

**Konzeption und Untersuchung videobasierter  
Aufgaben im Rahmen vorlesungsbegleitender  
Übungen zur Experimentalphysik (Mechanik)**

Dissertation

**Pascal Klein**

Vom Fachbereich Physik der Technischen Universität  
Kaiserslautern zur Verleihung des akademischen Grades  
„Doktor der Naturwissenschaften“ genehmigte Dissertation

Betreuer: Prof. Dr. Jochen Kuhn

Zweitgutachter: Prof. Dr. Andreas Müller

Datum der wissenschaftlichen Aussprache: 15. Juli 2016

D 386



# Kurzzusammenfassung

## Ausgangspunkte

Die Videoanalyse ist ein Verfahren zur berührungslosen Vermessung physikalischer Bewegungsvorgänge, dessen Messprinzip einen einfachen experimentellen Zugang zu physikalischen Größen ermöglicht. Im Bewusstsein des großen Potentials dieses Verfahrens hinsichtlich der Steigerung des Lernerfolgs von Lernenden in Schule und Hochschule werden in der Physikdidaktik seit über 20 Jahren Experimentierbeispiele entwickelt, die vor allem — aber nicht ausschließlich — in der Mechanik verortet sind. Trotz dieser reichhaltigen Entwicklungslage existieren nur wenige empirische Studien zur Effektivität dieses Medieneinsatzes sowie keine systematischen Implementationsstrategien des Mediums in universitären Lehr-/Lernszenarien. Dabei stellt die Verwendung der Videoanalyse in der Eingangsphase von Physikstudiengängen eine vielversprechende Gelegenheit dar, den Problemen hoher Schwundquoten (Studienabbruch und Studienfachwechsel) und der mehrfach diagnostizierten Ineffektivität traditioneller Lehr-/Lernmethoden zu begegnen, ohne dabei eine strukturelle Änderung des Lehrangebots in Kauf nehmen zu müssen.

## Vorgehen

Mit Blick auf die dargestellte Problematik der Studieneingangsphase wurden zwei **neue Aufgabenformate für Übungen zu Physikvorlesungen** eingeführt, welche die Möglichkeiten der Videoanalyse zur Steigerung der Lerneffektivität aufgreifen. Im Rahmen wöchentlