: v = s/t = 833m/61s = 14m/s
Unsicherheit der Geschwindigkeit Δv= 1m/s

Entscheide, welche Aussagen zutreffen.

	trifft zu	trifft nicht zu
Die Unsicherheit des Abstands zwischen den Stationen wurde mit der Wiederholungsmethode bestimmt.	○	○
Die Unsicherheit der Zeit wurde mit der Wiederholungsmethode bestimmt.	○	○
Die Unsicherheit der Zeit ist richtig angegeben.	○	○
Für die Unsicherheit der Strecke hat Yen die Genauigkeit von Google Maps recherchiert.	○	○

C.2 Indirektes Messen

*Addition von Messergebnissen

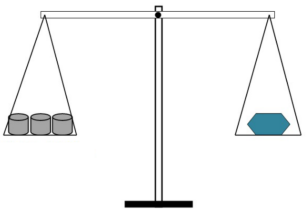
Wähle die richtige Regel aus, wenn zwei Messgrößen mit Unsicherheiten addiert oder subtrahiert werden.

Bitte wähle eine der folgenden Antworten:

- ○ Bei einer **Addition** werden die Messunsicherheiten der aufgenommenen Messwerte **addiert**, bei der **Subtraktion** werden diese **subtrahiert**.
- ○ Bei einer **Addition** werden die Messunsicherheiten der aufgenommenen Messwerte **subtrahiert**, bei der **Subtraktion** werden diese **addiert**.
- ○ Sowohl bei einer **Addition** als auch bei einer **Subtraktion** werden die Messunsicherheiten der aufgenommenen Messwerte **addiert**.
- ○ Sowohl bei einer **Addition** als auch bei einer **Subtraktion** werden die Messunsicherheiten der aufgenommenen Messwerte **subtrahiert**.

*Balkenwaage

Lotti bestimmt die Masse eines Steins mit Hilfe einer Balkenwaage. Um die Waage ins Gleichgewicht zu bringen, benötigt sie drei Massestücke. Jedes Massestück hat eine Masse von (50 ± 0,3) g.



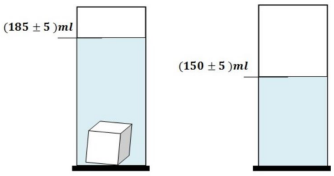
Bestimme die korrekte Aussage, wenn Lotti die Masse des Steins bestimmen möchte.

Bitte wähle eine der folgenden Antworten:

- ○ Die Messunsicherheit **verschwindet** bei der Addition der einzelnen Massen.
  Masse des Steins: 150g
- ○ Die Messunsicherheit wird bei der Addition von drei Massen nur noch **ein Drittel so groß**.
  Masse des Steins: (150 ± 0,1) g
- ○ Die Messunsicherheit bleibt **unverändert**, als würde man nur eine Masse angeben.
  Masse des Steins: (150 ± 0,3) g
- ○ Die Messunsicherheit ist **dreimal so groß** wie die Messunsicherheit eines Massestücks.
  Masse des Steins: (150 ± 0,9) g

*Wasserstand

Nicole bestimmt das Volumen eines Würfels, indem sie den Wasserstand ohne und mit dem Würfel notiert und die Differenz der beiden Werte bildet (siehe Abb.).



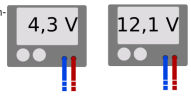
Wähle des korrekte Messergebnis aus, wenn Nicole die Differenz der beiden Wasserstände bildet.

Bitte wähle eine der folgenden Antworten:

- 35 ml ± 0 ml
- 35 ml ± 5 ml
- 35 ml ± 10 ml
- 35 ml ± 35 ml

*Stromkreis

In einem unverzweigten Stromkreis bestimmt Lenny die Gesamtspannung als Summe von zwei Teilspannungen U_1 und U_2. Für die Teilspannungen erhält er folgende Messwerte:



U_1 = (4,3 ± 0,3) V

U_2 = (12,1 ± 0,5) V

Wähle des korrekte Messergebnis aus, wenn Lenny $\mathbf{U_1 + U_2}$ bestmmt.

Bitte wähle eine der folgenden Antworten:

- (16,4 ± 0,2) V
- (16,4 ± 0,3) V
- (16,4 ± 0,5) V
- (16,4 ± 0,8) V

*Fortpflanzung der Unsicherheit

Es wurden mehrere physikalische Größen gemessen, die alle jeweils eine Unsicherheit haben. Mit diesen Größen wird nun in einer Formel weitergerechnet.

Wähle die korrekten Aussagen aus.

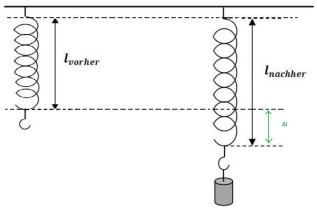
	trifft zu	trifft nicht zu
Die Messunsicherheiten der einzelnen physikalischen Größen haben **keinen Einfluss** auf die Gesamtunsicherheit.	○	○
Die Messunsicherheiten der einzelnen Größen müssen bei der Berechnung der Gesamtunsicherheit **berücksichtigt werden**.	○	○
Je **kleiner** die Unsicherheiten der einzelnen Größen, desto **größer** ist die Gesamtunsicherheit.	○	○
Je **größer** die absoluten Unsicherheiten der einzelnen Größen, desto **größer** die Gesamtunsicherheit.	○	○

*Längenänderung einer Feder

Um die Längenänderung einer Feder

$$\Delta l = l_{nachher} - l_{vorher}$$

durch das Anhängen von Massestücken zu bestimmen, hat sich Marleen die Länge l der Feder vor und während der Belastung mit einem Massestück notiert. Beide Längen haben dieselbe Messunsicherheit.

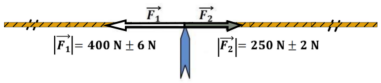


Entscheide, welche der folgenden Aussagen richtig sind.

	trifft zu	trifft nicht zu
Je **größer** die Unsicherheit von $l_{nachher}$, desto **kleiner** die Gesamtunsicherheit von Δl.	○	○
Je **größer** die Unsicherheit von $l_{nachher}$, desto **größer** die Gesamtunsicherheit von Δl.	○	○
Je **größer** die Unsicherheit von l_{vorher}, desto **kleiner** die Gesamtunsicherheit von Δl.	○	○
Je **größer** die Unsicherheit von l_{vorher}, desto **größer** die Gesamtunsicherheit von Δl.	○	○

*Tauziehen

Beim Tauziehen wirken zwei Kräfte in entgegengesetzte Richtungen.



Wähle die korrekte Rechnung aus.

	Berechnung des resultierenden Kraftbetrags	Berechnung der Gesamtunsicherheit	Resultierender Kraftbetrag und Gesamtunsicherheit.
Rechnung 1)	400 N – 250 N = 150 N	6 N – 2 N = 4 N	F = (150 ± 4) N
Rechnung 2)	400 N – 250 N = 150 N	6 N / 2 N = 3 N	F = (150 ± 3) N
Rechnung 3)	400 N – 250 N = 150 N	6 N + 2 N = 8 N	F = (150 ± 8) N
Rechnung 4)	400 N – 250 N = 150 N	6 N · 2 N = 12 N	F = (150 ± 12) N

Bitte wähle eine der folgenden Antworten:

- Rechnung 1
- Rechnung 2
- Rechnung 3
- Rechnung 4

C.3 Grafische Auswertung

***Kraftmesser**

Nora zieht beim Experimentieren im Physikunterricht einen Wagen über den Tisch. Sie misst dabei die Kraft, die sie aufbringen muss. Sie erhält folgende Messwerte:

Messung	Kraft in Newton
1	6,7
2	6,0
3	7,2
4	7,1
5	5,9
6	7,2

Ihr bester Freund Paul hat nur eine Messung durchgeführt und erhält 6,7 Newton.

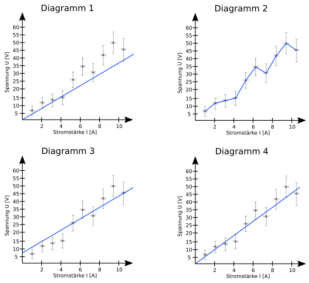
Entscheide, wie Nora vorgehen kann, um das Messergebnis zu bestimmen.

Bitte wähle eine der folgenden Antworten:

- Bei der Messung ist die Messunsicherheit **0 Newton**, da der erste Messwert mit dem Wert von Paul übereinstimmt. Die weiteren Werte sollte Nora vernachlässigen.
- Nora sollte den Wert 7,2, der **zweimal** vorkommt, als Ergebniswert verwenden. Als Messunsicherheit sollte sie die **Differenz** zu Pauls Wert angeben
- Sie sollte die drei Werte nehmen, die Pauls Wert **am nächsten liegen**. Aus diesen Werten sollte sie den **Mittelwert** bilden. Die Messunsicherheit ist dann der Abstand des kleinsten Werts zum Mittelwert.
- Sie sollte den Mittelwert aus **allen sechs** Werten bilden und den **Maximalfehler** als Messunsicherheit verwenden.

***Stromstärke-Spannungs-Diagramm**

Die Stromstärke *I* und Spannung *U* wurden im Unterricht während eines Experiments gemessen. Die Messwerte werden mit den dazugehörigen Unsicherheiten in einem Diagramm dargestellt.



Bestimme, welches Diagramm eine **Ausgleichsgerade** zeigt.

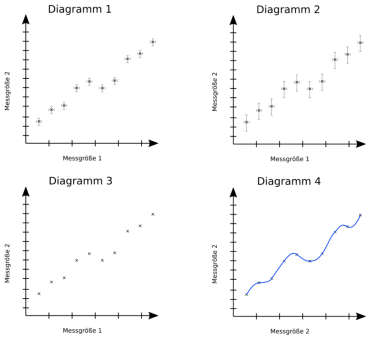
Bitte wähle eine der folgenden Antworten:

- Diagramm 1
- Diagramm 2
- Diagramm 3
- Diagramm 4

*Zwei Messgrößen

Zwischen zwei Messgrößen wird theoretisch ein linearer Zusammenhang erwartet.

Untersuche, in welches der folgenden Diagramme eine **Ausgleichsgerade** eingezeichnet werden kann, **ohne** einen Messwert auszuschließen?



Bitte wähle eine der folgenden Antworten:

- Diagramm 1
- Diagramm 2
- Diagramm 3
- Diagramm 4

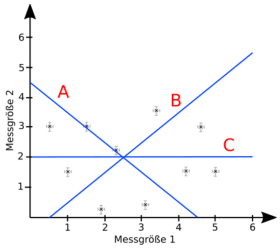
*Ausgleichsgerade

Entscheide, welche der folgenden Aussagen bezüglich einer **Ausgleichsgeraden** richtig sind?

	trifft zu	**trifft nicht zu**
Die Ausgleichsgerade erhält man, in dem man alle Messpunkte miteinander verbindet.		
Oberhalb und unterhalb der Ausgleichsgeraden müssen immer gleich viele Messwerte liegen.		
Damit eine Ausgleichsgerade sinnvoll eingezeichnet werden kann, müssen ausreichend viele Messwerte vorhanden sein.		
Eine Ausgleichsgerade kann nur dann sinnvoll gezeichnet werden, wenn ein linearer Zusammenhang vermutet wird.		

*Zusammenhang zwischen Messgrößen

Um einen möglichen Zusammenhang zwischen zwei Messgrößen zu untersuchen, werden Messwerte aus einem Experiment in einem Diagramm dargestellt.



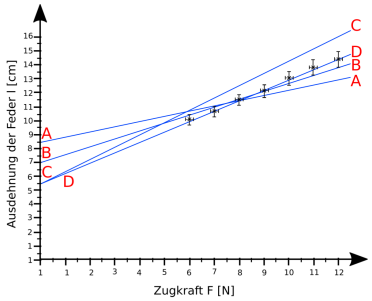
Gib an, welche der dargestellten Geraden eine **Ausgleichsgerade** ist.

Bitte wähle eine der folgenden Antworten:

- Gerade A
- Gerade B
- Gerade C
- Es besteht kein erkennbarer linearer Zusammenhang zwischen den Messgrößen.

*Zugkraft auf einer Feder

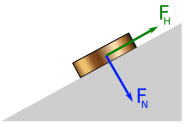
Zwischen der Länge *l* einer Feder und der Zugkraft F wird ein linearer Zusammenhang vermutet. Die Messergebnisse mit den zugehörigen Unsicherheitsintervallen werden in ein F-l-Diagramm eingetragen.



Gib an, welche der eingezeichneten Geraden eine **Ausgleichsgerade** des gemessenen Zusammenhangs zeigen.

	trifft zu	trifft nicht zu
Gerade A	○	○
Gerade B	○	○
Gerade C	○	○
Gerade D	○	○

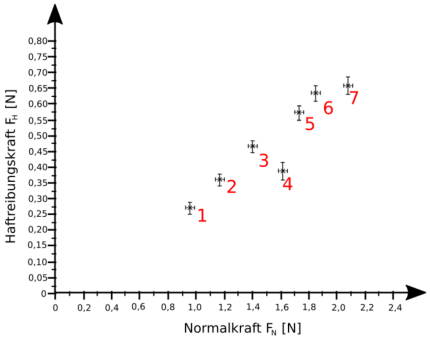
*Normalkraft und Haftreibung



Eine Schülerin vermutet einen linearen Zusammenhang zwischen der Haftreibungskraft F_{Haft} und der Normalkraft F_N an einer schiefen Ebene.

Beim Einzeichnen der Werte ins Diagramm fällt ihr auf, dass es einen Messpunkten gibt, der das Einzeichnen einer Ausgleichsgeraden nicht ermöglicht.

Gib an, um welchen Punkt es sich handelt.

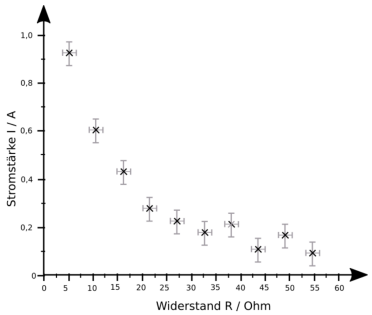


Bitte wähle eine der folgenden Antworten:

- ○ Messpunkt 2
- ○ Messpunkt 4
- ○ Messpunkt 6
- ○ Messpunkt 7

*Widerstand

In einem Stromkreis wird die Spannung *U* der Batterie als konstant angenommen. Der Widerstand *R* wird verändert und die Stromstärke *I* in dessen Abhängigkeit gemessen. Zu erwarten ist der theoretische Zusammenhang *I* = *U* / *R*. Die aufgenommenen Werte werden anschließend in ein *R*-*I*-Diagramm eingetragen.



Begründe, warum **keine Ausgleichsgerade** eingezeichnet werden kann.

	trifft zu	**trifft nicht zu**
Die gemessenen Werte liegen auch im Rahmen ihrer Unsicherheit nicht auf einer Geraden.	○	○
Es sind zu wenig Messwerte vorhanden.	○	○
Es scheint kein linearer Zusammenhang zu bestehen.	○	○
Die Messdaten stehen im Widerspruch zur theoretischen Formel I=U/R.	○	○

C.4 Signifikanz

*Zeitmessung

Ein Schüler erhält bei einer Zeitmessung ein Messergebnis von (3,5 ± 0,5) s. Der Vergleichswert für die Messung ist 3,2 s.

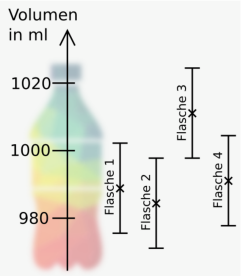
Entscheide und begründe, ob das Messergebnis und der Vergleichswert voneinander unterscheidbar sind.

Bitte wähle eine der folgenden Antworten:

- Das Messergebnis und der Vergleichswert sind **nicht** voneinander **unterscheidbar**, denn das Intervall der Unsicherheit geht von 3,2 s bis 3,8 s.
- Das Messergebnis und der Vergleichswert sind **nicht** voneinander **unterscheidbar**, denn das Intervall der Unsicherheit geht von 3,0 s bis 4,0 s.
- Das Messergebnis und der Vergleichswert sind voneinander **unterscheidbar**, denn das Intervall der Unsicherheit geht von 3,5 s bis 4 s.
- Das Messergebnis und der Vergleichswert sind voneinander **unterscheidbar**, denn das Intervall der Unsicherheit geht von 3,5 s bis 4,5 s.

*Wasserflaschen

In einer Fabrik werden Wasserflaschen mit 1000 ml Wasser befüllt. Bei der Befüllung kommt es zu Schwankungen. Liegt das geforderte Volumen von 1000 ml nicht innerhalb der Messunsicherheit, muss eine Flasche aussortiert werden. Die folgende Grafik zeigt die Messergebnisse von vier zufällig ausgewählten Flaschen.



Entscheide, welche Flaschen behalten werden.

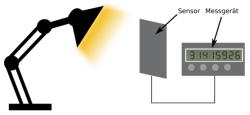
	Behalten	Aussortieren
Flasche 1	○	○
Flasche 2	○	○
Flasche 3	○	○
Flasche 4	○	○

*Schreibtischlampe

Die Helligkeit einer Schreibtischlampe wird untersucht. Die Helligkeit wird in Lux angegeben. Laut Hersteller hat Schreibtischlampe eine Helligkeit von 500 Lux betragen.

Drei Tester haben experimentell die Messwerte 492 Lux, 500 Lux und 506 Lux und jeweils eine Unsicherheit von 10 Lux ermittelt.

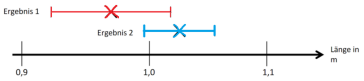
Entscheide, welche der folgenden Aussagen richtig sind.



	trifft zu	trifft nicht zu
Nur ein Messwert bestätigt innerhalb der Messunsicherheit die Herstellerangabe.	○	○
Die Angabe des Herstellers ist falsch, da bei zwei Messungen nicht 500 Lux gemessen wurde.	○	○
Innerhalb der Messunsicherheit sind alle Messwerte nicht voneinander unterscheidbar.	○	○
Da ein Messwert kleiner als 500 Lux ist, stimmt die Herstellerangabe nicht.	○	○

*Längenmessung

Die Länge eines Objekts wurde mit zwei unterschiedlichen Methoden bestimmt. In der Grafik sind die dazugehörigen Messergebnisse dargestellt.



Entscheide und begründe, ob die Messergebnisse voneinander unterscheidbar sind.

Bitte wähle eine der folgenden Antworten:

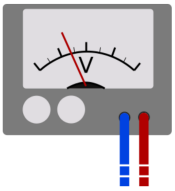
- Die Messergebnisse sind voneinander **unterscheidbar**, denn sie liegen nicht auf der gleichen Höhe.
- Die Messergebnisse sind voneinander **unterscheidbar**, da sie unterschiedlich große Unsicherheiten besitzen.
- Die Messergebnisse sind **nicht** voneinander **unterscheidbar**, denn die Intervalle der Unsicherheiten überschneiden sich.
- Die Messergebnisse sind **nicht** voneinander **unterscheidbar**, denn die Messwerte liegen jeweils außerhalb der Unsicherheitsintervalle.

*Zwei Voltmeter

Max möchte die Spannung mit zwei verschiedenen Messgeräten (Voltmeter) messen.

Er nimmt dazu mit beiden Voltmetern jeweils eine Messreihe mit der gleichen Schaltung auf. Mit dem einen Messgerät erhält er als Mittelwert 5,8 V und mit dem anderen erhält er als Mittelwert 6,3 V.

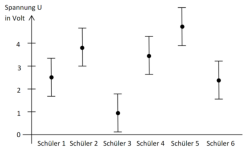
Entscheide, welche der folgenden Aussagen richtig sind.



	trifft zu	trifft nicht zu
Die beiden Messwerte sind voneinander **unterscheidbar**, weil sie nicht übereinstimmen.	○	○
Die beiden Messwerte sind **nicht** voneinander **unterscheidbar**, da sie gerundet identische Werte haben.	○	○
Es kann **nicht geprüft** werden, ob die beiden Messwerte voneinander unterscheidbar sind, da keine Unsicherheiten angegeben sind.	○	○
Da die Werte von verschiedenen Messeräten gemessen wurden, kann **nicht geprüft** werden, ob sie voneinander unterscheidbar sind.	○	○

*Spannung eines Akkus

Sechs Schüler haben nacheinander die Spannung desselben Akkus gemessen. Das Diagramm zeigt die einzelnen Messwerte mit ihren Unsicherheitsintervallen.



Entscheide, welche Messergebnisse **unterscheidbar** voneinander sind.

	trifft zu	trifft nicht zu
Schüler 1 und Schüler 2	○	○
Schüler 2 und Schüler 3	○	○
Schüler 3 und Schüler 4	○	○
Schüler 5 und Schüler 6	○	○

*Zwei Federkraftmesser

Zwei Schülerinnen bestimmen im Physikunterricht die Federkonstante einer Feder. Die Federkonstante legt fest, wie viel Kraft man aufwenden muss um eine Feder um eine 1cm auszulenken. Sie erhalten folgende Messergebnisse:

Schülerin 1: $D_1 = (0{,}06 \pm 0{,}01)$ N/cm

Schülerin 2: $D_2 = (0{,}07 \pm 0{,}02)$ N/cm

Auf der Verpackung der Feder wird ein Wert von 0,06 N/cm. angegeben Die Schülerinnen vergleichen ihre Messergebnisse mit diesem Wert.

Wähle die korrekte Antwort aus.

Bitte wähle eine der folgenden Antworten:

- Die Messergebnisse und der Wert sind **nicht unterscheidbar**, da die Werte dicht beieinander liegen.
- Die Messergebnisse und der Wert sind **nicht unterscheidbar**, da der Wert innerhalb der Unsicherheit der Messergebnisse liegt.
- Die Messergebnisse und der Wert sind **unterscheidbar**, denn das Ergebnis D_2 stimmt nicht mit dem Wert überein.
- Die Messergebnisse und der Wert sind **unterscheidbar**, denn die Schülerinnen haben unterschiedliche Messwerte erhalten.

*Außentemperatur

Im Radio wird die aktuelle Temperatur von 18 °C für Marks Wohnort angesagt. Zu Hause liest er von einem Thermometer die Außentemperatur von 16,7 °C ab. Laut Hersteller beträgt die Unsicherheit dieses Geräts 2,0 °C.

Entscheide, welche der folgenden Aussagen korrekt ist.

Bitte wähle eine der folgenden Antworten:

- Marks Messung ist von der Angabe im Radio **unterscheidbar**, denn gerundet erhält er einen Wert von 17,0 °C.
- Die beiden Werte sind **nicht unterscheidbar**, weil die genauen Orte der Messung nicht bekannt sind.
- Die beiden Werte sind **unterscheidbar**, da der gemessene Temperatur Wert von Mark zu klein ist.
- Marks Messung ist von der Angabe im Radio **nicht unterscheidbar**, denn der im Radio genannte Wert liegt innerhalb der Unsicherheit.

*Taschenlampe

Anna überprüft die Spannung der Batterie für ihre Taschenlampe. Anna misst eine Spannung von 1,5 V. In der Herstellerangabe des Messgeräts ist eine Unsicherheit von 0,2 V angegeben. Die Taschenlampe benötigt eine Spannung zwischen 1,4 V und 1,7 V. Ist die Spannung der Batterie kleiner, so geht die Lampe nicht an. Ist die Spannung höher als 1,7 V, wird die Lampe beschädigt.

Entscheide, welche der folgenden Aussagen richtig ist.

Bitte wähle eine der folgenden Antworten:

- Anna kann die Batterie **verwenden**, weil die gemessene Spannung innerhalb der Unsicherheit liegt.
- Anna kann die Batterie **nicht verwenden,** weil die Batterie die Taschenlampe beschädigt.
- Anna kann die Batterie **verwenden**, weil die Batterie sonst nicht hätte verkauft werden dürfen.
- Anna kann die Batterie **nicht verwenden**, da die Taschenlampe nicht angehen wird.

Anhang D

Fragebogen zu Studie 3

Vincent ist 16 Jahre alt und geht in die 9. Klasse. Nach dem Frühstück checkt er manchmal die Nachrichten auf seinem Handy. Er ist sehr erstaunt, als er folgende Meldung sieht (das Foto rechts ist eine Vergrößerung der Abbildung von seinem Handy). „Wow, das macht bestimmt Spaß!" Auf dem Bild sieht er einen Springer am Seil durch die Luft hin und her pendeln. „Was passiert eigentlich, wenn man sich alleine nicht traut und zu zweit pendeln würde? Dauert es dann vielleicht länger oder kürzer, bis man einmal hin und wieder zurück geschwungen ist?" fragt sich Vincent.

Mindener Tageblatt

Kletterer springen von Mindener Glacisbrücke

Von Max Stark und Daniel Mehlkopf vom 01.02.2014, 00:00 Uhr | Update: 01.02.2014, 01:19 Uhr | Lesezeit 3 Min.

Minden (dm/ms). Mit einem lauten Schrei stürzt sich der Mann von der Glacisbrücke. Nur ein Seil schützt ihn vor dem Aufprall. Nach einem kurzen Fall fängt die Sicherung ihn auf, er schwingt knapp über der Wasseroberfläche wie ein Pendel hin und her.

Ausgelassen schwingen die Kletterer unter der Brücke hin und her. - © Foto: Paul Offermann

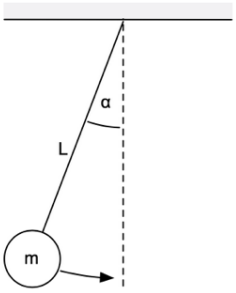
Man kann den Springer am Seil als ein Fadenpendel betrachten (siehe Abbildung unten rechts). Ein Fadenpendel besteht aus einer Masse, die an einem frei beweglichen Faden aufgehängt ist. Diese angehangene Masse entspricht der Masse Person, der Faden entspricht dem Seil. Ein Fadenpendel kann durch die folgenden Größen beschrieben werden:

T Schwingungsdauer: die Zeit, die ein Pendel benötigt, um eine komplette Schwingung durchzuführen, z. B. von der Ruhelage (Seil hängt senkrecht nach unten) zunächst ganz nach rechts, dann ganz nach links und wieder zurück zur Ruhelage. **Je größer** also die Schwingungsdauer, **desto länger** benötigt das Pendel für eine komplette Schwingung.

L Länge des Fadens, an dem die Masse aufgehängt ist

m Die angehangene Masse

α Auslenkung: der Winkel, um den das Pendel zu Beginn der Schwingung aus der Ruhelage gehoben wird



Es ist anzunehmen, dass Schwingungsdauer, Fadenlänge, Masse und Auslenkung in irgendeiner Weise miteinander zusammenhänge.

Vincent weiß, dass sich die Masse des Pendels ändert, wenn z.B. zwei Personen anstatt einer Person zusammen schwingen. Vincent möchte nun ein Experiment machen, um herauszufinden, wie sich die Schwingungsdauer ändert, wenn man die Masse ändert. Vor dem Experiment stellt er zunächst eine Vermutung auf. Dabei berücksichtigt er Effekte der Luftreibung und ganz große Winkel (größer als 90°) nicht.

*Stelle bitte auch du eine Vermutung auf, wie die Masse des Pendels mit der Schwingungsdauer zusammenhängt.

Ein Vergrößern der Pendelmasse bewirkt, dass die Schwingungsdauer, also die Zeit für eine komplette Schwingung...

Bitte wähle eine der folgenden Antworten:

- ○ größer wird.
- ○ gleich bleibt.
- ○ kleiner wird.

Begründe **kurz** deine Vermutung!

(etwa 2-4 Sätze)

	unsicher 1	2	3	sicher 4
Wie sicher bist du dir, dass deine Vermutung richtig ist?	○	○	○	○

Vincent hat die Messwerte in einer Tabelle aufgetragen. Für jede angehangene Masse hat er die Messung **sechsmal** wiederholt. Die gemessene Schwingungsdauer **T** steht jeweils in den Spalten T_1 bis T_6. Dabei steht T_1 für die erste Messung, T_2 für die zweite Messung, usw.

	T_1	T_2	T_3	T_4	T_5	T_6
m = 50 g	2,02 s	2,05 s	2,01 s	2,02 s	2,05 s	2,03 s
m = 100 g	2,05 s	2,06 s	2,07 s	2,03 s	2,02 s	2,01 s
m = 150 g	2,03 s	2,06 s	2,06 s	2,04 s	2,05 s	2,06 s
m = 200 g	2,06 s	2,07 s	2,02 s	2,07 s	2,08 s	2,06 s

Werte die Messwerte aus und überprüfe anhand der Messwerte deine gerade aufgestellte Vermutung zu diesem Experiment. Du kannst die folgende Box für Notizen verwenden.

*Behältst du deine Vermutung von Beginn bei?

✔ Ja ⊘ Nein

	unsicher 1	2	3	sicher 4
Wie sicher bist du dir, dass es richtig ist, deine Vermutung beizubehalten / zu verwerfen	○	○	○	○

	trifft gar nicht zu	trifft wenig zu	trifft teils/teils zu	trifft eher zu	trifft völlig zu
Bei der Entscheidung habe ich mich ganz auf mein Gefühl verlassen.	○	○	○	○	○
Bei meiner Entscheidung berücksichtige ich auch, dass das Experiment ungenau ist.	○	○	○	○	○
Die Auswertung meiner Messung ist der Grund für meine Entscheidung zum Beibehalten/Wechseln meiner Vermutung.	○	○	○	○	○
Ich vertraue demjenigen, der das Experiment aufgebaut hat. Das berücksichtige ich bei meiner Entscheidung.	○	○	○	○	○
Ich habe mich gefühlsmäßig für die Antwort entschieden, die mir am meisten zusagt.	○	○	○	○	○
Die Messdaten spielen bei meiner Entscheidung die größte Rolle.	○	○	○	○	○
Bei der Entscheidung berücksichtige ich stark mein Gefühl.	○	○	○	○	○
Bei meiner Entscheidung berücksichtige ich, dass dieses Experiment von den Fachleuten bestimmt richtig funktioniert.	○	○	○	○	○
Ich habe die Messungenauigkeiten beim Experimentieren in hohem Maße berücksichtigt.	○	○	○	○	○

Meine Entscheidung, die Vermutung beizubehalten/zu verwerfen ist eine Reaktion auf die Messdaten.	○	○	○	○	○
Bei meiner Entscheidung berücksichtige ich, dass das Experiment bereits von Wissenschaftlern erdacht und überprüft wurde.	○	○	○	○	○
Bei diesem Experiment gibt es Messungenauigkeiten, die ich bei meiner Schlussfolgerung berücksichtige.	○	○	○	○	○
Meine Messdaten sind der Grund, warum ich meine Vermutung verwerfe/beibehalte.	○	○	○	○	○
Bei meiner Entscheidung spielen Gefühle eine große Rolle.	○	○	○	○	○
Bei der Entscheidung zum Wechseln/Beibehalten meiner Vermutung berücksichtige ich, dass man immer Ungenauigkeiten beim Experimentieren hat.	○	○	○	○	○
Bei meiner Entscheidung berücksichtige ich, dass die Fachleute bestimmt keine Fehler im Aufbau des Experiments gemacht haben.	○	○	○	○	○
Ich höre stark auf mein Bauchgefühl, wenn ich aus dem Experiment Schlüsse ziehe.	○	○	○	○	○
Um für mich diese Entscheidung zum Beibehalten/Verwerfen meiner Vermutung zu begründen, denke ich besonders intensiv über die Messdaten nach.	○	○	○	○	○
Die Ungenauigkeiten beim Experimentieren erschweren meine Entscheidung.	○	○	○	○	○
Ich berücksichtige, dass das Experiment von Wissenschaftlern mitgebracht wurde, die an der Pädagogischen Hochschule Karlsruhe arbeiten.	○	○	○	○	○

Anhang E

Ausführungen zu den statistischen Methoden

E.1 Konvergenz der MCMC-Chains

Die Konvergenz der in Abschnitt 5.1.6 beschriebenen Verfahren zugrunde liegenden MCMC-Chains wird anhand der Traceplots (s. Abb. E.1) und des Parameters $\hat{R}$ sowie der effektiven Stichprobengröße *ESS* bewertet.

Bei den Traceplots handelt es sich um Darstellungen der MCMC-Chains. Die Elemente der Kette werden sequenziell dargestellt. Für die Auswertung werden für jeden zu schätzenden Parameter vier MCMC-Chains berechnet (Vehtari et al., 2021). Streuen die dargestellten Werte der Chains eines Parameters um dieselben Werte, so kann von einer Konvergenz des Verfahrens ausgegangen werden (McElreath, 2020). Die in Abbildung E.1 dargestellten Traceplots zeigen diesem Verständnis nach eine Konvergenz der MCMC-Chains.

Bei $\hat{R}$ handelt es sich um den sogenannten *Potential-Scale-Reduction-Factor* (McElreath, 2020). Der Faktor wird gebildet, indem die Streuung aller MCMC-Chains ins Verhältnis zur Streuung innerhalb der MCMC-Chains gesetzt wird (Vehtari et al., 2021). Ein Faktor über 1,00 kann dabei bedeuten, dass die MCMC-Chains (noch) nicht konvergiert sind. In der Literatur werden Cut-off-Werte von 1,01 (Vehtari et al., 2021) vorgeschlagen. In Tabelle E.1 zeigt sich, dass sämtliche $\hat{R}$ Parameter unter dem Cut-off-Wert liegen. Die Parameter deuten somit auf eine Konvergenz der jeweiligen MCMC-Chains. Die effektive Stichprobengröße einer MCMC-Chain kann verwendet werden, um die Effizienz einer MCMC-Chain zu beurteilen (McElreath, 2020). Dabei schlägt Kruschke (2015) vor, mindestens eine effektive Stichprobengröße von 10000 zu verwenden, wenn Parameterschätzungen vorgenommen werden sollen (Kruschke, 2015, S. 184). Dabei ist zu berücksichtigen, dass es verschiedene Algorithmen zur Berechnung der *ESS* gibt (Kruschke, 2015, S. 184) und es sich bei dem Kennwert um eine Schätzung handelt (McElreath, 2020, S. 293). Die in Tabelle E.1 dargestellten

Tab. E.1: Die Tabelle zeigt die Konvergenzparameter für die MCMC-Chains für den jeweiligen Kompetenzbereich und die dazugehörigen Mittelwerte μ und Standardabweichungen σ.

	μ		σ	
Kompetenzbereich	$\hat{R}$	ESS	$\hat{R}$	ESS
Direktes Messen	1,00	61546	1,00	45945
Indirektes Messen	1,00	60272	1,00	42231
Grafische Auswertung	1,01	60452	1,00	32972
Signifikanz	1,00	59755	1,00	40257

effektiven Stichprobengrößen lassen darauf schließen, dass die MCMC-Chains konvergiert sind.

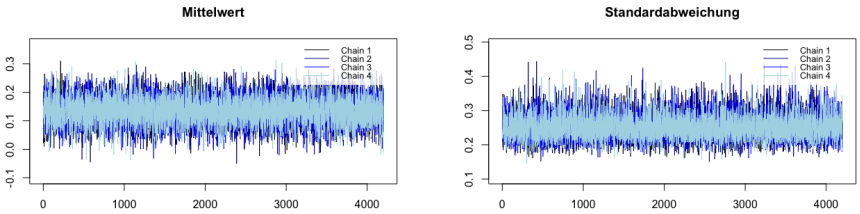
In Anbetracht der Tatsache, dass die Traceplots, der Parameter $\hat{R}$ und die effektive Stichprobengröße *ESS* für eine Konvergenz der MCMC-Chains sprechen, wird für die vorliegende Arbeit Konvergenz angenommen. Eine Auswertung der Ergebnisse ist somit valide und kann in Abschnitt 5.2 eingesehen werden.

E.2 Item Response Theorie

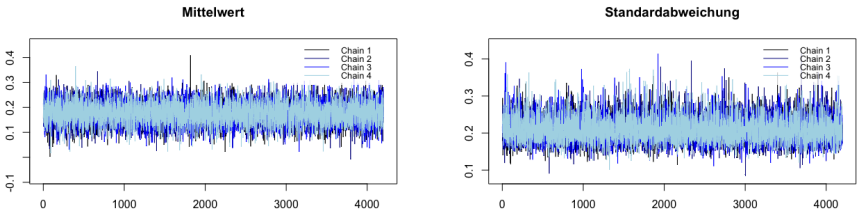
Im nachfolgenden wird das Messmodell für die Erfassung der Kompetenz im Umgang mit Daten und Messunsicherheiten ausführlich analysiert. Zunächst werden die Modelle zu dem jeweiligen Erfassungszeitpunkt untersucht, bevor anschließend das zusammengefügte Modell verwendet wird, welches die Grundlage für die Untersuchung der Forschungsfragen darstellt.

E.2.1 Analyse des Rasch-Modells für die Prä- und Posterhebung

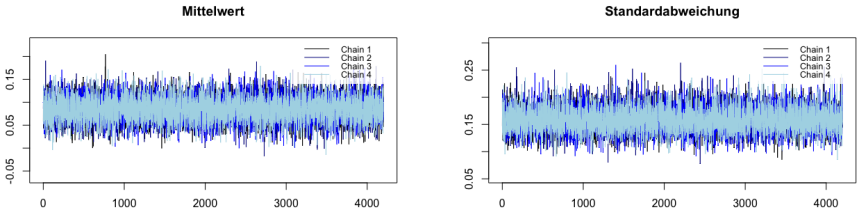
Als Testinstrument kommen die Items nach Schulz (2022) in seiner überarbeiteten Form nach Abschnitt 5.1 und C zum Einsatz. Die Evaluation des Testinstruments erfolgt auf Grundlage von drei Parametern und Darstellungen: MNSQ-Fit-Werte, Item Characteristic Curves und Wright-Maps. Auf Grundlage der Gesamtheit wird das Testinstrument evaluiert.



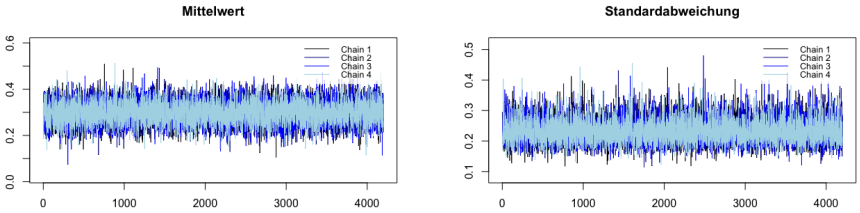
(a) Direktes Messen



(b) Indirektes Messen



(c) Grafische Auswertung



(d) Signifikanz

Abb. E.1: Traceplots der Parameterschätzungen für die Kompetenzbereiche Direktes Messen, Indirektes Messen, Grafische Auswertung und Signifikanz. Es zeigt sich eine gleichmäßige Streuung der MCMC-Chains um einen gemeinsamen Wert. Es kann von einer vorliegenden Konvergenz ausgegangen werden.

Passung des Modells

Die Rasch-Analyse wird unter Verwendung des `TAM`-Pakets (Robitzsch et al., 2021) und `WrightMap`-Pakets (Irribarra & Freund, 2014) für die Programmiersprache `R` realisiert (R Core Team, 2021). Die Rasch-Analyse entspricht einem 1-*pl*-Modell, bei welchem der Diskriminationsparameter auf 1 fixiert ist. Die Resultate für die Fit-Werte sind in Tabelle E.2 dargestellt.

Für den Test zum Präzeitpunkt zeigt sich eine Auffälligkeit bei Item S14. Nach Wu et al. (2016, S. 191) liegt ein Underfit vor und dieses Items weist eine geringere diskriminante Stärke auf als die restlichen Items des Tests zum Präzeitpunkt. Dennoch liegt dieses Items wie die restlichen Items des Präzeitpunkts in dem von Bond und Fox (2007) vorgeschlagenen Cut-off-Intervall zwischen 0,8 und 1,2. Solange die Fit-Werte des Items in dem vorgeschlagenen Intervall liegen, kann die signifikante Abweichung vernachlässigt werden (Boone et al., 2014, S. 166). Diesbezüglich motivieren und disktuieren Wright und Linacre (1994) die Bevorzugung der MSQ-Fit-Werte gegenüber der Signifikanztestung.

Im Posttest für die Erfassung der Kompetenz im Umgang mit Daten und Messunsicherheiten zeigt sich ebenfalls ein auffälliges Item. Das Item S29 weist mit einem Infit-Wert von 1,14 einen leichten Underfit auf. Zudem liegt der Outfit-Wert mit 1,27 außerhalb des Cut-off-Werts. Somit liegen alle Items bezüglich der Infit-Werte im Intervall von 0,8 bis 1,2. Bei den Outfit-Werten stellt das Item S29 einen Ausreißer dar. Für den vorliegenden Test wurde trotzdem beschlossen, das Item zu behalten, da einerseits der Infit-Wert innerhalb des Cut-off-Intervalls liegt und der Test trotz dieses Items eine akzeptable WLE-Reliabilität von 0,59 aufweist.

ICCs

Die ICCs zeigen, wie die Lösungswahrscheinlichkeit eines Items in Abhängigkeit der Schüler:innenfähigkeit verläuft. Die ICCs können dazu verwendet werden, die Passung des empirischen und theoretischen Modells zu bewerten. Dabei wird nach Wu et al. (2016) vorgegangen. Anhand der Evaluation der ICCs kann daher evaluiert werden, inwiefern eine Auswertung der Daten unter Verwendung des Rasch-Modells legitim ist.

Die ICCs für den Prä- und Posterhebungszeitraum können den Abbildungen E.2 und E.3 entnommen werden. Für den Präzeitpunkt zeigt sich, dass die empirischen und theoretischen ICC für die Items S11, S12, S13 und S17 gleich sind. Für Item S14 zeigt sich, dass das Item nicht gut zwischen leistungsstarken Schüler:innen im Bereich von 1 bis 3 Logit unterscheidet. Das Item S15 weist eine fehlende Diskrimination zwischen 0 und 1 Logit auf, kann jedoch außerhalb dieser Intervalle besser zwischen den Schüler:innen unterscheiden. Ebenso zeigt sich für das Item S16, dass es zwischen leistungsschwachen und durchschnittlichen Schüler:innen besser differenziert als die theoretische ICC. Für leistungsstarke

Tab. E.2: Übersicht der Itemschwierigkeiten und Itemparameter für den Prä- (S11-S19) und Postzeitpunkt (S21-S29).

			Outfit			Infit		
Item	ξ	SE_ξ	Wert	t	p	Wert	t	p
S11	−0,78	0,13	1,00	0,08	0,94	1,02	0,39	0,70
S12	1,18	0,13	0,96	−0,40	0,69	0,96	−0,61	0,54
S13	0,59	0,12	0,96	−0,55	0,58	0,98	−0,51	0,61
S14	0,54	0,12	1,18	2,51	0,01	1,11	2,23	0,02
S15	3,03	0,23	1,11	0,46	0,65	1,01	0,13	0,90
S16	0,02	0,12	0,91	−1,52	0,13	0,93	−1,68	0,09
S17	−0,41	0,12	0,96	−0,58	0,56	0,98	−0,51	0,61
S18	−0,04	0,12	0,95	−0,58	0,56	0,98	−0,51	0,61
S19	−0,76	0,13	1,09	1,17	0,24	1,05	1,01	0,31
S21	−1,30	0,14	0,90	−0,63	0,53	0,99	−0,17	0,86
S22	0,36	0,13	1,04	0,43	0,66	1,02	0,33	0,74
S23	0,25	0,13	1,03	0,34	0,73	1,03	0,49	0,62
S24	−1,84	0,16	0,85	−0,70	0,48	0,96	−0,44	0,66
S25	1,79	0,15	0,94	−0,20	0,84	1,04	0,53	0,60
S26	−0,25	0,13	1,02	0,26	0,79	0,98	−0,27	0,79
S27	1,67	0,15	0,86	−0,68	0,50	0,97	−0,37	0,70
S28	−0,74	0,13	0,88	−0,98	0,33	0,89	−1,82	0,07
S29	−1,10	0,14	1,27	1,82	0,07	1,14	0,33	0,74

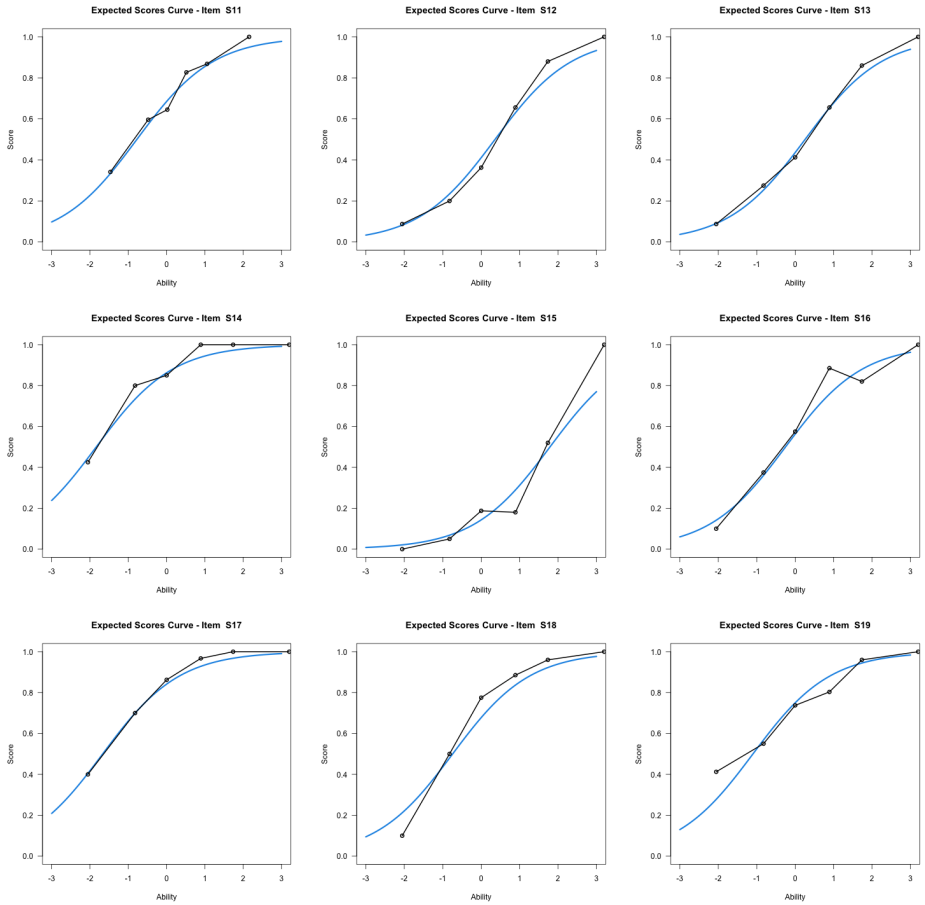


Abb. E.2: Darstellung der ICCs für die Items S11 bis S19 für den Prätest.

Schüler:innen weist das Item jedoch Schwächen auf. Die empirische ICC von S18 zeigt, dass das Item durchgängig besser zwischen den Schüler:innen differenziert, als theoretisch erwartbar ist. Beim Vergleich der empirischen und theoretischen ICCs von Item S19 ist erkennbar, dass die empirische Kurve etwas flacher verläuft. Somit löst das Item etwas schlechter als theoretisch erwartet zwischen unterschiedlichen Schüler:innenfähigkeit auf. Insgesamt lässt sich festhalten, dass der Test im Gesamten gut zwischen den leistungsschwachen und -starken Schüler:innen differenzieren kann.

Im Posttest zeigen die Items S22, S23, S25 und S28 eine starke Übereinstimmung zwischen empirischer und modellbasierter ICC. Das Item S21 kann empirisch unter leistungsstarken Schüler:innen etwas besser differenzieren. Im Bereich von leistungsschwachen bis durchschnittlichen Schüler:innen kann das Item S24 besser unterscheiden als die modellbasierte ICC. Für das restliche Spektrum der Schüler:innenfähigkeit deckt sich die modellbasierte mit der empirischen

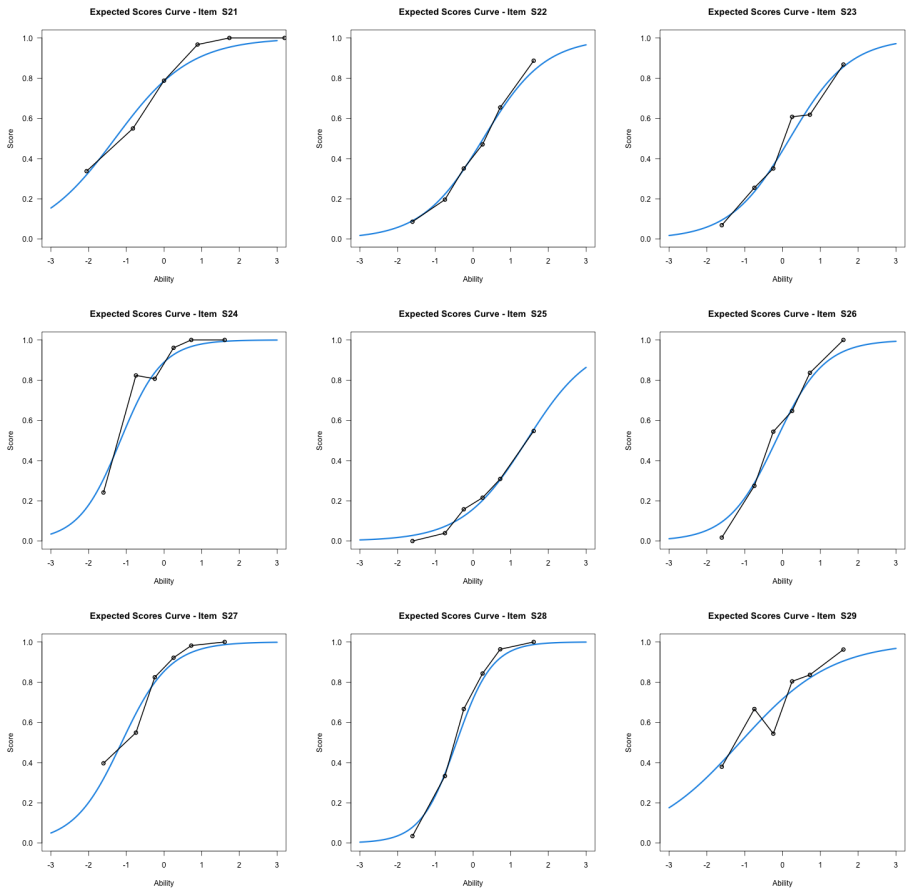


Abb. E.3: Darstellung der ICCs für die Items S21 bis S29 für den Posttest.

ICC. Besonders gut zwischen leistungsstarken und -schwachen Schüler:innen unterscheidet das Item S26, welches dann wiederum im durchschnittlichen Bereich eine Übereinstimmung zwischen empirischer und modellbasierter ICC aufweist. Bei der Funktionsweise von Item S27 zeigen sich ebenso Unterschiede zwischen der empirischen und der theoretischen ICC im Bereich von -2 bis 0 Logits. Zwischen -2 und -1 Logits unterscheidet die empirische ICC deutlich schlechter als die theoretische ICC zwischen den Schüler:innen. Im Bereich von -1 bis 0 Logits ist die theoretische ICC etwas steiler und nähert sich daraufhin der theoretischen ICC an. Abschließend bleibt das Item S29, welches bereits bei den Fit-Werten Auffäligkeiten zeigt. Es zeigt sich, dass das Item bzw. die empirische ICC in der unteren Hälfte der Schüler:innenfähigkeiten nicht gut differenzieren kann. Erst ab einer Fähigkeit von 0 Logits und aufwärts funktioniert das Item und weist anfangs sogar eine bessere Differenzierung auf als die theoretische ICCs. Dieses Verhalten des Items könnte darauf hinweisen, dass das Item deutlich schwerer als der Durchschnitt ist. Hierzu wird im folgenden Abschnitt Stellung genommen.

Zusammenfassend lässt sich sagen, dass es in einigen Bereichen Unterschiede in der diskriminanten Funktion zwischen den empirischen ICCs und den theoretischen ICCs gibt. In seiner Gesamtheit kann der Test jedoch gut zwischen den unterschiedlichen Schüler:innenfähigkeiten differenzieren. Zudem sei angemerkt, dass die zuvor dokumentierten Unterschiede in der diskriminanten Funktionsweise durchaus gering ausfallen und nur in Ausnahmefällen sehr stark vorhanden sind.

Unter Berücksichtigung der vorangehenden Abschnitte wird von einer komplexeren 2-*pl*-Modellierung des Testinstruments abgesehen. Es konnte eine hinreichend gute Passung zwischen den Daten und dem Rasch-Modell gefunden werden, sodass die in dieser Analyse erhaltenen Schüler:innenfähigkeiten sowie Itemschwierigkeiten für die nachfolgenden Untersuchungen verwendet werden.

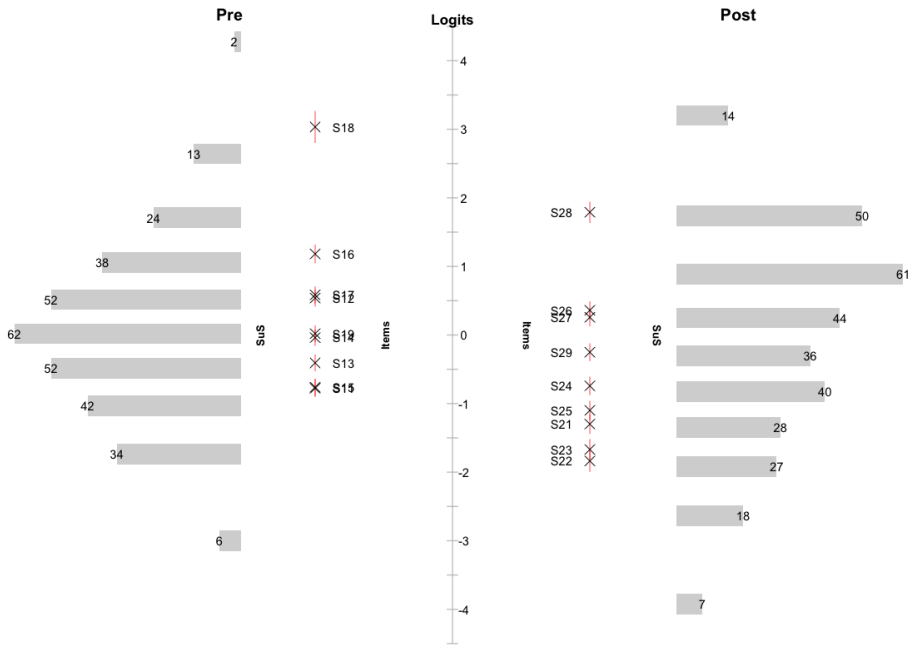


Abb. E.4: Darstellung der Schüler:innenfähigkeit θ_i und Itemschwierigkeiten ξ_j mit dazugehörigen Standardfehlern (rot) für den Prä- und Posterhebunszeitpunkt. Die Itemschwierigkeiten nehmen im Postzeitraum ab, während die Schüler:innenfähigkeitenn zunehmen.

Wright-Maps

Die Wright-Maps stellen die Schüler:innenfähigkeiten sowie die Itemschwierigkeiten in einem Diagramm dar. Eine Darstellung für die Prä- und Posterhebung kann in Abb. E.4 gesehen werden. Die durchschnittliche Schwierigkeit der Items betrug im Prätest $\overline{\xi_{\text{pre}}} = (0{,}38 \pm 1{,}19)$ Logits und im Posttest $\overline{\xi_{\text{post}}} = (-0{,}50 \pm 1{,}16)$ Logits. Im Abschnitt E.2.2 wird untersucht, ob die unterschiedlichen Schwierigkeitsgrade auf eine veränderte Funktionsweise des Tests oder auf eine durch die durchgeführte Intervention bedingte Fähigkeitszunahme der Lernenden zurückzuführen ist.

Abbildung E.4 zeigt, dass die Items zum Präzeitpunkt größtenteils einen Schwierigkeitsgrad von -1 bis 1 Logit aufweisen. Einzig das Item S18 weist mit 3 Logit einen deutlich höheren Schwierigkeitsgrad auf. Im Postzeitraum liegt ebenfalls eine Häufung vor, jedoch ist diese etwas verschoben im Bereich -2 bis 0,5 Logit. Das Item S28 weist wie im Prätest bereits einen deutlich höheren Schwierigkeitsgrad auf.

Grundlegend ist der Schwierigkeitsgrad des Tests im durchschnittlichen Schwierigkeitsgrad gut verteilt. Sowohl im Prä- als auch im Posttest gibt es bis auf das Item S18 bzw. S28 keine schweren Items. In diesem Bereich kann der Test somit nicht so gut differenzieren wie im durchschnittlichen Spektrum.

E.2.2 Messinvarianz

Für den Vergleich der Schüler:innenfähigkeiten zwischen dem Prä- und Posttest muss zunächst untersucht werden, ob die Items gleich funktionieren. Im vorigen Abschnitt konnte gezeigt werden, dass die Items im Prä- und Posttest unterschiedliche Schwierigkeitsgrade aufweisen (vgl. Abb. E.5). Die nachfolgende Analyse untersucht die Invarianz des Testinstruments zu den unterschiedlichen Erhebungszeitpunkten.

In Abbildung E.5 ist zu erkennen, dass die Itemkurven im Posttest weiter links zu finden sind und somit einen niedrigeren Schwierigkeitsgrad aufweisen. Ein genaueres Bild der Situation zeigt Tabelle E.2, die alle Itemschwierigkeiten für den Prä- und Posttest darstellt. Es ist zu erkennen, dass jedes Item im Posttest als einfacher identifiziert wird. In einer DIF-Analyse, in der die Itemschwierigkeiten für den gesamten Datensatz kalkuliert und nachträglich für den Prä- und Posttest bestimmt und gegeneinander getestet wurden, haben sich für die meisten Items signifikante Unterschiede gezeigt. Da jedoch die Itemschwierigkeit an der Schüler:innenfähigkeit skaliert wird, ist aufgrund der Intervention eher davon auszugehen, dass die durchschnittliche Schüler:innenfähigkeit absolut gestiegen ist. Da die Schüler:innenfähigkeit in der Rasch-Analyse auf 0 zentriert wird, muss eine weitere Analyse erfolgen, anhand derer untersucht werden kann,

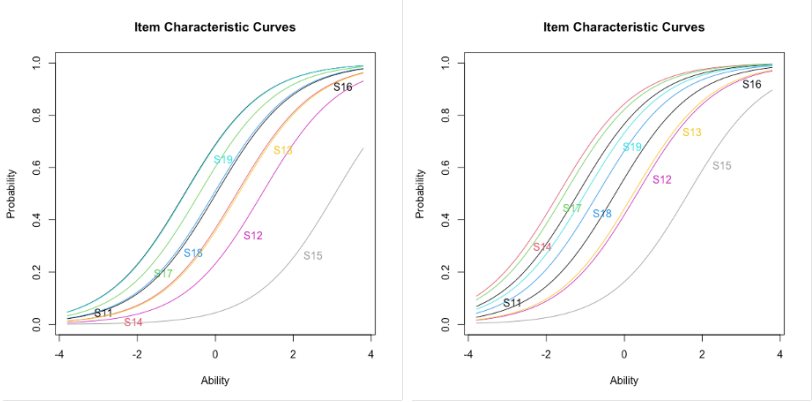


Abb. E.5: Darstellung aller ICCs Kurven für den Prä- (links) und Posterhebungszeitraum (rechts).

wie die Fähigkeit der Lernenden sich geändert hat. Dabei wurde sich an dem Racking- und Stacking–Verfahren von Wright (2003) orientiert, welches in diesem Kontext oftmals Verwendung findet (Cunningham & Bradley, 2010; Zhao et al., 2017). Dieses Verfahren hat gegenüber der DIF-Analyse den Vorteil, dass der Effekt der Intervention und die Veränderung der Schüler:innenfähigkeiten in der Auswertung berücksichtigt werden können. Beim Stacking werden die Datensätze des Prä- und Posttests übereinander geschachtelt (Cunningham & Bradley, 2010, S. 8). Abbildung E.6 zeigt den Vergleich der ermittelten Schwierigkeitsgrade.

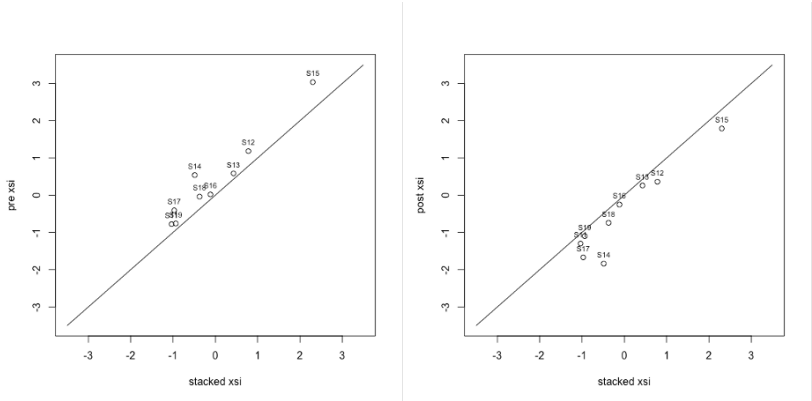
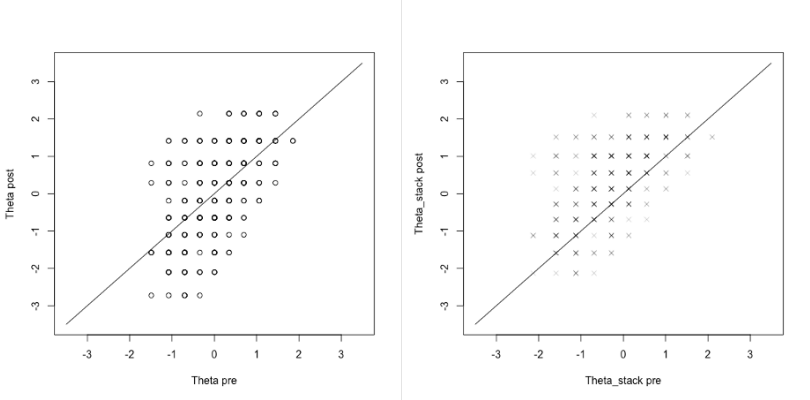
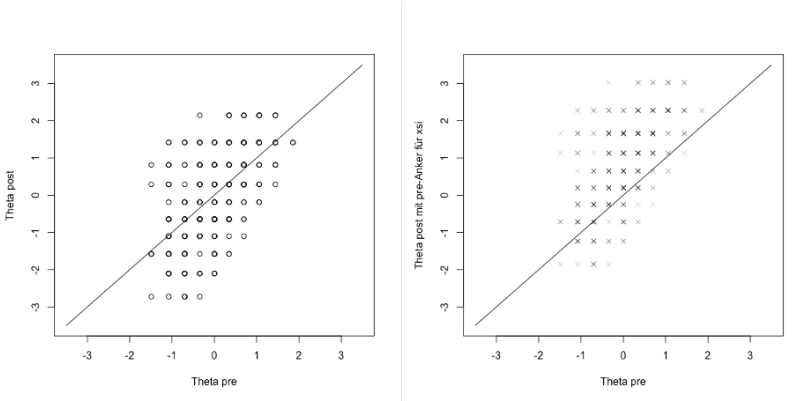


Abb. E.6: Vergleich der separat ermittelten Schwierigkeitsgerade mit dem auf Grundlage der Stacked-Method erhaltenen Schwierigkeitsgrad für den Prä- (links) und Posterhebungszeitraum (rechts).

Es ist zu erkennen, dass ein Großteil der Items an der Ursprungsgerade liegt. Die Ursprungsgerade charakterisiert in diesem Fall theoretische Werte für Items, die einen gleichen Schwierigkeitsgrad im Pre- und im Posttest aufweisen. Je näher



(a) Vergleich Schüler:innenfähigkeit zwischen separater Modellierung und der Stacked-Methode für den Prä- (links) und Posterhebungszeitraum (rechts).



(b) Vergleich der ermittelten Schüler:innenfähigkeit mit dem ursprünglichen (links) und verankerten (rechts) Schwierigkeitsgrad für den Posterhebungszeitraum (rechts).

Abb. E.7: Gegenüberstellung der geschätzten Schüler:innenfähigkeiten.

ein Item an dieser Geraden liegt, desto ähnlicher sind die Schwierigkeitsgrade in den jeweiligen Situationen. Während einige Items stark abweichen, beispielsweise S14 und S15 im Pretst und S14 im Posttest, und einige Items moderat abweichen (S17, S17 im Pretest, S17 und S15 im Posttest), zeigt sich, dass einige Items nahezu über gleiche Itemschwierigkeiten verfügen.

Durch das Stacking wurde eine Referenzschwierigkeit für jedes Item bestimmt, die verdeutlicht, dass die Items zu den unterschiedlichen Testungszeitpunkten unterschiedlich schwer waren. Gleichzeitig darf jedoch nicht vergessen werden, dass die Intervention die Kompetenzen im Umgang mit Daten und Unsicherheiten adressiert hat und damit die Schüler:innenfähigkeit sich geändert haben kann. Eine Gegenüberstellung der Schüler:innenfähigkeit zwischen den Erhebungszeitpunkten mittels beider Datensätze kann Abbildung E.7a betrachtet werden.

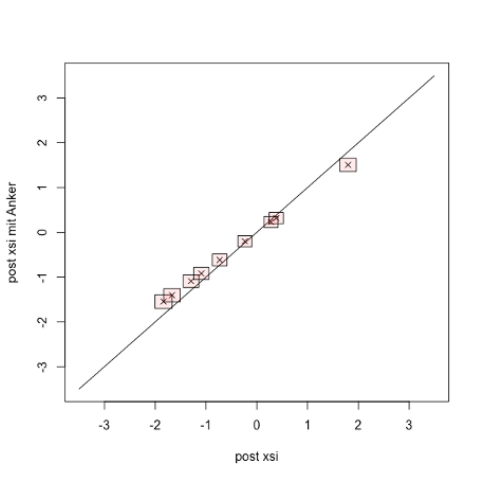


Abb. E.8: Darstellung des ursprünglich ermittelten und verankerten Schwierigkeitsgrad für den Posttest.

Der gestackte Datensatz zeigt im Vergleich zur Gegenüberstellung der separierten Datensätze eine Tendenz dazu, dass die Schüler:innenfähigkeit durch die Intervention gesteigert werden könnte. Um diesen Punkt deutlicher zu untersuchen, wird eine Rasch-Analyse auf Grundlage der Skala des Präzeitpunkts durchgeführt. Die erhaltenen Itemschwierigkeiten werden nun als Ankerpunkte für die Postanalyse gesetzt. Auf diese Weise wird die Fähigkeit der Schüler:innen mit demselben Instrument wie zum Präzeitpunkt erfasst. Somit ist eine Aussage bezüglich der Entwicklung der Schüler:innen möglich. Abbildung E.7b ist eine Gegenüberstellung der separierten und nicht verankerten Resultate gegenüber der verankerten Rasch-Analyse.

Abschließend wird nun untersucht, wie sich die Itemschwierigkeit ändert, wenn die Schüler:innenfähigkeit verankert wird. Als Verankerung werden die Fähigkeitswerte genommen, die mit dem Messintrument und der Skala des Präzeitpunkts bestimmt worden sind. Die auf diese Weise bestimmten Itemschwierigkeiten ξ_j werden mit den ursprünglich für den Postzeitraum bestimmten Itemschwierigkeiten verglichen.

Die Abbildung demonstriert die beiden unterschiedlichen Szenarien: Auf der x-Achse ist die Itemschwierigkeit aus dem Posttest dargestellt. Auf der y-Achse die Itemschwierigkeit des Posttests mit verankerten Schüler:innenfähigkeiten, die sich aus dem Skalenniveau des Pretests ergeben. Die Itemschwierigkeiten unterscheiden sich geringfügig in den jeweiligen Szenarien, lassen sich jedoch im Rahmen des Standardfehlers als gleich schwer bezeichnen. Als Nächstes wird die mittlere Itemschwierigkeit aus separaten Datensätzen für den Pre- und Posttest miteinander verglichen. Der Mittelwert des Unterschieds der Itemschwierigkeit liegt bei $-0{,}87$ Logits. Dieser Wert stimmt nahezu mit dem Zuwachs der

Tab. E.3: Übersicht der Itemschwierigkeiten und Itemparameter für die Items S1 bis S9 unter Verwendung des Stacked-Modells.

			Outfit			Infit		
Item	ξ	SE_ξ	Wert	t	p	Wert	t	p
S1	−1,03	0,10	0,96	−0,40	0,69	1,02	0,37	0,71
S2	0,78	0,09	0,98	−0,30	0,76	0,97	−0,70	0,48
S3	0,43	0,09	1,00	0,04	0,97	1,01	0,20	0,84
S4	−0,49	0,09	1,07	1,15	0,25	1,06	1,50	0,13
S5	2,30	0,12	0,94	−0,29	0,78	1,00	0,00	1,00
S6	−0,12	0,09	0,97	−0,43	0,67	0,96	−1,01	0,31
S7	−0,97	0,09	0,91	−1,17	0,24	0,96	−0,82	0,41
S8	−0,37	0,09	0,91	−1,49	0,14	0,93	−2,03	0,04
S9	−0,94	0,09	1,20	2,38	0,02	1,11	2,44	0,01

Schüler:innenfähigkeit überein. Da die Itemschwierigkeit des Posttests unabhängig von der zugrunde liegenden Schüler:innenfähigkeit ist (vgl. Abbildung E.8), kann der Unterschied der Schwierigkeitsgrade der Items mit dem Zuwachs der Schüler:innenfähigkeit erklärt werden. Für die weiterführenden Analysen kann somit davon ausgegangen werden, dass das Testinstrument zur Erfassung der Kompetenz im Umgang mit Daten und Messunsicherheiten longitudinal messinvariant ist.

E.2.3 Analyse des Rasch-Modells für beide Erhebungszeitpunkte

Nach Untersuchung der Passung zwischen den Modellen und Daten sowie dem Nachweis der Messinvarianz wird das Stacked-Modell für die weiterführenden Analysen für Abschnitt 6.2 verwendet. Nachfolgend werden die Kenngrößen und Parameter des Modells detailliert dargestellt.

Die Itemparameter des Stacked-Modells sind in Tabelle E.3 dargestellt. Für die Items S1 bis S7 zeigen sich keinerlei Auffälligkeiten bezüglich der Infit- und Outfit-Werte. Das Item S8 weist einen signifikanten Infit-Wert auf ($t = -2{,}09, p = 0,04$). Der Wert liegt jedoch innerhalb der Cut-off-Kriterien. Die größte Auffälligkeit zeigt das Item S9, welches signifikante Abweichungen im Outfit ($t = 2{,}38, p = 0,02$) und Infit ($t = 2{,}44, p = 0{,}01$) aufweist, jedoch liegen die Werte erneut innerhalb der Cut-off-Werte, sodass die Items im Test beibehalten werden.

Zwischen der empirischen und modellbasierten ICC zeigt sich eine stärkere Übereinstimmung als in der Modellierung für den Prä- und Postzeitpunkt (s.

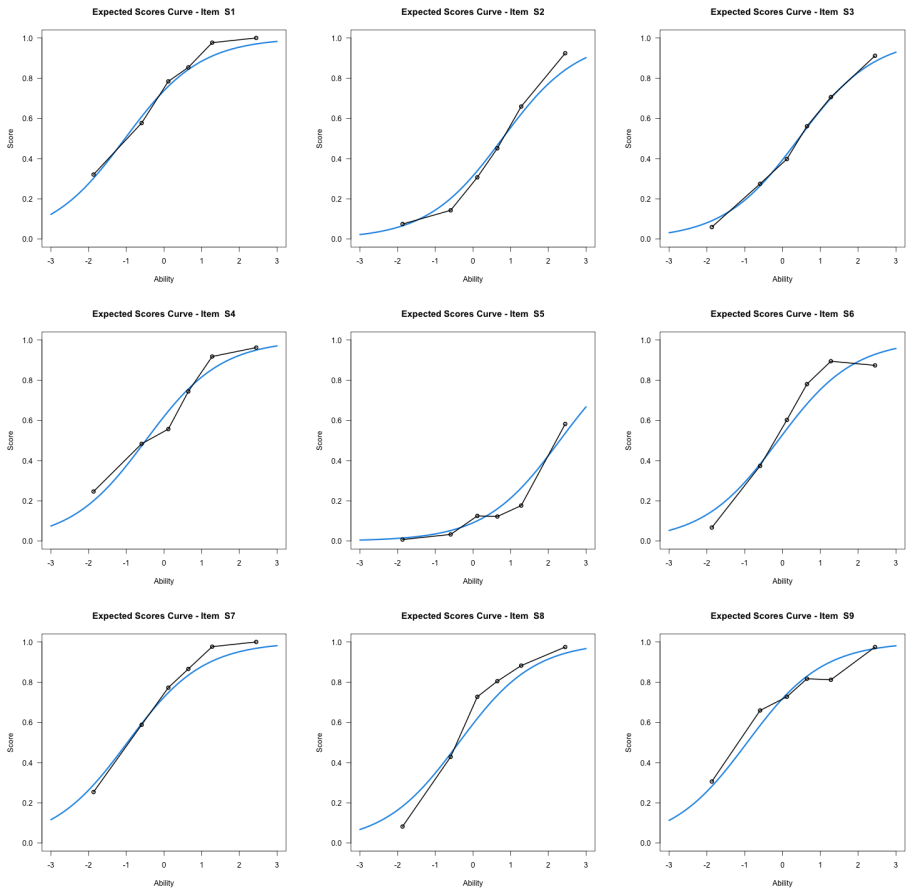


Abb. E.9: Darstellung der ICCs für die Items S1 bis S9 des *stacked*-Modells

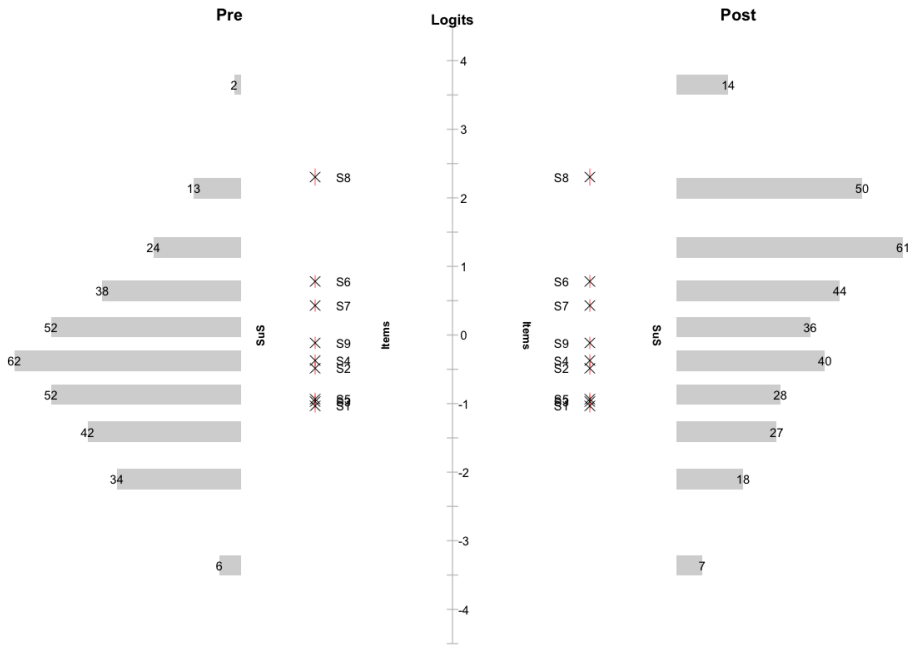


Abb. E.10: Darstellung der Schüler:innenfähigkeit θ_i und Itemschwierigkeiten ξ_j mit dazugehörigen Standardfehlern (rot) für den Prä- und Posterhebunszeitpunkt. Die Itemschwierigkeiten sind in diesem Fall konstant, da die Schüler:innenfähigkeiten für beide Zeitpunkte gleichzeitig unter Verwendung des Stacked-Modells ermittelt wurden.

Abb. E.2, E.3 und E.9). Es gibt durchaus leichte Unterschiede zwischen den ICCs. Beispielsweise zeigt sich beim Item `S6`, dass die empirische ICC nicht so gut zwischen leistungsstarken Schüler:innen differenziert wie die modellbasierte. Ein ähnlicher Unterschied zeigt sich beim Item `S9`, wo es Abweichungen zwischen den empirischen und modellbasierten ICCs im Bereich 0,5 Logit bis 1,5 Logit gibt. Die genannten Unterschiede sind jedoch sehr gering und können für die Praxis vernachlässigt werden. Die ICCs sprechen daher für eine gute Passung zwischen den Daten und dem Rasch-Modell.

Abschließend wird die Wright-Map des Stacked-Modells untersucht (s. Abb. E.10). Im Vergleich zu Abbildung E.4 wird der Vorteil der Stacked-Methode deutlich: Der Schwerpunkt der Modellierung liegt in dem Unterschied zwischen den beiden Zeitpunkten, da die Itemschwierigkeit auf Grundlage beider Datensätze ermittelt wurde. Es zeigt sich, dass der verwendete Test zum Präzeitpunkt etwas zu schwer ist und etwa bei 80 Schüler:innen die latente Fähigkeit unterhalb des leichtesten Items liegt. Für den Großteil der Schüler:innen zeigt sich jedoch eine gute Passung zwischen Schüler:innenfähigkeit $\theta_i$und Itemschwierigkeit ξ_i. Im Posttest zeigt sich ein anderes Bild. Es gibt weiterhin Schüler:innen, deren Fähigkeiten unterhalb des leichtesten Items liegen (ca. 50). Jedoch ist eine deutliche Verschiebung der Schüler:innenfähigkeit nach oben zu beobachten. Bei der Gegenüberstellung der Schüler:innenfähigkeiten und Itemschwierigkeiten lässt sich die Tendenz erkennen, dass der Test den Schüler:innen in diesem Fall leichter fiel. Da die Itemschwierigkeit auf Grundlage des gemeinsamen Datensatzes von Prä- und Posterhebungszeitraum bestimmt wurde, wird der Kompetenzzuwachs der Schüler:innen anhand der Wright-Map verdeutlicht.

E.3 Latente Modellierung

Im nachfolgenden Abschnitt werden grundlegende Eigenschaften der latenten Modellierung präsentiert. Dabei wird auf die Konstruktion der einzelnen Messmodelle, die Güte der Modelle sowie die Untersuchung auf longitudinale Messinvarianz eingegangen. Die Ergebnisse des vorigen Abschnitts E.2 werden verwendet, um die Kompetenz in Abschnitt E.3.3 in das latente Modell einzuspeisen.

E.3.1 Evaluation der Passung zwischen Daten und Modellen

„Because there is no single statistical significance test that identifies a correct model given the sample data, it is necessary to take multiple criteria into consideration and to evaluate model fit on the basis of various measures simultaneously.“

Schermelleh-Engel et al., 2003, S. 31

Die Evaluation der Passung zwischen den Daten und dem theoretischen Modell kann auf unterschiedliche Arten durchgeführt werden. Im Folgenden werden drei übergeordnete Vergleichsarten vorgestellt und die Relevanz für diese Arbeit herausgearbeitet (Schermelleh-Engel et al., 2003). Aufgrund der unterschiedlichen Konzepte der Fit-Indizes ist es sinnvoll, die Modellevaluation auf einer Vielzahl dieser zu realisieren. Eine Übersicht der Eigenschaften der Fit-Indizes kann den folgenden Abschnitten entnommen werden.

Inferenzstatistische Evaluation der Modellgüte

Die Bewertung der Passung zwischen dem implizierten Modell und den empirischen Daten wird anhand eines Vergleichs der Varianz-Kovarianz-Matrizen durchgeführt. Dabei wird überprüft, ob die wahre Varianz-Kovarianz-Matrix Σ gleich der modellimplizierten Varianz-Kovarianz-Matrix $\Sigma(\theta)$ ist, wobei θ die exakten Modellparameter beinhaltet. Beide Matrizen können jedoch nicht exakt erfasst werden und werden daher mit empirischen Schätzungen $\hat{\Sigma}$ und $\Sigma(\hat{\theta})$ approximiert, wobei $\hat{\theta}$ die geschätzten Modellparameter beinhaltet. Eine inferenzstatistische Evaluation des Modells kann daraufhin anhand eines χ^2-Tests durchgeführt werden, wobei die Nullhypothese

$$H_0: \quad \hat{\Sigma} = \Sigma(\hat{\theta})$$

überprüft wird (Brown, 2006, S. 81). Der Test sollte für eine gute Modellpassung nicht signifikant ausfallen.

Deskriptive Evaluation der Gesamtmodellgüte

Aufgrund der bekannten Limitierungen des χ^2-Tests wurden alternative Fit-Indizes entwickelt, die auf deskriptiven Statistiken beruhen und weniger anfällig für die Stichprobengröße sind (Meade et al., 2008). In der vorliegenden Arbeit werden die Fit-Indizes RMSEA und SRMR verwendet.

Der Fit–Index RMSEA stellt ein Maß für die Abweichung der approximierten Differenz $\hat{\Sigma} - \Sigma(\hat{\theta})$ dar und evaluiert somit eine ungefähre Gleichheit der Varianz-Kovarianz-Matrix. Eine gute Passung des Modells liegt vor, wenn der RMSEA

Wert kleiner gleich 0,05 ist. Werte kleiner als 0,08 zeigen einen zufriedenstellenden Fit zwischen dem Modell und den Daten an (Borowne & Cudeck, 1993). Hu und Bentler (1999) schlagen einen strengeren Cut-off-Wert von 0,06 vor. Das 90 %-Konfidenzintervall kann verwendet werden, um die Präzision der Schätzung zu evaluieren. Beinhaltet die linke Grenze des Intervalls die Null, so ist im Rahmen der Unsicherheit eine exakte Passung zwischen Daten und Modell denkbar (Schermelleh-Engel et al., 2003). Zudem wird ein Signifikanztest bezüglich der Nullhypothese RMSEA$\leq$ 0,05 durchgeführt, welcher für eine gute Passung keine Signifikanz zeigen sollte.

Einen anderen Ansatz verfolgt der Fit-Index SRMR, welcher die standardisierten Residuen der Modellanpassung evaluiert. In diesem Fall gilt als Faustregeln, dass Werte unter 0,05 eine gute Passung zwischen Daten und Modellpassung zeigen und Werte unter 0,10 eine akzeptable Passung suggerieren (Hu & Bentler, 1999).

Deskriptive Evaluation basierend auf Modellvergleichen

Eine weitere Art, die Güte von Modellen deskriptiv zu beschreiben, ist die Verwendung von Indizes, die auf dem Vergleich von Modellen beruhen. In der vorliegenden Arbeit wird sich auf den CFI berufen. Die Berechnung des CFIs basiert auf den χ^2 und df Parametern des spezifizierten Modells sowie entsprechenden Nullmodells und kann durch

$$1 - \frac{\max\left(\chi_z^2 - df_z, 0\right)}{\max\left(\chi_z^2 - df_z, \chi_n^2 - df_n, 0\right)}$$

bestimmt werden, wobei die χ_z^2 und df_z zum spezifizierten Zielmodell gehören und χ_n^2 und df_n zum entsprechenden Nullmodell (Bentler, 1990). Werte nahe bei 0,95 und darüber sind Indikatoren für einen guten Modellfit (Hu & Bentler, 1999). Werte zwischen 0,90 und 0,95 sprechen für eine akzeptable Passung zwischen Modell und Daten (Bentler, 1990). Der CFI weist eine geringe Abhängigkeit von der Stichprobengröße auf (Schermelleh-Engel et al., 2003).

Für die Evaluation wird auf den χ^2-Test sowie die Fit-Indizes zurückgegriffen. Wie in Abschnitt 6.4 dargestellt, besitzen letzte den Vorteil, dass die Abweichung der Varianz hinsichtlich der Relevanz bewertet werden kann. Eine Übersicht bezüglich der Cut-off-Werte der Fit-Parameter findet sich in Tabelle E.4. Bei der Evaluation der Modellpassung werden alle zuvor genannten Parameter betrachtet.

Bei widersprüchlichen Aussagen bezüglich der Modellpassung werden die deskriptiven Parameter gegenüber dem χ^2-Tests stärker bewertet. Dies begründet sich darin, dass die Abweichung der Varianz-Kovarianz-Matrix durch die deskriptiven Fit-Indizes auf eine inhaltliche Relevanz untersucht werden kann. Weiterhin sind diese Parameter größtenteils unabhängig von der Stichprobengröße und weisen somit nicht die Schwäche von inferenzstatistischen Fit-Indizes auf.

hierarchische Modelle

In der vorliegenden Arbeit werden insbesondere im Zusammenhang mit der Untersuchung auf Messinvarianz hierarchische Modelle untersucht. Hierarchisch bedeutet hierbei, „dass ein Modell durch zusätzliche Restriktionen als Spezialfall aus einem allgemeineren Modell hervorgeht“ (Geiser, 2010, S. 60). In der vorliegenden Arbeit finden sich Restriktionen in Form von Autokorrelationen zwischen denselben Indikatoren zu verschiedenen Messzeitpunkten, korrelierten Residualvarianzen von unterschiedlichen Indikatoren und Fixierungen von Faktorladungen und Achsenabschnitten für die Untersuchung von schwacher und starker Messinvarianz.

Die komplexeren und restriktiveren Modelle folgen oftmals einem Forschungsinteresse oder sind eine Grundlage für die Verwendung von bestimmten statistischen Methoden bzw. Vergleichen. Ein hierarchisches Modell sollte für die Auswertung nur dann verwendet werden, wenn die Modellpassung im Vergleich zum nicht restringierten Modell nicht schlechter wird.

Die empirische Bewertung der Modellpassungsdifferenz kann durch den χ^2-Differenztest, auch *Likelihood-Ratio*-Test genannt, bestimmt werden. Dabei werden die Differenz $\Delta\chi^2$ sowie Δdf aus den Parameterin χ^2 und df der beiden zu vergleichenden Modelle bestimmt. Ist der resultierende χ^2-Test mit gegebenem $\chi^2 = \Delta\chi^2$ und $df = \Delta df$ signifikant, so liegt eine signifikante Verschlechterung der Modellpassung vor und das weniger restriktive Modell sollte beibehalten werden. Jedoch ist anzumerken, dass dieses Verfahren ebenfalls einen Nullhypothesentest darstellt und somit die gleichen Schwächen aufweist wie der χ^2-Test (vgl. Abschnitt E.3.1).

Dem gegenüber steht die Bewertung des Modellpassungsunterschieds auf Grundlage deskriptiver Fit-Indizes (Cheung & Rensvold, 2002, S. 239). Dabei wird die Differenz

$$\Delta\mathrm{FI} = \mathrm{FI}_r - \mathrm{FI}_{ur} \tag{E.1}$$

mit FI_r als Fit-Index des restriktiven Modells und FI_{ur} als Fit-IKndizes des nicht (bzw. weniger) restriktiven Modells. Bezüglich eines Cut-off-Werts gibt es verschiedene Abschätzungen. Cheung und Rensvold (2002) vertreten die Ansicht, dass bei einem $\Delta\mathrm{CFI} < -0{,}01$ von einer signifikanten Verschlechterung des Modells auszugehen ist und die Nullhypothese entsprechend abgelehnt werden sollte. Strengere Vorgaben machen Meade et al. (2008), die die allgemeine Abschätzung von Cheung und Rensvold kritisch sehen, da diesem Verfahren in einigen Szenarien die Teststärke fehlt. Sie schlagen daher ein strengeres Kriterium von $\Delta\mathrm{CFI} < -0{,}002$ als allgemeine Faustregel vor. Weitere Untersuchungen wurden von F. F. Chen (2007) durchgeführt. Für Stichproben mit $n > 300$ empfiehlt er die Cut-off-Werte $\Delta\mathrm{CFI} \leq -0{,}010$ und zusätzlich $\Delta\mathrm{RMSEA} \geq 0{,}015$.

Beim Vergleich von hierarchischen Modellen wird sowohl auf inferenz- als auch auf deskriptivstatistische Fit-Indizes zurückgegriffen. Bei sich widerspre-

chenden Ergebnissen hinsichtlich der Passung wird die Aussage der Fit-Indizes bevorzugt bewertet, da diese eine Bewertung der Abweichung darstellt. Insbesondere bei einer Untersuchung der Messinvarianz wird dieses Vorgehen auch von Meade et al. (2008, S. 509) vorgeschlagen. Bei der Evaluation der Modelle werden die Cut-off-Werte von F. F. Chen (2007) verwendet, da die vorliegende Arbeit ebenfalls Mplus[1] verwendet wird und somit die gleiche Berechnung des CFIs gewährleistet werden kann. Die Diskrepanz der Cut-off-Werte für das ΔCFI zwischen Cheung und Rensvold (2002) sowie Meade et al. (2008) lässt sich durch eine veränderte Berechnung des CFIs in dem Statistikprogramm LISREL[2] (Meade et al., 2008, S. 590) erklären.

E.3.2 Schätzmethoden

Die inferenzstatistische Evaluation der Modellgüte basiert auf dem Vergleich der wahren Varianz-Kovarianzmatrix (Σ) mit der modellimplizierten Varianz-Kovarianz-Matrix ($\Sigma(\theta)$. Beide Matrizen werden empirisch geschätzt ($\hat{\Sigma}, \Sigma(\hat{\theta})$). Für die Schätzung der Matrizen können verschiedene Verfahren herangezogen werden (Eid et al., 2010, S. 899). Die vorliegende Arbeit verwendet das Maximum-Likelihood (ML)-Verfahren, um die Parameter zu schätzen. Dabei ist das Ziel, die Abweichung der modellimplizierten Varianz-Kovarianz-Matrix zur Stichprobenmatrix unter Berücksichtigung der Modellrestriktionen zu minimieren. Das ML-Verfahren verwendet die Gleichung

$$F_{ML} = \ln \Sigma(\hat{\theta}) - \ln |S| + tr[S \cdot \Sigma(\hat{\theta})^{-1}] - p, \tag{E.2}$$

wobei S die Stichprobenmatrix, $\Sigma(\hat{\theta})$ die modellimplizierte auf der empirischen Schätzung der Parameter beruhende Varianz-Kovarianz-Matrix und p die Anzahl der manifesten Variablen bezeichnet (Eid et al., 2010, S. 899). Die Fit-Funktion F_{ML} wird dabei so minimiert, dass die Wahrscheinlichkeit für das Erhalten der Stichprobenmatrix S maximiert wird.

Dem ML-Verfahren liegen einige Voraussetzungen zugrunde, die getestet werden müssen, um keine verzerrten Schätzwerte zu erhalten (Brown, 2006, S. 75). Vorausgesetzt wird, dass die analysierten Variablen eine multivariate Normalverteilung aufweisen. Zudem werden eine Unabhängigkeit der Beobachtungen und eine korrekte Modellspezifikation gefordert (Eid et al., 2010, S. 899).

Die Familie der ML-Verfahren weist den Vorteil auf, dass zusätzlich zu den Parametern Standardfehler geschätzt werden können, die weiterführend zur Berechnung von Konfidenzintervallen verwendet werden können (Eid et al., 2010, S. 899). Dies ermöglicht die Durchführung von statistischen Signifikanztests

[1] Muthén und Muthén, 2017
[2] Jöreskog, 2001

bezüglich der geschätzten Parameter (Brown, 2006, S. 75). Außerdem wird die Gleichung E.2 in der Berechnung verschiedener Fit-Indizes aus Abschnitt E.3.1 benötigt. Auf Grundlage dieser Vorteile wurde in der vorliegenden Arbeit auf die ML-Schätzverfahren zurückgegriffen.

Die Verletzung der Voraussetzungen beeinflusst die Parameterschätzungen kaum, sondern äußerst sich insbesondere in verzerrten Standardfehlern. Eine ausführliche Diskussion dieses Aspekts findet sich in dem Werk von Schweizer und DiStefano (2016). Die Erfassung der Zustimmung zu den Argumentkategorien erfolgt anhand von fünfstufigen Likert-Skalen (vgl. Abschnitt 6.1.3). Ähnliche Skalen finden häufig Anwendung in der empirischen Forschung (Finney & DiStefano, 2013), obwohl die Betrachtung als intervallskalierte Variable an Voraussetzungen geknüpft ist und Likert-Skalen in einigen Fällen keine Normalverteilung aufweisen können. In diesen Fällen ist von der Verwendung von ML-Verfahren abzusehen (Kline, 2015, S. 257 f.).

Da in der vorliegenden Arbeit fünfstufige Likert-Skalen verwendet werden und das Testinstrument asymmetrische Verteilungen aufweisen kann (Ludwig, 2017), wird auf robuste Schätzverfahren zurückgegriffen. Es wird auf den MLR-Schätzer des mPlus Programms zurückgegriffen (Muthén & Muthén, 2017), welcher zur Familie der ML-Verfahren zu zählen ist, jedoch robuste Standardfehler aufweist (Kline, 2015, S. 238). Bei robusten Standardfehlern handelt es sich um Schätzungen der Standardfehler, welche durch die Verletzung der Normalität nicht beeinflusst werden und dadurch keine Verzerrung aufweisen. Die Robustheit von Schätzverfahren unterliegt oftmals demselben Prinzip und kann auf die meisten verbreiteten Schätzmethoden angewandt werden (Savalei, 2014).

E.3.3 Messmodelle

Die Evaluation der Messmodelle wird exemplarisch für den latenten Faktor Intuition dargestellt. Die restlichen Messmodelle wurden analog evaluiert und angepasst. Bevor ein Messmodell für beide Erhebungszeitpunkte betrachtet werden kann, werden zunächst die Messmodelle für den Prä- und Postzeitpunkt einzeln evaluiert.

Bis auf das Modell `MeMoMuPost` zeigt sich für alle Messmodelle eine gute Passung zu den Daten. Es zeigen sich keine Auffälligkeiten im χ^2-Test und beim 90 % Konfidenzintervall des RMSEA Fit-Indizes. Auf Grundlage von Gleichung **??** ergibt sich theoretisch ein Freiheitsgrad von $\mathrm{d}f = 5$ für jedes Modell. Die Abweichungen resultieren aus Anpassungen der Modelle bezüglich der korrelierten Residualvarianzen. Auf diese Weise konnte eine zufriedenstellende Passung zwischen Daten und Modell gewährleistet werden. Eine Korrelation der Residuen wurde nur durchgeführt, wenn sich diese theoretisch auf Grundlage des

Tab. E.4: Fit-Indizes für die jeweiligen Messmodelle (`MeMo[...]`) für die entsprechende Zustimmung der Argumentkategorien Intuition ([...]`Int`), Expertenwissen ([...]`Exp`), Messunsicherheit ([...]`MU`) und Daten als Evidenz ([...]`Evi`) basierend auf dem Testinstrument von Ludwig (2017) für den Prä- und Posterhebungszeitpunkt. Die Darstellung der Modelle kann in Abschnitt E.4 betrachtet werden.

Modell	χ^2	df	p	CFI	RMSEA [90% CI]; p	SRMR
`MeMoIntPre`	3,05	4	0,55	1	0 [0;0,07];0,83	0,01
`MeMoIntPost`	2,09	5	0,84	1	0 [0;0,05];0,96	0,01
`MeMoExpPre`	5,47	3	0,14	0,99	0,05 [0;0,12];0,41	0,02
`MeMoExpPost`	1,83	4	0,77	1	0 [0;0,06];0,93	0,01
`MeMoMuPre`	2,65	2	0,27	1	0,03 [0;0,12];0,51	0,02
`MeMoMuPost`	4,73	2	0,09	0,99	0,07 [0;0,14];0,28	0,01
`MeMoEviPre`	2,57	4	0,63	1	0 [0;0,07];0,87	0,01
`MeMoEviPost`	2,63	3	0,45	1	0 [0;0,09];0,73	0,01

Testinstruments begründen ließ. Die Vorgehensweise wird exemplarisch für das Modell `MeMoIntPre` demonstriert. Die Anpassung erfolgte auf analoge Weise. Der auffällige RMSEA Parameter des Modells `MeMoMuPost` (RMSEA = 0,07), der im Rahmen seines Konfidenzintervalls einen möglichen Wert von 0,14 annimmt, kann vernachlässigt werden, da alle anderen Fit-Indizes eine gute Passung zwischen Daten und Modell aufweisen (Brown, 2006, S. 89). Im weiteren Vorgehen wird dieses Modell daher auf die gleiche Art behandelt, wie die zuvor genannten.

Als Ansatzpunkt für die Überarbeitung der Modelle wurden die Modifikations-Indizes betrachtet. Dabei handelt es sich um Parameter, die eine Aussage darüber ermöglichen, wie stark ein bestimmter geschätzter Parameter bzw. eine Restriktion die Passung des Modells zu den Daten negativ beeinflusst.

Bei der Analyse des Messmodells zur Erfassung der Zustimmung zur Argumentkategorie Intuition hat sich gezeigt, dass eine zusätzliche Modellierung einer korrelierten Residualvarianz für die Indikatoren 4 und 5 (vgl. Anhang Testinstrument Ludwig) besteht. Bei Betrachtung der Indikatoren hat sich gezeigt, dass beide Items das Maß des Gefühls qualitativ beschreiben und somit eine korrelierte Residualvarianz auf Grundlage der Itemformulierungen nachvollziehbar ist (Kline, 2015). Im Rahmen von longitudinalen Analysen ist die Verwendung von korrelierten Residualvarianzen plausibel (Pentz & Chou, 1994). Die Fit-Indizes dieses Modells sind in Tabelle E.4 als `MeMoIntPre` dargestellt. Weitere Residualvarianzen wurden im Sinne der Parametersparsamkeit nicht berücksichtigt, wenn die Passung zwischen dem Modell und den Daten zufriedenstellend war. Ein möglicher Overfit des Modells an die Daten soll so verhindert werden. Diese Tatsache führt auch zu den unterschiedlichen Freiheitsgraden des Messmodells für die Zustimmung zur Argumentkategorie Intuition zum Prä- und Postzeitpunkt (vgl. Abschnitt E.4).

Single-Item Indikator

E.3.4 Messinvarianz

Die einzelnen Messmodelle aus Tabelle E.4 werden im nachfolgenden zu longitudinalen Messmodellen zusammengefasst und hinsichtlich ihrer Eigenschaft der Messinvarianz untersucht. Der Vergleich der hierarchischen Modelle geschieht auf Grundlage der in Abschnitt E.3.3 dargestellten Fit-Indizes.

Das longitudinale Messmodelle (Int-konfigural) für die Erfassung der Zustimmung zur Argumentkategorie Intuition weist auf Grundlage der deskriptiven Fit-Indizes eine gute Passung zu den Daten auf. Der inferenzstatistische Fit-Index χ^2 zeigt eine signifikante Abweichung zwischen der theoretischen und empirischen Varianz-Kovarianz-Matrix an. Eine Untersuchung der Modifikations-Indizies zeigt, dass eine Autokorrelation der Indikatoren die Passung des Modells verbessern könnte. Vor dem Hintergrund der durchgeführten longitudinalen Untersuchung erscheint es sinnvoll, eine Autokorrelation der Residualvarianzen in das Modell einzubeziehen. Das entsprechende Modell (Int-konfigural-`AR`) zeigt eine bessere Passung der Daten hinsichtlich der deskriptiven Fit-Indizes. Der χ^2-Test fällt für dieses Modell nicht mehr signifikant aus. Für die weiteren Analysen wird das konfigurale Modell mit Autokorrelation der Residualvarianzen (Int-konfigural-`AR`) als Vergleichsgrundlage für die Vergleiche der hierarchischen Modelle verwendet.

Hinsichtlich der schwachen Messinvarianz zeigt sich eine gute Passung zwischen den Daten und dem Modell (Int-schwach-`AR`) bezüglich der deskriptiven Fit-Indizes. Der χ^2-Test fällt signifikant aus. Ein Vergleich zwischen den Modellen

mit konfiguraler und schwacher Messinvarianz zeigt, dass die Verschlechterung des Modells inferenzstatistisch signifikant ist. Da die Veränderung der deskriptiven Fit-Indizes unterhalb der Cut-off-Werte liegt (Δ CFI = -0,006 und ΔRMSEA = 0,010) wird das Modell akzeptiert und für die weiteren Analysen von einer schwachen Messinvarianz ausgegangen.

Die Untersuchung bezüglich starker Messinvarianz (Int-stark-`AR`) zeigt eine akzeptable Passung auf. Es zeigt sich jedoch, dass das Modell im Vergleich zur schwachen Messinvarianz eine Verschlechterung darstellt. Die starke Messinvarianz wird daher für das Messmodell abgelehnt.

Auf Grundlage der Modifikations-Indizes zeigte sich, dass die Bedingung, alle Achsenabschnitte der Indikatoren zu fixieren, zu restriktiv ist. Für die Etablierung von partieller Messinvarianz wird der Achsenabschnitt für Indikator 1 für den Prä- und Postzeitpunkt separat geschätzt. Dieses Modell (IKnt-partiell-stark-`AR`) weist trotz des signifikanten χ^2-Tests eine gute Passung zu den Daten auf. Im Vergleich zur schwachen Messinvarianz (Int-schwach-`AR`) zeigt sich, dass die partielle starke Messinvarianz angenommen werden kann, da die Verschlechterung des Modells unterhalb der Cut-off-Werte für die deskriptiven Fit-Indizes liegt.

Das longitudinale Messmodell für die Zustimmung zur Argumentkategorie Expertenwissen (Exp-konfigural) weist keine gute Passung zu den Daten auf. Der χ^2-Test ist signifikanz und der RMSEA Wert von 0,094 spricht auch für eine schlechte Passung. Eine Berücksichtigung der Autokorrelation von Residualvarianzen für die Indikatoren des latenten Faktors zeigt sich eine deutliche Verbesserung (Int-konfigural-`AR`). Trotz des signifikanten χ^2-Tests zeigt sich eine gute Passung auf Grundlage der deskriptiven Fit-Indizes. Die konfigurale Messinvarianz mit autokorrelierten Residualvarianzen wird für die weiteren Analysen angenommen.

Im nächsten Schritt wurde das Vorliegen von starker Messinvarianz untersucht. Das Modell weist eine signifikante Abweichung bezüglich der Varianz-Kovarianz-Matrix auf, die sich jedoch nicht auf die Fit-IKndizes niederschlägt. Dort zeigt sich eine gute (CFI = 0,982) bis zufriedenstellende Passung (RMSEA = 0,055; SRMR = 0,055). Beim Vergleich zur schwachen Messinvarianz zeigt sich eine deutliche Verschlechterung des Modells, sodass in diesem Fall ein Vorliegen von starker Messinvarianz abgelehnt wird.

Die Restriktion der konstanten Achsenabschnitte für die beiden Erhebungszeitpunkte wird für einen Indikator aufgehoben, um das Vorliegen von partieller starker Messinvarianz zu untersuchen (Exp-partiell-stark-`AR`). Das Modell zeigt eine gute Passung zu den Daten auf und der Vergleich zum Modell mit schwacher Messinvarianz (Exp-schwach-`AR`) zeigt, dass die Verschlechterung der deskriptiven Fit-Indizes unterhalb der Cut-off-Indizes liegt. Es wird somit ein Vorliegen von partieller starker Messinvarianz für die weiteren Analysen angenommen.

Für die Zustimmung zur Argumentkategorie Messunsicherheit wurde zu-

nächst im longitudinalen Messmodell das Vorliegen von konfiguraler Messinvarianz überprüft. Alle Fit-Indizes deuten in diesem Zusammenhang auf eine gute Passung des Modells (Mu-konfigural) hin. Aufgrund der longitudinalen Untersuchung wurde ebenfalls eine Autokorrelation der Indikatorresidualvarianzen untersucht (Int-konfigural-`AR`). Alle Fit-Indizes deuten auf eine verbesserte Passung zwischen Daten und Modell hin, sodass für die weiteren Analysen dieses Modell verwendet wird.

Das Modell mit implementierter schwacher Messinvarianz zeigt ebenfalls eine gute Passung. Ein Vergleich zwischen den Modellen mit schwacher (Mu-schwach-`AR`) und konfiguraler (Mu-konfigural-`AR`) Messinvarianz zeigt, dass keine Verschlechterung des Modells vorliegt. Die schwache Messinvarianz kann für das Messmodell daher angenommen werden.

Die nächste Stufe der Messinvarianz wird über die starke Messinvarianz des Messmodells etabliert. Dieses Modell (Mu-stark-`AR`) weist bezüglich aller Parameter eine gute Passung zu den Daten auf. Der Vergleich zwischen schwacher und starker Messinvarianz zeigt, dass trotz des signifikanten Resultats bezüglich des χ^2-Tests auf Grundlage der deskriptiven Fit-Indizes keine wesentliche Verschlechterung festgestellt werden kann. Für dieses Messmodell wird somit die starke Messinvarianz angenommen.

Abschließend wird das Messmodell für die Erfassung der Zustimmung zur Argumentkategorie Daten als Evidenz evaluiert. Das konfigurale Messmodell (Evi-konfigural) weist eine gute Passung auf. Nach Modellierung der Autokorrelation zwischen den Residualvarianzen der Indikatoren zeigt sich eine Verbesserung der Fit-Indizes. Insbesondere der χ^2-Test zeigt, dass die Abweichung zwischen der modellbasierten und empirischen Varianz-Kovarianz vernachlässigbar ist. Auf Grundlage dieser Parameter wird dieses Modell für als Grundlage für die kommenden Modelle verwendet.

Eine Fixierung der Faktorladungen zwischen den Erhebungszeitpunkten führt zu keiner Verschlechterung des Modells (Evi-schwach-`AR`). Die Modellpassung weist gute Fit-Indizes auf und im Vergleich zum konfiguralen Modell mit autokorrelierten Residualvarianzen zeigt sich keine Verschlechterung. Es wird daher davon ausgegangen, dass eine schwache Messinvarianz vorliegt.

Die Modellierung der starken Messinvarianz weist eine auffällige χ^2-Statistik auf. Bezüglich der deskriptiven Fit–Indizes zeigt sich eine gute bis zufriedenstellende Passung des Modells. Der Vergleich der Modelle mit schwacher (Evi-schwach-`AR`) und starker (Evi-stark-`AR`) Messinvarianz zeigt eine Verschlechterung des Modells bezüglich deskriptiver und inferenzstatistischer Fit-Indizes. Starke Messinvarianz wird für dieses Messmodell daher abgelehnt.

Eine Abschwächung der Restriktionen mündet in dem Modell mit implementierter partieller starker Messinvarianz (Evi-partiell-stark-`AR`). Das Modell weist eine bessere Passung zu den Daten auf als das Modell mit vollständiger starker Messinvarianz. Im Vergleich zum Modell mit schwacher Messinvarianz zeigt

sich, dass eine leichte Verschlechterung vorliegen könnte, da die Veränderung der deskriptiven Fit-Indizes unterschiedliche Schlussfolgerungen zulässt. Vor dem Hintergrund der sehr guten Passung des Modells zu den Daten wird sich dazu entschieden, die partielle skalare Messinvarianz für dieses Messmodell anzunehmen.

Tab. E.5: Übersicht der Modellvergleiche zwischen den hierarchischen Modellen im Kontext der Messinvarianz.

Fak	Messinvarianz	χ^2	df	p	CFI	RMSEA [90 %CI]; p	SRMR	$\Delta\chi^2$	p	ΔCFI	ΔRMSEA
Int											
	konfigural	56,6	33	< 0,05	0,984	0,046 [0,025; 0,067]; 0,57	0,029				
	konfigural-AR	35,9	28	0,14	0,995	0,030 [0; 0,055]; 0,90	0,025				
	schwach-AR	48,8	32	< 0,05	0,989	0,040 [0,013; 0,062]; 0,75	0,043				
	stark-AR	83,3	36	< 0,05	0,968	0,064 [0,046; 0,082]; 0,10	0,063				
	partiell stark-AR	60,8	35	< 0,05	0,983	0,048 [0,026; 0,067]; 0,55	0,046				
	konfigural-AR vs. schwach-AR							15,0	< 0,05	-0,006	0,010
	schwach-AR vs. stark-AR							40,8	< 0,05	-0,021	0,024
	schwach-AR vs.partiell stark-AR							10,2	< 0,05	-0,006	0,010

AR weist darauf hin, dass eine Autokorrelation zwischen den entsprechenden Indikatorresiduen vorliegt.

... Tabelle geht auf der nächsten Seite weiter

Tab. E.5: Übersicht der Modellvergleiche zwischen den hierarchischen Modellen im Kontext der Messinvarianz.

Fak	Messinvarianz	χ^2	df	p	CFI	RMSEA [90 %CI]; p	SRMR	$\Delta\chi^2$	p	ΔCFI	ΔRMSEA
Exp											
	konfigural	119,5	31	< 0,05	0,922	0,094 [0,076; 0,112]; 0,00	0,055				
	konfigural-AR	47,8	26	< 0,05	0,980	0,052 [0,029; 0,074]; 0,42	0,038				
	schwach-AR	50,6	30	< 0,05	0,982	0,046 [0,022; 0,067]; 0,59	0,037				
	stark-AR	76,1	34	< 0,05	0,963	0,062 [0,043; 0,080]; 0,14	0,046				
	partiell stark-AR	63,2	33	< 0,05	0,974	0,053 [0,033; 0,073]; 0,38	0,041				
	konfigural-AR vs. schwach-AR							2,0	0,74	0,002	-0,001
	schwach-AR vs. stark-AR							30,7	< 0,05	-0,019	0,009
	schwach-AR vs. partiell stark-AR							14,0	< 0,05	-0,008	0,004

AR weist darauf hin, dass eine Autokorrelation zwischen den entsprechenden Indikatorresiduen vorliegt.

... Tabelle geht auf der nächsten Seite weiter

Tab. E.5: Übersicht der Modellvergleiche zwischen den hierarchischen Modellen im Kontext der Messinvarianz.

Fak	Messinvarianz	χ^2	df	p	CFI	RMSEA [90 %CI]; p	SRMR	$\Delta\chi^2$	p	ΔCFI	ΔRMSEA
MU											
	konfigural	27,2	19	0,10	0,988	0,037 [0; 0,065]; 0,75	0,025				
	konfigural-AR	20,2	15	0,16	0,992	0,033 [0; 0,066]	0,021				
	schwach-AR	22,1	18	0,23	0,994	0,026 [0; 0,059]; 0,87	0,028				
	stark-AR	31,5	21	0,07	0,985	0,039 [0; 0,066];0,13	0,033				
	konfigural-AR vs. schwach-AR							1,9	0,60	0,002	-0,007
	schwach-AR vs. stark-AR							9,4	$< 0{,}05$	-0,009	0,013

AR weist darauf hin, dass eine Autokorrelation zwischen den entsprechenden Indikatorresiduen vorliegt.

. . . Tabelle geht auf der nächsten Seite weiter

Tab. E.5: Übersicht der Modellvergleiche zwischen den hierarchischen Modellen im Kontext der Messinvarianz.

Fak	Messinvarianz	χ^2	df	p	CFI	RMSEA [90 %CI]; p	SRMR	$\Delta\chi^2$	p	ΔCFI	ΔRMSEA
Evi											
	konfigural	57,9	33	< 0,05	0,973	0,048 [0,027; 0,068]; 0,53	0,038				
	konfigural-AR	29,9	28	0,37	0,998	0,014 [0; 0,046]; 0,97	0,028				
	schwach-AR	34,1	32	0,37	0,998	0,014 [0; 0,04]; 0,98	0,036				
	stark-AR	80,9	36	< 0,05	0,951	0,062 [0,044;0,080]	0,047				
	partiell stark-AR	50,4	35	< 0,05	0,983	0,037 [0,01; 0,058]; 0,83	0,045				
	konfigural-AR vs. schwach-AR							4,2	0,38	0,000	0,008
	schwach-AR vs. stark-AR							52,8	< 0,05	-0,037	0,048
	schwach-AR vs partiell stark-AR							17,5	< 0,05	-0,015	0,023

AR weist darauf hin, dass eine Autokorrelation zwischen den entsprechenden Indikatorresiduen vorliegt.

E.4 Strukturgleichungsmodelle

Die Ergebnisse in Abschnitt 6.2 basieren auf latenten Modellen. Nachfolgend werden die in den Tabellen E.4 und 6.10 genannten Modelle dargestellt.

E.4.1 Globales Messmodell

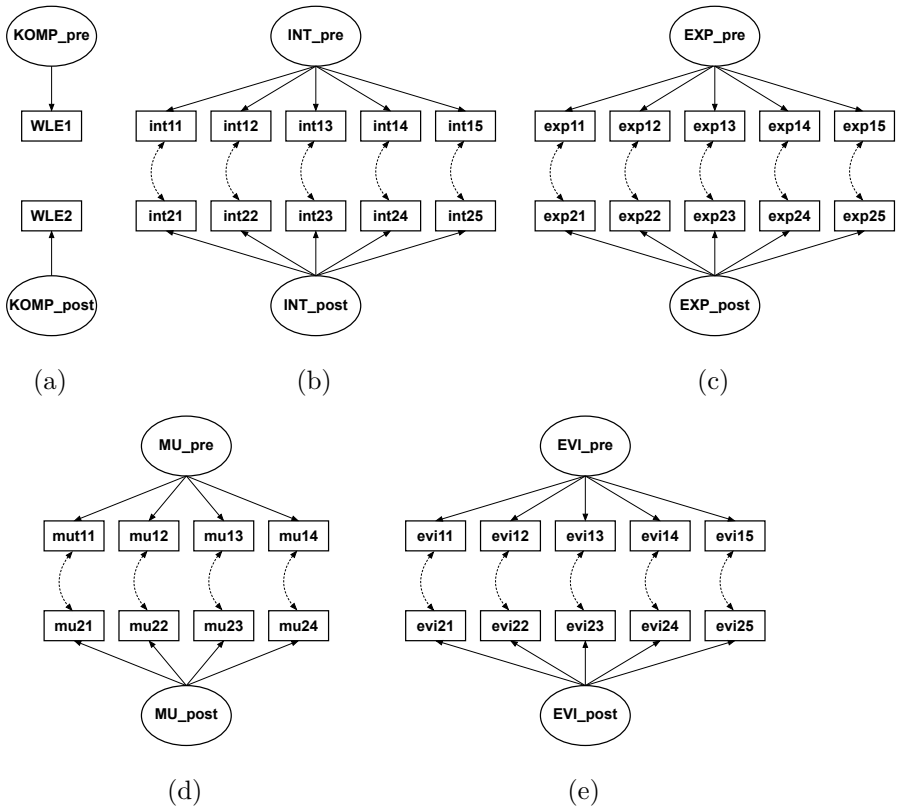
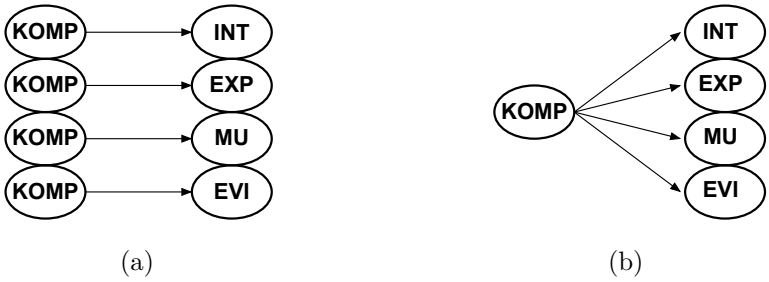


Abb. E.11: Darstellung der verwendeten Messmodelle (s. Tab. E.4): (a) Kompetenz , Zustimmung zu den Argumentkategorien (b) Intuition, (c) Expertenwissen, (d) Messunsicherheiten und (e) Daten als Evidenz.

E.4.2 Strukturgleichungsmodell zur Untersuchung von FF2



(a) (b)

Abb. E.12: Strukturgleichungsmodelle zur Untersuchung der Wirkung von Datenkompetenz auf die Zustimmung zu den Argumentkategorien (s. Tab. 6.10). (a) Einzelne Modellierung für jede Argumentkategorie, (b) gemeinsame Modellierung aller Argumentkategorien (`MoKompArg`).

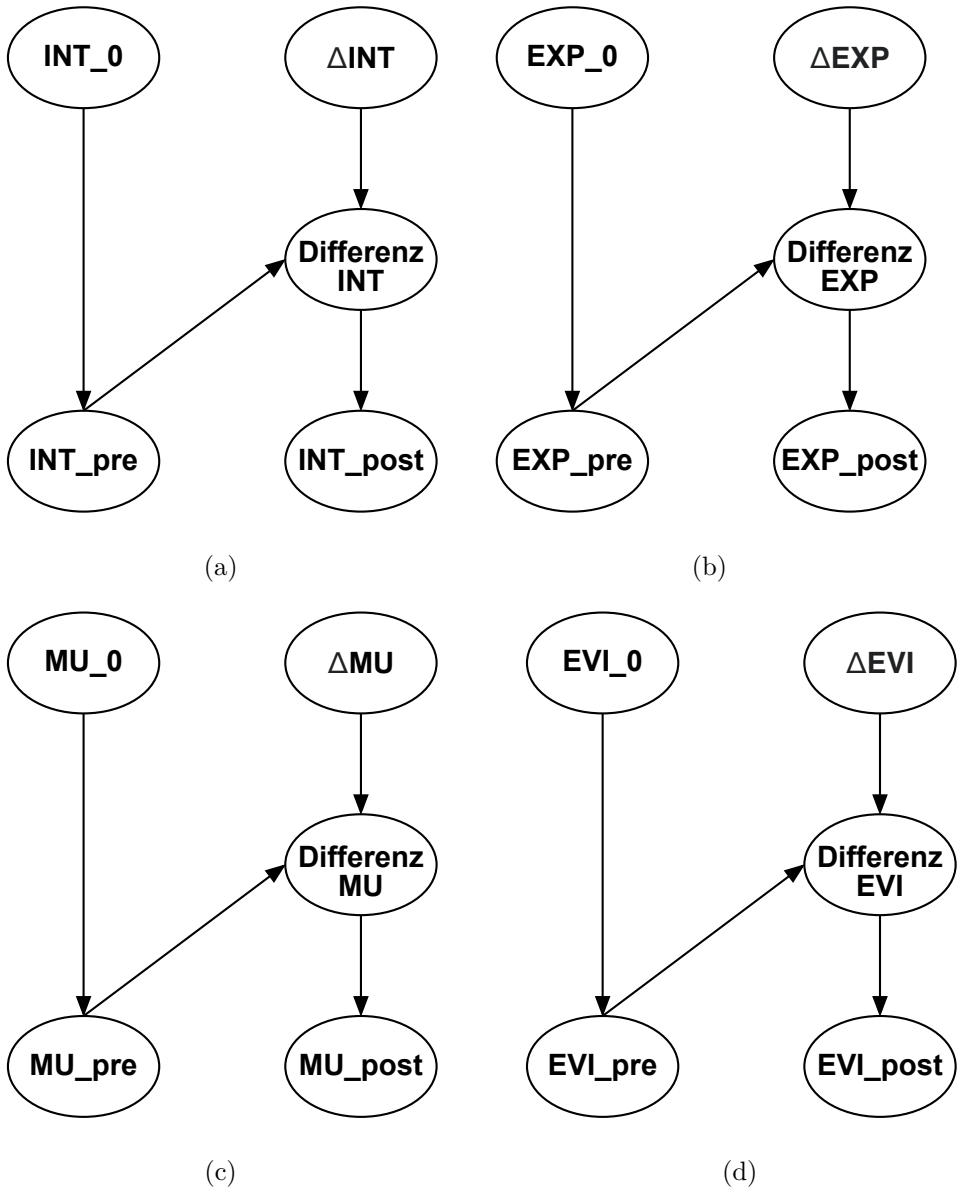


Abb. E.13: Modellierung des Latent-change-score-Modells (`MoKompArgLCSM`) (s. Tab. 6.10). (a) Inuition, (b) Expertenwissen, (c) Messunsicherheiten und (d) Daten als Evidenz.

E.4.3 Strukturgleichungsmodell zur Untersuchung von FF3

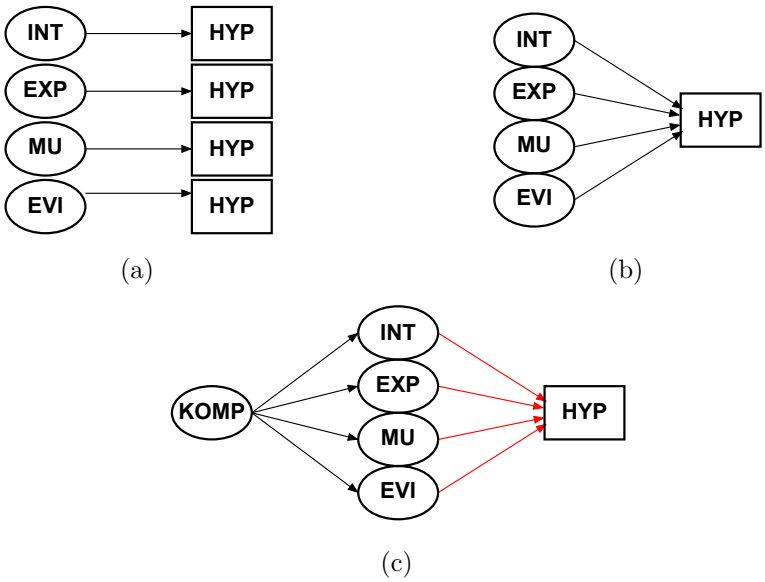


Abb. E.14: Strukturgleichungsmodelle zur Untersuchung der Wirkung der Zustimmung zu den Argumentkategorien auf die Richtigkeit der Hypothese (s. Tab. 6.10). (a) Einzelne Modellierung für jede Argumentkategorie, (b) gemeinsame Modellierung aller Argumentkategorien, (c) gemeinsame Modellierung aller Argumentkategorien unter Berücksichtigung der Datenkompetenz (`MoArgHyp`).

E.4.4 Strukturgleichungsmodell zur Untersuchung von FF4

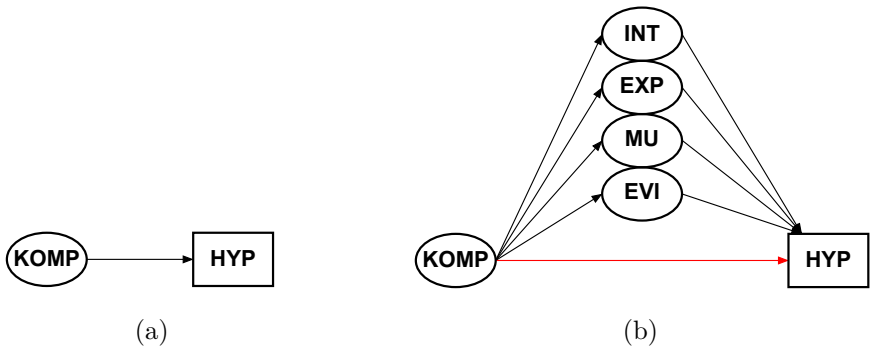


Abb. E.15: Strukturgleichungsmodelle zur Untersuchung der Datenkompetenz auf die Richtigkeit der Hypothese (s. Tab. 6.10). (a) Direkte Modellierung, (b) direkte und indirekte Modellierung unter Berücksichtigung der Zustimmung zu den Argumentkategorien (`MoKompArgHyp`).

Literatur

Abd-El-Khalick, F. (2008). *Modeling Science Classrooms after Scientific Laboratories: Sketching Some Affordances and Constraints Drawn from Examining Underlying Assumptions* [Pages: 80-85 Section: Teaching Scientific Inquiry]. Brill.

Abrahams, I. & Reiss, M. J. (2012). Practical work: Its effectiveness in primary and secondary schools in England. *Journal of Research in Science Teaching*, *49*(8), 1035–1055.

Acar, O. (2016). Examination of Science Learning Equity by Argumentation Instruction between Students Having Different Socio-Economic Status and Attending Different Achievement Level Schools. *Journal of Turkish Science Education.*

Aikenhead, G. S. (2005). Science-based occupations and the science curriculum: Concepts of evidence. *Science Education*, *89*(2), 242–275.

Allie, S. & Buffler, A. (1998). A course in tools and procedures for Physics I [MAG ID: 1968824985]. *American Journal of Physics*, *66*(7), 613–624.

Allie, S., Buffler, A., Campbell, B., Lubben, F., Evangelinos, D., Psillos, D. & Valassiades, O. (2003). Teaching Measurement in the Introductory Physics Laboratory. *The Physics Teacher*, *41*(7), 394–401.

Altenkirch, C. . (2023, 27.). *Anmeldung Fortbildung Messunsicherheiten.* https://www.uni-potsdam.de/de/physikdidaktik/mitarbeiter/phillip-schossau/anmeldung-fortbildung-messunsicherheiten

Ananiadou, K. & Claro, M. (2009). *21st Century Skills and Competences for New Millennium Learners in OECD Countries* (Techn. Ber.). OECD. Paris.

Anderson, J. C. & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach [Place: US Publisher: American Psychological Association]. *Psychological Bulletin*, *103*(3), 411–423.

Asterhan, C. S. C. (2018). Exploring enablers and inhibitors of productive peer argumentation: The role of individual achievement goals and of gender. *Contemporary Educational Psychology*, *54*, 66–78.

Asterhan, C. S. C. & Schwarz, B. B. (2016). Argumentation for Learning: Well-Trodden Paths and Unexplored Territories. *Educational Psychologist*, *51*(2), 164–187.

Aydeniz, M. & Ozdilek, Z. (2015). Assessing Pre-Service Science Teachers' Understanding of Scientific Argumentation: What Do They Know about Argumentation after Four Years of College Science? *Science Education International*, *26*(2), 217–239. Verfügbar 8. August 2019 unter https://eric.ed.gov/?id=EJ1064033

Baig, S. A. (2021). Bayesian Inference: Parameter Estimation for Inference in Small Samples. *Nicotine & Tobacco Research*.

Bano, M., Zowghi, D., Kearney, M., Schuck, S. & Aubusson, P. (2018). Mobile learning for science and mathematics school education: A systematic review of empirical evidence. *Computers & Education*, *121*, 30–58.

Barzilai, S. & Chinn, C. A. (2020). A review of educational responses to the "post-truth" condition: Four lenses on "post-truth" problems. *Educational Psychologist*, *55*(3), 107–119.

Bauer, J., Berthold, K., Hefter, M. H., Prenzel, M. & Renkl, A. (2017). Wie können Lehrkräfte und ihre Schülerinnen und Schüler lernen, fragile Evidenz zu verstehen und zu nutzen? *Psychologische Rundschau*, *68*(3), 188–192.

Becker, J. (2004). *Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT)* (Diss.) [Accepted: 2018-06-07T20:59:15Z]. Freie Universität Berlin. Accepted: 2018-06-07T20:59:15Z.

Bell, P. & Linn, M. C. (2000). Scientific arguments as learning artifacts: designing for learning from the web with KIE. *International Journal of Science Education*, *22*(8), 797–817.

Bennett, J. (2003). *Teaching and Learning Science: A Guide to Recent Research and Its Applications*. A&C Black.

Bentler, P. M. (1990). Comparative fit indexes in structural models [MAG ID: 2018201949]. *Psychological Bulletin*, *107*(2), 238–246.

Bentler, P. M. & Chou, C.-P. (1987). Practical Issues in Structural Equation Modeling. *Sociological Methods & Research*, *16*.

Berland, L. K. & Lee, V. R. (2010). Anomalous graph data and claim revision during argumentation, *2*, 2.

Berland, L. K. & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts: Learning Progression for Scientific Argumentation. *Science Education*, *94*(5), 765–793.

Berland, L. K. & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55.

Bilican, K. & Aydeniz, M. (2016). Appropriating Epistemic Norms of Science through Sustained Practice with Argumentation: Can It Happen? A Learning Progressions Perspective. *Journal of Education in Science, Environment and Health*, *2*(2), 125–133.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences, 2nd ed* [Pages: xvi, 340]. Mahwah, NJ, US, Lawrence Erlbaum Associates Publishers.

Boone, W. J., Staver, J. R. & Yale, M. S. (2014). The Rasch model and item response theory models: identical, similar, or unique? In *Rasch analysis in the human sciences* (S. 449–458). Springer.

Borowne, M. & Cudeck, R. (1993). Alternative ways of assessing model fit. In: KA Bollen and JS Long. *Testing Structural Equation Models*, 136–162.

Bos, W. (Hrsg.). (2003). *Erste Ergebnisse aus IGLU: Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich.* Münster München Berlin, Waxmann.

Briggs, S. R. & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, *54*(1), 106–148.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, Guilford Press.

Buffler, A., Allie, S. & Lubben, F. (2001). The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education*, *23*(11), 1137–1156.

Buffler, A., Allie, S. & Lubben, F. (2008). Teaching Measurement and Uncertainty the GUM Way. *The Physics Teacher*, *46*(9), 539–543.

Buffler, A., Lubben, F. & Ibrahim, B. (2009). The Relationship between Students' Views of the Nature of Science and their Views of the Nature of Scientific Measurement. *International Journal of Science Education*, *31*(9), 1137–1156.

Byrne, B. M. (2011). Structural Equation Modeling with Mplus, 431.

Byrne, B. M., Shavelson, R. J. & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466.

CERN. (2022, 16.). *LHCb discovers three new exotic particles.* Verfügbar 23. Juli 2022 unter https://home.cern/news/news/physics/lhcb-discovers-three-new-exotic-particles

Cetin, P. S., Dogan, N. & Kutluca, A. Y. (2014). The Quality of Pre-service Science Teachers' Argumentation: Influence of Content Knowledge. *Journal of Science Teacher Education*, *25*(3), 309–331.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A. & Borges, B. (2021). *shiny: Web Application Framework for R* [R package version 1.6.0]. R package version 1.6.0. https://CRAN.R-project.org/package=shiny

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Chen, H.-T., Wang, H.-H., Lu, Y.-Y., Lin, H.-s. & Hong, Z.-R. (2016). Using a modified argument-driven inquiry to promote elementary school students' engagement in learning science and argumentation. *International Journal of Science Education*, *38*(2), 170–191.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling*, *9*(2), 233–255.

Chinn, C. A. & Brewer, W. F. (1993). The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction. *Review of Educational Research*, *63*(1), 1–49.

Chinn, C. A. & Brewer, W. F. (2001). Models of Data: A Theory of How People Evaluate Data. *Cognition and Instruction*, *19*(3), 323–393.

Cho, E. (2016). Making Reliability Reliable: A Systematic Approach to Reliability Coefficients. *Organizational Research Methods*, *19*(4), 651–682.

Choi, A., Klein, V. & Hershberger, S. (2015). Success, difficulty, and instructional strategy to enact an argument-based inquiry approach: experiences of elementary teachers. *International Journal of Science and Mathematics Education*, *13*(5), 991–1011.

Christodoulou, A. & Osborne, J. (2014). The science classroom as a site of epistemic talk: A case study of a teacher's attempts to teach science based on argument. *Journal of Research in Science Teaching*, *51*(10), 1275–1300.

Clark, D., Nelson, B., Sengupta, P. & D'Angelo, C. (2009). Rethinking Science Learning Through Digital Games and Simulations: Genres, Examples, and Evidence.

Clark, D. & Sampson, V. (2008). Assessing dialogic argumentation in online environments to relate structure, grounds, and conceptual quality. *Journal of Research in Science Teaching*, *45*(3), 293–321.

Clark, D., Sampson, V., Weinberger, A. & Erkens, G. (2007). Analytic Frameworks for Assessing Dialogic Argumentation in Online Learning Environments. *Educational Psychology Review*, *19*(3), 343–374.

Clarke, S., Gerritsen, D., Grainger, R. & Ogan, A. (2016). Uncovering Teachers' Pedagogical Reasoning in Science Discussions. https://repository.isls.org//handle/1/362

Cohen, D. (1995). Argument is War...and War is Hell: Philosophy, Education, and Metaphors for Argumentation. *Informal Logic*, *17*(2).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed). New York, Routledge.

Comrey, A. L. & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed). Hillsdale, N.J, L. Erlbaum Associates.

Council, N. R. et al. (2008). Research on future skill demands: A workshop summary [Publisher: National Academies Press].

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Cunningham, J. D. & Bradley, K. D. (2010). Rasch Model to Measure Change, 33.

Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik*, *39*(2), 223–238. Verfügbar 15. Februar 2023 unter http://nbn-resolving.de/urn:nbn:de:0111-pedocs-111739

de Vries, E., Lund, K. & Baker, M. (2002). Computer-Mediated Epistemic Dialogue: Explanation and Argumentation as Vehicles for Understanding Scientific Notions. *Journal of the Learning Sciences*, *11*(1), 63–103.

DIN 1319-1:1995-01, Grundlagen der Meßtechnik- Teil 1: Grundbegriffe (Techn. Ber.). (1995). Beuth Verlag GmbH.

DIN 1319-2:2005-10, Grundlagen der Messtechnik - Teil 2: Begriffe für Messmittel (Techn. Ber.). (2005). Beuth Verlag GmbH.

DIN 1319-3:1996-05, Grundlagen der Meßtechnik - Teil 3: Auswertung von Messungen einer einzelnen Meßgröße, Meßunsicherheit (Techn. Ber.). (1996). Beuth Verlag GmbH.

DIN 1319-4:1999-02, Grundlagen der Meßtechnik - Teil 4: Auswertung von Messungen; Meßunsicherheit (Techn. Ber.). (1999). Beuth Verlag GmbH.

Donner, T. (2019). *Evaluation digitaler Lernumgebungen zum Vergleich von Daten mit Messunsicherheiten* (Masterarbeit).

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.* Berlin, Heidelberg, Springer Berlin Heidelberg.

Driver, R., Newton, P. & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287–312.

Duden. (2023). *Argumentation | Duden.* Verfügbar 5. Januar 2023 unter https://www.duden.de/node/8529/revision/538946

Duggan, S. & Gott, R. (2002). What sort of science education do we really need? *International Journal of Science Education*, *24*(7), 661–679.

Duschl, R. A. (2008). Science Education in Three-Part Harmony: Balancing Conceptual, Epistemic, and Social Learning Goals. *Review of Research in Education*, *32*(1), 268–291.

Duschl, R. A. & Osborne, J. (2002). Supporting and Promoting Argumentation Discourse in Science Education. *Studies in Science Education*, *38*(1), 39–72.

Duschl, R. A., Schweingruber, H. A. & Shouse, A. W. (2007). Taking science to school: Learning and teaching science in grades K-8. *Eurasia Journal of Mathematics, Science & Technology Education*, *3*(2), 163–166.

Ebner, B. & Henze, N. (2020). Tests for multivariate normality—a critical review with emphasis on weighted L2-statistics. *TEST*, *29*(4), 845–892.

Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden: mit Online-Materialien* (1. Auflage). Weinheim Basel, Beltz.

Erduran, S. & Jiménez-Aleixandre, M. P. (Hrsg.). (2008). *Argumentation in science education: perspectives from classroom-based research.* Dordrecht, Springer.

Evangelinos, D., Psillos, D. & Valassiades, O. (2002). An Investigation of Teaching and Learning about Measurement Data and their Treatment in the Introductory Physics Laboratory, 179–190.

Feinstein, N. (2011). Salvaging science literacy. *Science Education*, *95*(1), 168–185.

Feinstein, N., Allen, S. & Jenkins, E. (2013). Outside the Pipeline: Reimagining Science Education for Nonscientists. *Science*, *340*(6130), 314–317.

Field, A., Miles, J. & Field, Z. (2014). *Discovering statistics using R* (2nd ed). Los Angeles, CA, USA, Sage.

Finkelstein, N. D., Adams, W. K., Keller, C. J., Kohl, P. B., Perkins, K. K., Podolefsky, N. S., Reid, S. & LeMaster, R. (2005). When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment. *Physical Review Special Topics - Physics Education Research*, *1*(1), 010103.

Finn, L., Kuhn, L., Whitcomb, J., Bruozas, M. & Reiser, B. (2006). Where have all the creatures gone. *IQWST: Investigating and questioning our world through science and technology. Evanston, IL: Northwestern University.*

Finney, S. J. & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In *Structural equation modeling: A second course, 2nd ed* (S. 439–492). Charlotte, NC, US, IAP Information Age Publishing.

Fishman, E. J., Borko, H., Osborne, J., Gomez, F., Rafanelli, S., Reigh, E., Tseng, A., Million, S. & Berson, E. (2017). A Practice-Based Professional Development Program to Support Scientific Argumentation From Evidence in the Elementary Classroom. *Journal of Science Teacher Education*, *28*(3), 222–249.

Fleischer, T., Deibl, I., Strahl, A., Moser, S., Maier, S. & Zumbach, J. (2020). EXBOX-Digital - Praxisorientiertes Unterrichtskonzept zum Einsatz digitaler Medien im Chemie- und Physikunterricht, *8*, 211–224.

Ford, M. J. (2012). A Dialogic Account of Sense-Making in Scientific Argumentation and Reasoning. *Cognition and Instruction*, *30*(3), 207–245.

for Standardization, I. O. (2004). *Guide to the expression of uncertainty in measurement (GUM)-Supplement 1: Numerical methods for the propagation of distributions* (Bd. ISO draft guide DGUIDE99998). Geneva, International Organization for Standardization.

Geis, G. L. (1966). Variety and programed instruction or what can't be programed? *AV Communication Review*, *14*(1), 109–116.

Geiser, C. (2010). *Datenanalyse mit Mplus: eine anwendungsorientierte Einführung* (1. Auflage). Wiesbaden, VS Verlag.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2015). *Bayesian Data Analysis* (3. Aufl.). New York, Chapman; Hall/CRC.

Ghisletta, P. & McArdle, J. J. (2012). Latent Curve Models and Latent Change Score Models Estimated in R. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(4), 651–682.

Gibson, J. P. & Mourad, T. (2018). The growing importance of data literacy in life science education. *American Journal of Botany*, *105*(12), 1953–1956.

Gott, R. & Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. *International Journal of Science Education*, *18*(7), 791–806.

Gott, R. & Duggan, S. (2007). A framework for practical work in science and scientific literacy through argumentation. *Research in Science & Technological Education*, *25*(3), 271–291.

Gott, R. & Duggan, S. (1995). *Investigative work in the science curriculum*. Buckingham ; Philadelphia, Open University Press.

Gott, R. & Duggan, S. (2003). *Understanding and using scientific evidence: How to critically evaluate data*. Sage.

Gould, R. (2017). Data Literacy is Statistical Literacy [Number: 1]. *Statistics Education Research Journal*, *16*(1), 22–25.

Grabe, M. (2001). Estimation of measurement uncertainties – an alternative to the ISO Guide. *Metrologia*, *38*, 97–106.

Greenwald, A. G. (1968). Cognitive Learning, Cognitive Response to Persuasion, and Attitude Change. In *Psychological Foundations of Attitudes* (S. 147–170). Elsevier.

Gromadecki, U. (2008). Argumente in physikalischen Kontexten, 346.

Grooms, J., Sampson, V. & Enderle, P. (2018). How concept familiarity and experience with scientific argumentation are related to the way groups participate in an episode of argumentation. *Journal of Research in Science Teaching*, *55*(9), 1264–1286.

Hamacher, J., Erkelenz, J. & Heinke, H. (2015). Messunsicherheiten mit Hilfe von Lehrvideos verstehen.

Hamilton, E. R., Rosenberg, J. M. & Akcaoglu, M. (2016). The Substitution Augmentation Modification Redefinition (SAMR) Model: a Critical Review and Suggestions for its Use. *TechTrends*, *60*(5), 433–441.

Hammann, M. (2004). Kompetenzentwicklungsmodelle Merkmale und ihre Bedeutung-dargestellt anhand von Kompetenzen beim Experimentieren [Publisher: DUEMMLER]. *Mathematische und Naturwissenschaftliche Unterricht*, 196–203.

Hammann, M., Phan, T. T. H., Ehmer, M. & Grimm, T. (2008). Assessing pupils' skills in experimentation. *Journal of Biological Education*, *42*(2), 66–72.

Hanekamp, G. (2014). Zahlen und Fakten: Allensbach-Studie 2013 der Deutsche Telekom Stiftung. In *Digitale Medien im naturwissenschaftlichem Unterricht* (S. 21–28).

Haßler, B., Major, L. & Hennessy, S. (2016). Tablet use in schools: a critical review of the evidence for learning outcomes. *Journal of Computer Assisted Learning*, *32*(2), 139–156.

Heim, A. B., Walsh, C., Esparza, D., Smith, M. K. & Holmes, N. G. (2022). What influences students' abilities to critically evaluate scientific investigations? *PLOS ONE*, *17*(8).

Heinicke, S., Glomski, J., Priemer, B. & Rieß, F. (2010). Aus Fehlern wird man klug-Über die Relevanz eines adäquaten Verständnisses von"Messfehlernïm Physikunterricht. *Praxis der Naturwissenschaften-Physik in der Schule*, *59*(5), 26–33.

Hellwig, J. (2012). Messunsicherheiten verstehen : Entwicklung eines normativen Sachstrukturmodells am Beispiel des Unterrichtsfaches Physik, 343.

Hemmert, G. A. J., Schons, L. M., Wieseke, J. & Schimmelpfennig, H. (2018). Log-likelihood-based Pseudo-R2 in Logistic Regression: Deriving Sample-sensitive Benchmarks [Publisher: SAGE Publications Inc]. *Sociological Methods & Research*, *47*(3), 507–531.

Hendershot, C. H. (1971). *Programmed Learning: A Bibliography of Programs and Presentation Devices. Fourth Edition with Supplements Including the 1971 Release* (Techn. Ber.).

Hiebert, J. & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. *Conceptual and procedural knowledge: The case of mathematics*, *2*, 1–27.

Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I. & Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education*, *153*, 103897.

Hilton, A., Hilton, G., Dole, S. & Goos, M. (2016). Promoting middle school students' proportional reasoning skills through an ongoing professional development programme for teachers. *Educational Studies in Mathematics*, *92*(2), 193–219.

Hintikka, J. (1989). The Role of Logic in Argumentation. *The Monist*, *72*(1), 3–24.

Hofstein, A. (2017). The Role of Laboratory in Science Teaching and Learning [MAG ID: 2562318930], 357–368.

Hofstein, A. & Lunetta, V. N. (1982). The Role of the Laboratory in Science Teaching: Neglected Aspects of Research. *Review of Educational Research*, *52*(2), 201–217.

Hofstein, A. & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, *88*(1), 28–54.

Hsiao, Y.-Y. & Lai, M. H. C. (2018). The Impact of Partial Measurement Invariance on Testing Moderation for Single and Multi-Level Data. *Frontiers in Psychology*, *9*, 740.

Hu, L.-t. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis : Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Humes, V. (2017). *Eighth Grade Language Art and Mathematics* (Diss.). Youngtown State University.

Iordanou, K. & Constantinou, C. P. (2015). Supporting Use of Evidence in Argumentation Through Practice in Argumentation and Reflection in the Context of SOCRATES Learning Environment. *Science Education*, *99*(2), 282–311.

Irribarra, D. T. & Freund, R. (2014). *Wright Map: IRT item-person map with ConQuest integration*. https://github.com/david-ti/wrightmap

ISO. (2022). *ISO - 17.020 - Metrology and measurement in general*. Verfügbar 22. Juli 2022 unter https://www.iso.org/ics/17.020/x/

ISO/TMBG. (2008). *Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)* (Standard). International Organization for Standardization. Geneva, CH.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. John Wiley & Sons.

Jak, S., Jorgensen, T. D., Verdam, M. G. E., Oort, F. J. & Elffers, L. (2021). Analytical power calculations for structural equation modeling: A tutorial and Shiny app. *Behavior Research Methods*, *53*(4), 1385–1406.

JCGM. (2008). *100: 2008—Evaluation of measurement data—Guide to the expression of uncertainty in measurement (GUM)* (Techn. Ber.).

JCGM. (2012). *200: 2012—International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM)* (Techn. Ber.).

JCGM. (2020). *GUM-6:2020—Evaluation of measurement data—Guide to the expression of uncertainty in measurement (GUM) - Part 6: Developing and using measurement models* (Techn. Ber.).

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, *186*(1007), 453–461.

Jeffreys, H. (1998). *Theory of probability* (3rd ed). Oxford: New York, Clarendon Press ; Oxford University Press.

Jiménez-Aleixandre, M. P. & Crujeiras, B. (2017). *Epistemic Practices and Scientific Practices in Science Education*. Brill.

Jobst, L. J., Bader, M. & Moshagen, M. (2021). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods*.

Jones, M. & Sugden, R. (2001). Positive confirmation bias in the acquisition of information, 41.

Jöreskog, K. G. (2001). LISREL 8.50. Student edition. Mooresville, Ind. : Scientific Software. https://search.library.wisc.edu/catalog/999924848502121

Kanari, Z. & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748–769.

Kang, H., Scharmann, L. C., Kang, S. & Noh, T. (2010). Cognitive conflict and situational interest as factors influencing conceptual change, 24.

Kardas, E. (2019). *Entwicklung digitaler Lernumgebungen zum Vergleich von Daten mit Messunsicherheiten* (Masterarbeit). Berlin.

Kardas, E. & Ludwig, T. (2021). Den Umgang mit Messunsicherheiten lernen - Digitale Apps für ein wichtiges Thema. *plusLucis*, (4), 24–32.

Kass, R. E. (1993). Bayes Factors in Practice. *The Statistician*, *42*(5), 551.

Kass, R. E. & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben*. Berlin, Logos Berlin.

Kelly, G. J. (2008). Kelly, G. J. (2008). Inquiry, Activity, and Epistemic Practice. In R. Duschl & R. Grandy (Eds.) Teaching Scientific Inquiry: Recommendations for Research and Implementation (pp. 99-117; 288-291). Rotterdam: Sense Publishers.

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

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A.-L., de Mooij, S. M., Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., Lindenberger, U. & Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, *33*, 99–117.

Kim, H. & Song, J. (2006). The Features of Peer Argumentation in Middle School Students' Scientific Inquiry. *Research in Science Education*, *36*(3), 211–233.

Kind, P. M. (2013). Establishing Assessment Scales Using a Novel Disciplinary Rationale for Scientific Reasoning. *Journal of Research in Science Teaching*, *50*(5), 530–560.

Kind, P. M., Kind, V., Hofstein, A. & Wilson, J. (2011). Peer Argumentation in the School Science Laboratory—Exploring effects of task features. *International Journal of Science Education*, *33*(18), 2527–2558.

Kind, P. M. & Osborne, J. (2017). Styles of Scientific Reasoning: A Cultural Rationale for Science Education? *Science Education*, *101*(1), 8–31.

Kircher, E., Girwidz, R. & Häußler, P. (Hrsg.). (2020). *Physikdidaktik | Grundlagen*. Berlin, Heidelberg, Springer Berlin Heidelberg.

Kirkup, L. & Frenkel, R. B. (2006). An introduction to uncertainty in measurement using the GUM (Guide to the expression of Unvertainty in Measurement).

Kirschner, P. A., Sweller, J. & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist*, *41*(2), 75–86.

Klahr, D. & Dunbar, K. (1988). Dual Space Search During Scientific Reasoning. *Cognitive Science*, *12*(1), 1–48.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. Guilford Publications.

Klinger, W. & Wörlen, F. (1991). Messwertstreuung dennoch exakte Gesetze?, 26–30.

Klopack, E. T. & Wickrama, K. ((2020). Modeling Latent Change Score Analysis and Extensions in Mplus: A Practical Guide for Researchers. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 97–110.

KMK. (2003). Bildungsstandards im Fach Mathematik für Mittleren Schulabschluss.

KMK. (2004). Bildungsstandards im Fach Physik für den Mittleren Schulabschluss.

KMK. (2012). Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife.

KMK. (2020). Bildungsstandards im Fach Physik für die Allgemeine Hochschulreife.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing Research*, *39*(2), 121–123.

Kok, K. (2022). *Certain about uncertainty* (PhD Thesis). Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät.

Kok, K., Priemer, B., Musold, W. & Masnick, A. (2019). Students' conclusions from measurement data: The more decimal places, the better? *Physical Review Physics Education Research*, *15*(1).

Kopperschmidt, J. (1989). *Methodik der Argumentationsanalyse*. Stuttgart-Bad Cannstatt, Frommann-Holzboog.

Koslowski, B. (1996). *Theory and Evidence: The Development of Scientific Reasoning*. Cambridge, Massachusetts, MIT Press.

Kramer, M., Förtsch, C., Aufleger, M. & Neuhaus, B. J. (2019). Der Einsatz digitaler Medien im gymnasialen Biologieunterricht: Eine deskriptive Auswertung einer quantitativen Videostudie. *Zeitschrift für Didaktik der Naturwissenschaften*, *25*(1), 131–160.

Krouwer, J. S. (2003). Critique of the Guide to the Expression of Uncertainty in Measurement Method of Estimating and Reporting Uncertainty in Diagnostic Assays. *Clinical Chemistry*, *49*(11), 1818–1821.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*(7), 293–300.

Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280.

Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291.

Kruschke, J. K. & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206.

Krystek, M. (2020). *Berechnung der Messunsicherheit: Grundlagen und Anleitung für die praktische Anwendung*. Beuth Verlag.

Kuhn, D. (1991). *The Skills of Argument*. Cambridge, Cambridge University Press.

Kuhn, D. (1992). Thinking as Argument. *Harvard Educational Review*, 25.

Kuhn, D., Cheney, R. & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development*, *15*(3), 309–328.

Kuhn, D. & Lerman, D. (2021). Yes but: developing a critical stance toward evidence. *International Journal of Science Education*, *43*(7), 1036–1053.

Kuhn, D. & Udell, W. (2003). The Development of Argument Skills. *Child Development*, *74*(5), 1245–1260.

Kung, R. L. (2005). Teaching the concepts of measurement: An example of a concept-based laboratory course. *American Journal of Physics*, *73*(8), 771–777.

Kung, R. L. & Linder, C. (2012). University students' ideas about data processing and data comparison in a physics laboratory course. *Nordic Studies in Science Education*, *4*(2), 40–53.

Laube, H. (1966). Programmierte Instruktion in der US-amerikanischen Wirtschaft. *Mensch und Arbeit*, *18*(3), 69–73. https://www.jstor.org/stable/24376573

Lauer, P. & Lenz, L. (2021). *Pilotierung von Testitems zur Erfassung von Kompetenzen im Umgang mit Daten und Messunsicherheiten und Evaluation von digtialen Lernumgebungen zum Thema direktes Messen, indirektes Messen, graphische Regression und Signifikanz* (Masterarbeit).

Lehesvuori, S., Viiri, J., Rasku-Puttonen, H., Moate, J. & Helaakoski, J. (2013). Visualizing communication structures in science classrooms: Tracing cumulativity in teacher-led whole class discussions. *Journal of Research in Science Teaching*, *50*(8), 912–939.

Lemke, J. L. (1990). *Talking Science: Language, Learning, and Values.* New Jersey, Ablex Publishing Corporation.

Lira, I. (2016). The GUM revision: the Bayesian view toward the expression of measurement uncertainty. *European Journal of Physics*, *37*(2), 025803.

LISUM. (2015). Teil C Physik Jahrgangsstufe 7-10 (J. u. S. d. L. B. Senatsverwaltung für Bildung Jugend und Wissenschaft Berlin Ministerium für Bildung, Hrsg.).

Liu, S. & Roehrig, G. (2017). Exploring Science Teachers' Argumentation and Personal Epistemology About Global Climate Change. *Research in Science Education*, *49*(1), 173–189.

Lorenz, R., Bos, W., Endberg, M., Eickelmann, B., Grafe, S. & Vahrenhold, J. (Hrsg.). (2017). *Schule digital – der Länderindikator 2017: schulische Medienbildung in der Sekundarstufe I mit besonderem Fokus auf MINT-Fächer im Bundesländervergleich und Trends von 2015 bis 2017.* Münster New York, Waxmann.

Lubben, F., Campbell, B., Buffler, A. & Allie, S. (2001). Point and Set Reasoning in Practical Science Measurement by Entering University Freshmen. *Science Education*, *85*(4), 311–327.

Lubben, F. & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education*, *18*(8), 955–968.

Ludwig, T. (2017). Argumentieren beim Experimentieren in der Physik - Die Bedeutung personaler und situationaler Faktoren.

Ludwig, T., Priemer, B. & Lewalter, D. (2019). Assessing Secondary School Students' Justifications for Supporting or Rejecting a Scientific Hypothesis in the Physics Lab. *Research in Science Education*, *51*(3), 819–844.

Lunetta, V. N., Hofstein, A. & Clough, M. P. (2007). *Learning and teaching in the school science laboratory: An analysis of research, theory and practice.* Lawrence Erlbaum Associates: Mahwah, NJ, USA.

Lyu, K., Zhang, X. & Church, J. (2020). Impacts of Model Mean Biases and Climate Sensitivity on Regional Dynamic Sea Level Projections in the

CMIP5 and CMIP6 models, *2020*, OS007–06. https://ui.adsabs.harvard.edu/abs/2020AGUFMOS007..06L

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), 130–149.

Maiseyenka, V., Schecker, H. & Nawrath, D. (2013). Kompetenzorientierung des naturwissenschaftlichen Unterrichts - Symbiotische Kooperation bei der Entwicklung eines Modells experimenteller Kompetenz [Number: 12]. *PhyDid A - Physik und Didaktik in Schule und Hochschule*, *1*(12), 1–17.

Manz, E. (2015). Representing Student Argumentation as Functionally Emergent From Scientific Activity. *Review of Educational Research*, *85*(4), 553–590.

Manz, E. & Renga, I. P. (2016). Understanding how teachers guide evidence construction conversations. *Science Education*, *101*(4), 584–615.

Marsh, E. J. & Yang, B. W. (2017). A call to think broadly about information literacy. *Journal of Applied Research in Memory and Cognition*, *6*, 401–404.

Masnick, A. M. & Klahr, D. (2003). Error Matters: An Initial Exploration of Elementary School Children's Understanding of Experimental Error. *Journal of Cognition and Development*, *4*(1), 67–98.

Masnick, A. M. & Morris, B. J. (2022). A Model of Scientific Data Reasoning. *Education Sciences*, *12*.

McComas, W. F. (2004). Keys to teaching the nature of science. *The science teacher*, *71*(9), 24.

McDonald, S. P. & Kelly, G. J. (2012). Beyond Argumentation: Sense-Making Discourse in the Science Classroom. In M. S. Khine (Hrsg.), *Perspectives on Scientific Argumentation* (S. 265–281). Dordrecht, Springer Netherlands.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2. Aufl.). New York, Chapman; Hall/CRC.

McKelvey, R. D. & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120.

McNeill, K. L., González-Howard, M., Katsh-Singer, R. & Loper, S. (2016). Pedagogical content knowledge of argumentation: Using classroom contexts to assess high-quality PCK rather than pseudoargumentation. *Journal of Research in Science Teaching*, *53*(2), 261–290.

McNeill, K. L. & Krajcik, J. S. (2011). *Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing*. Pearson.

Meade, A. W., Johnson, E. C. & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Millar, R. (2004). The role of practical work in the teaching and learning of science, 25.

Millar, R., Lubben, F., Gott, R. & Duggan, S. (1994). Investigating in the school science laboratory: conceptual and procedural knowledge and their influence on performance. *Research Papers in Education*, *9*(2), 207–248.

Millar, R. & Osborne, J. (1998). *Beyond 2000: science education for the future.* London, King's College London.

Millsap, R. E. & Kwok, O.-M. (2004). Evaluating the Impact of Partial Factorial Invariance on Selection in Two Populations. *Psychological Methods*, *9*(1), 93–115.

Ministerium für Kultus, Jugend und Sport Baden-Württemberg. (2016). *Gemeinsamer Bildungsplan der Sekundrstufe 1 Physik.* http://www.bildungsplaene-bw.de/site/bildungsplan/get/documents/lsbw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_SEK1_PH.pdf

MINT Zukunft e.V. (2022). *MINT-freundliche Schulen in Baden-Württemberg.* Verfügbar 12. Februar 2023 unter https://mintzukunftschaffen.de/baden-wuerttemberg

Möhrke, P. & Runge, B.-U. (2020). *Arbeiten mit Messdaten: Eine praktische Kurzeinführung nach GUM* (1st ed). Berlin, Springer Spektrum.

Moreno, R. & Mayer, R. E. (2005). Role of Guidance, Reflection, and Interactivity in an Agent-Based Multimedia Game. *Journal of Educational Psychology*, *97*(1), 117–128.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D. & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.

Morey, R. D. & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419.

Morey, R. D. & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs* [R package version 0.9.12-4.2]. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

Munier, V., Merle, H. & Brehelin, D. (2013). Teaching Scientific Measurement and Uncertainty in Elementary School. *International Journal of Science Education*, *35*(16), 2752–2783.

Murphy, P. K. & Mason, L. (2006). Changing Knowledge and Beliefs. In *Handbook of educational psychology* (S. 305–324). Mahwah, NJ, US, Lawrence Erlbaum Associates Publishers.

Murphy, P. & Gott, R. (1984). *Science: Assessment Framework Age 13 & 15.* Assessment of Performance Unit.

Muthén, L. & Muthén, B. (2017). Mplus user's guide (eight edition). *Los Angeles, CA.*

Nagel, C. (2021). Sicher ist sicher! Fachliche Klärung für die didaktische Rekonstruktion von Messunsicherheiten im Unterricht. *plusLucis*, (4), 7–11.

Nagel, C., Lux, B. & Steindl, S. (2021). Die Thematisierung von Messunsicherheiten im Physikunterricht - Eine Umfrage. *plusLucis*, (4), 4–6.

Nawrath, D., Maiseyenka, V. & Schecker, H. (2011). Experimentelle Kompetenz–Ein Modell für die Unterrichtspraxis. *Praxis der Naturwissenschaften–Physik in der Schule*, *60*(6), 42–49.

Neumann, K., Viering, T., Boone, W. J. & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching*, *50*(2), 162–188.

Neumann, S. (2021). Bin ich wirklich schneller als mein Sitznachbar? *plusLucis*, (4), 36–38.

Newton, P., Driver, R. & Osborne, J. (1999). The place of argumentation in the pedagogy of school science. *International Journal of Science Education*, *21*(5), 553–576.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220.

Norris, S., Philips, L. & Osborne, J. (2007). Scientific inquiry: The place of interpretation and argumentation. *Science as inquiry in the secondary setting*, 87–98.

Norris, S. P., Phillips, L. M., Smith, M. L., Guilbert, S. M., Stange, D. M., Baker, J. J. & Weber, A. C. (2008). Learning to read scientific text: Do elementary school commercial reading programs help?: Learning to Read Scientific Text. *Science Education*, *92*(5), 765–798.

NRC. (2013). *Next Generation Science Standards: For States, By States* (N. R. Council, Hrsg.).

Nussbaum, M. E. & Bendixen, L. D. (2003). Approaching and avoiding arguments: The role of epistemological beliefs, need for cognition, and extraverted personality traits. *Contemporary Educational Psychology*, *28*(4), 573–595.

OECD. (2015). PISA 2015 science framework. In *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving.*

OECD. (2018). *The future of education and skills.*

Ohlsson, S. (1996). Learning to do and learning to understand : a lesson and a challenge for cognitive modeling. In P. Reiman & H. Spade (Hrsg.),

Learning in humans and machines towards an interdisciplinary learning science (S. 37–62). Pergamon.

Osborne, J. (2010). Arguing to Learn in Science: The Role of Collaborative, Critical Discourse. *Science*, *328*(5977), 463–466.

Osborne, J. (2012). The Role of Argument: Learning How to Learn in School Science. In K. Tobin, C. J. McRobbie & B. J. Fraser (Hrsg.), *Second International Handbook of Science Education* (S. 933–949). Dordrecht, Springer Netherlands. Verfügbar 8. August 2019 unter http://link.springer.com/10.1007/978-1-4020-9041-7_62

Osborne, J., Erduran, S. & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994–1020.

Osborne, J. & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, *95*(4), 627–638.

Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C. & Richardson, K. (2013). Learning to argue: A study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching*, *50*(3), 315–347.

Oser, F. & Patry, J.-L. (1994). Sichtstruktur und Basismodelle des Unterrichts: Über den Zusammenhang von Lehren und Lernen unter dem Gesichtspunkt psychologischer Lernverläufe. In R. Olechowski & B. Rollett (Hrsg.), *Theorie und Praxis* (S. 138–146). Lang.

Ozdem Yilmaz, Y., ertepınar, H., Cakiroglu, J. & Erduran, S. (2011). The Nature of Pre-service Science Teachers' Argumentation in Inquiry-oriented Laboratory Context. *International Journal of Science Education*, *2011*.

Patterson, E. W. (2001). Structuring the composition process in scientific writing. *International Journal of Science Education*, *23*(1), 1–16.

Pehmer, A.-K., Gröschner, A. & Seidel, T. (2015). How teacher professional development regarding classroom dialogue affects students' higher-order learning. *Teaching and Teacher Education*, *47*, 108–119.

Pentz, M. A. & Chou, C.-P. (1994). Measurement Invariance in Longitudinal Clinical Research Assuming Change from Development and Intervention, 13.

Petrescu, M. (2013). Marketing research using single-item indicators in structural equation models. *Journal of Marketing Analytics*, *1*(2), 99–117.

Petty, R. E. & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. In *Advances in Experimental Social Psychology* (S. 123–205). Elsevier.

Pfeiler, S. (2019). Die Evaluation von Daten aus erster und zweiter Hand im naturwissenschaftlichen Unterricht.

Pfost, M., Neuenhaus, N., Kuntner, P., Becker, S., Goppert, S. & Werner, A. (2020). Selbstständiges Lernen an der Hochschule: Diskussion eines niedrigschwelligen Förderansatzes. Pfost, M., Neuenhaus, N., Kuntner, P., Becker, S., Goppert, S. & Werner, A. (2020)., 83–101.

Pillay, S., Buffler, A., Lubben, F. & Allie, S. (2008). Effectiveness of a GUM-compliant course for teaching measurement in the introductory physics laboratory. *European Journal of Physics*, *29*, 647–659.

Pohl, R., Gilman, R., Miller, G. A. & Pachucki, K. (2013). Muonic Hydrogen and the Proton Radius Puzzle. *Annual Review of Nuclear and Particle Science*, *63*(1), 175–204.

Pollard, B., Werth, A., Hobbs, R. & Lewandowski, H. J. (2020). Impact of a course transformation on students' reasoning about measurement uncertainty [Publisher: American Physical Society]. *Physical Review Physics Education Research*, *16*(2), 020160.

Pols, F. (2019). Students' ability to analyse empirical data in practical work. *Journal of Physics: Conference Series*, *1287*.

Pols, F., Dekkers, P. & de VRIES, M. (2021). Data-analysis in practical work, what do students know?, 2.

Puentedura, R. R. (2006). Transformation, Technology, and Education, 8.

Putnick, D. L. & Bornstein, M. H. (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental review : DR*, *41*, 71–90.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rasch, G. (1966). An Item Analysis Which Takes Individual Differences into Account. *British Journal of Mathematical and Statistical Psychology*, *19*(1), 49–57.

Ratcliffe, M. (1996). Adolescent decision-making, by individuals and groups, about science-related societal issues. *Research in science education in Europe*, 126–140.

Rickey, D. & Stacy, A. M. (2000). The Role of Metacognition in Learning Chemistry. *Journal of Chemical Education*, *77*(7), 915.

Robitzsch, A., Kiefer, T. & Wu, M. (2021). *TAM: Test Analysis Modules* [R package version 3.7-16]. R package version 3.7-16. https://CRAN.R-project.org/package=TAM

Rocci, A. (2005). Connective predicates in monologic and dialogic argumentation. *Studies in communication sciences*, *5*(1), 97–118.

Romrell, D., Kidder, L. C. & Wood, E. (2014). The SAMR Model as a Framework for Evaluating mLearning. *Online Learning*, *18*(2).

Ropohl, M. & Scheuermann, H. (2018). Welche Rückmeldungen wirken am besten?: Ergebnisse einer empirischen Untersuchung von Rückmelde-

formen beim Planen von Experimenten. *Zeitschrift für Didaktik der Naturwissenschaften*, *24*(1), 151–165.

Rouder, J. N., Haaf, J. M. & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113.

Rutten, N., van Joolingen, W. R. & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education*, *58*(1), 136–153.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *Journal of Research in Science Teaching*, *41*(5), 513–536.

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

Sampson, V. & Clark, D. (2009). The impact of collaboration on the outcomes of scientific argumentation. *Science Education*, *93*(3), 448–484.

Sampson, V., Hutner, T. L., FitzPatrick, D., LaMee, A. & Grooms, J. (2017). *Argument-driven inquiry in physics: volume 1: mechanics lab investigations for grades 9-12*. NSTA Press.

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

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, *21*, 149–160.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *MPR-online*, *8*(2), 23–74.

Schield, M. (2004). Information Literacy, Statistical Literacy and Data Literacy. *IASSIST quarterly / International Association for Social Science Information Service and Technology*, *28*, 7–14.

Schiepe-Tiska, A., Schöps, K., Rönnebeck, S., Köller, O. & Prenzel, M. (2012). 6 Naturwissenschaftliche Kompetenz in PISA 2012: Ergebnisse und Herausforderungen.

Schmidt, H. (2003). Warum GUM? - Kritische Anmerkungen zur Normdefinition der 'Messunsicherheit' und zu verzerrten 'Elementarfehlermodellen'. *zfv – Zeitschrift für Geodäsie, Geoinformation und Landmanagement*, *128*(5), 303–312.

Schmitt, N. & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222.

Schnotz, W. (2008). Why Multimedia Learning is not Always Helpful. In J.-F. Rouet, R. Lowe & W. Schnotz (Hrsg.), *Understanding Multimedia Docu-*

ments (S. 17–41). Boston, MA, Springer US. http://link.springer.com/10.1007/978-0-387-73337-1_2

Schulz, J. (2022). Entwicklung eines Testinstrumentes zur Erfassung von Kompetenzen im Umgang mit Messunsicherheiten.

Schweizer, K. & DiStefano, C. (2016). *Principles and Methods of Test Construction: Standards and Recent Advances*. Hogrefe Publishing.

Séré, M.-G. & Journeaux, R. (1995). Le traitement statistique des mesures en travaux pratiques de physique de DEUG : une innovation à Orsay. *Didaskalia*, *6*(1), 165–177.

Séré, M.-G., Journeaux, R. & Larcher, C. (1993). Learning the statistical analysis of measurement errors. *International Journal of Science Education*, *15*(4), 427–438.

Shakespeare, D. (2004). Starting an argument in science lessons. *School Science Review*.

Solomon, J. (1980). *Teaching Children in the Laboratory*. Croom Helm.

Solomon, J. (1992). The classroom discussion of science-based social issues presented on television: knowledge, attitudes and values. *International Journal of Science Education*, *14*(4), 431–444.

Steele, A. G. & Douglas, R. J. (2006). Extending E_{n} for measurement science. *Metrologia*, *43*(4), S235–S243.

Steinmetz, H. (2013). Analyzing Observed Composite Differences Across Groups. *Methodology*. https://econtent.hogrefe.com/doi/abs/10.1027/1614-2241/a000049

Stern, E. (2017). Anschlussfähiges Wissen aufbauen: Spiralcurricula für den Physikunterricht.

Stolurow, L. M. & Davis, J. D. (1963). *Teaching Machines and Computer-based Systems* (Techn. Ber.). University of Illinois. Urbana, Illinois.

Tabbers, H. K. (2002). *The modality of text in multimedia instruction: Refining the design guidelines* (PhD Thesis).

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296.

Taylor, J. R., Thompson, W., Thompson, W. J. & William Thompson. (1998). An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, *51*(1), 57–58.

Toulmin, S. E. (1958). *The Uses of Argument*. New York, NY, Cambridge University Press.

van de Pol, J., Volman, M. & Beishuizen, J. (2010). Scaffolding in Teacher–Student Interaction: A Decade of Research. *Educational Psychology Review*, *22*(3), 271–296.

Van Joolingen, W. R. & De Jong, T. (1997). An extended dual search space model of scientific discovery learning. *Instructional Science*, *25*(5), 307–346. https://www.jstor.org/stable/23371247

Vandenberg, R. J. & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70.

Veall, M. R. & Zimmermann, K. F. (1994). Evaluating Pseudo-R2's for binary probit models. *Quality & Quantity*, *28*(2), 151–164.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved R^ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2), 667–718.

Volkwyn, T., Trevor S. Volkwyn, Allie, S., Buffler, A., Lubben, F. & Campbell, B. (2004). First year Physic Students' Understanding of Measurement in the context of laboratory practicals.

von Aufschnaiter, C., Erduran, S., Osborne, J. & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching*, *45*(1), 101–131.

Wagner, S. (2018). *Erklärung physikalischer Phänomene mit Modellen* (Diss.). Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät.

Wang, J. & Buck, G. A. (2016). Understanding a High School Physics Teacher's Pedagogical Content Knowledge of Argumentation. *Journal of Science Teacher Education*, *27*(5), 577–604.

Weiji, Z. (1980). *The synthesis of random error and systematic error* (Techn. Ber. Nr. 02). J. Nanjing Univ. Sci. Technol.

Weinberger, A. & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, *46*(1), 71–95.

Wolf, E. J., Harrington, K. M., Clark, S. L. & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and psychological measurement*, *76*(6), 913–934.

Wright, B. (1996). Time 1 to Time 2 (Pre-test to Post-test) comparison: Racking and Stacking. *Rasch measurement transactions*, *10*(1), 478.

Wright, B. (2003). Rack and stack: time 1 vs. time 2. *Rasch measurement transactions*, *17*(1), 905–906.

Wright, B. & Linacre, J. (1994). easonable mean-square fit values. *Rasch measurement transactions*, *(*(3), 370.

Wu, M., Tam, H. P. & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers*. Singapore, Springer Singapore.

Xu, H. & Talanquer, V. (2013). Effect of the Level of Inquiry on Student Interactions in Chemistry Laboratories. *Journal of Chemical Education*, *90*(1), 29–36.

Yardimci Capkinoglu, E., Leblebicioglu, G., Metin, D. & Cetin, P. (2022). The impact of peer review on pre-service science teachers' written arguments about socioscientific issues related to chemistry.

Ye, X., Xiao, X., Shi, J. & Ling, M. (2016). The new concepts of measurement error theory. *Measurement*, *83*, 96–105.

Yu, C.-Y. (2002). *Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes* (Diss.). University of California. Los Angeles.

Zaccarelli, F. G., Schindler, A.-K., Borko, H. & Osborne, J. (2018). Learning from professional development: a case study of the challenges of enacting productive science discourse in the classroom. *Professional Development in Education*, *44*, 1–17.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C. & Shi, L. (2020). Applying machine learning in science assessment: a systematic review [Publisher: Routledge _eprint: https://doi.org/10.1080/03057267.2020.1735757]. *Studies in Science Education*, *56*(1), 111–151.

Zhang, L. (2016). Is Inquiry-Based Science Teaching Worth the Effort? Some Thoughts Worth Considering. *Science Education*, *25*(7), 897–915.

Zhao, Y., Huen, J. & Chan, Y. (2017). Measuring Longitudinal Gains in Student Learning: A Comparison of Rasch Scoring and Summative Scoring Approaches. *Research in Higher Education*.

Zohar, A. (2007). Science Teacher Education and Professional Development in Argumentation. In S. Erduran & M. P. Jiménez-Aleixandre (Hrsg.), *Argumentation in Science Education* (S. 245–268). Dordrecht, Springer Netherlands. http://link.springer.com/10.1007/978-1-4020-6670-2_12

Zohar, A. & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*(1), 35–62.

Danksagung

Mein aufrichtiger Dank gilt Jun.-Prof. Dr. Tobias Ludwig, der mir im Rahmen meiner Masterarbeit einen spannenden Einblick in die physikdidaktische Forschung und daran anschließend eine Möglichkeit zur wissenschaftlichen Qualifikation in Karlsruhe bot. Sein unermüdliches kritisches Fragen, die konstruktiven Diskussionen sowie zahlreichen Ratschläge haben mich dazu gebracht, meine Erkenntnisse immer weiter zu vertiefen, neue Perspektiven zu eröffnen und mich stets weiterzuentwickeln. Besonders dankbar bin ich ihm für die Freiheiten während meiner Betreuung sowie sein stets offenes Ohr und seine Unterstützung in allen Phasen der Promotion sowohl in Berlin als auch in Karlsruhe.

Ich bin besonders dankbar für Prof. Dr. Burkhard Priemer, der mein Interesse für die Fachdidaktik schon früh geweckt hat. Seine lebendigen und kritischen Diskussionen, insbesondere im Rahmen unserer Arbeitsgruppentreffen, haben mir wertvolle Einblicke eröffnet. Durch seine Anregungen während meines Promotionsprojekts hat er mich dazu ermutigt, meine Arbeit kontinuierlich zu hinterfragen und zu verbessern.

Außerdem möchte ich mich bei meiner Arbeitsgruppe erkenntlich zeigen: Vielen Dank an Kay Burger, Anja Fäger und Max Grünling, die mich in der Lehre und dem Lehr-Lern-Labor unterstützt haben und somit den Alltag an der PH sehr entlastet haben, sowie Dr. Ralph Hansmann, der mich bei der Anpassung des Kompetenztests für die Sekundarstufe I unterstützt hat. Ich möchte mich bei Lena Lenz, die mir vor allem bei der Pilotierung meiner Studie eine große Hilfe war, und Dr. Tina Schulze bedanken, die mir besonders in der Anfangszeit zur Seite stand und mir die Eingewöhnung in Karlsruhe sehr erleichtert hat. Mein besonderer Dank gilt Gregor Benz, der im Rahmen intensiver Diskussionen unzählige Tassen Kaffee mit mir getrunken und mir insbesondere in der Endphase meiner Promotion mit scharfsinnigen Anregungen weitergeholfen hat, und Olga Walter, mit der ich die wissenschaftliche Reise gemeinsam begonnen habe und deren Motivation und Unterstützung mich stets ermutigt haben und mir eine Stütze waren.

Darüber hinaus konnte ich von vielen spannenden und interessanten Treffen mit Kolleg:innen aus Berlin profitieren. An dieser Stelle bedanke ich mich bei Sophia Chroszczinsky, Dr. Karel Kok und Dr. Johannes Schulz.

Obwohl ich nicht zur Arbeitsgruppe gehörte, erhielt ich trotzdem Unterstützung von Prof. Dr. Samuel Merk, ohne den der Adler niemals gelandet wäre,

und Kirstin Schmidt, die bei zahlreichen Problemen und Diskussionen mit interessanten Lösungsideen beiseite stand. Vielen Dank für Eure Unterstützung.

Die Arbeit kann als Kind der Corona Pandemie angesehen werden. Umso mehr hat es mich erfreut, mich nach den Lockerungen der Bestimmungen mit Mitgliedern des Doktorand:innenkonvents sowohl beruflich als auch privat austauschen zu können. Bedanken möchte ich mich bei Kristina Bohrer, Tatjana Hille und Tamara Schilling.

Außerdem möchte ich den Expert:innen danken, die im Rahmen des Expert:innenratings die didaktische Struktur evaluiert haben. Zudem bedanke ich mich bei den Schulen, Lehrkräften und Schüler:innen, die mich bei der Durchführung und Planung der Studie unterstützt haben bzw. an der Studie teilgenommen haben.

Schließlich möchte ich meiner Familie und meinen Freunden einen besonderen Dank aussprechen, die mich auf dem Weg der Promotion begleitet haben. Vielen Dank an Tina, Florian, Madeleine und Niklas, die mit fortwährendem Austausch und Nachfragen zum Aufrechterhalten meiner Motivation beigetragen haben und mir eine große Unterstützung beim Korrekturlesen der Arbeit waren.

Besonders erkenntlich möchte ich mich bei meinen Brüdern zeigen, die vermutlich am meisten unter meiner angespannten und gestressten Laune gelitten haben und trotzdem stets für mich da waren und mir Zuspruch für die Arbeit gegeben haben. Vielen Dank für Eure Geduld und Eure Unterstützung.

Abschließend möchte ich von ganzem Herzen meinen Eltern danken, die stets geduldig hinter mir standen, mich bei jeder Entscheidung unterstützt und mir die wertvolle Möglichkeit gegeben haben, mich frei nach meinen eigenen Interessen zu entfalten.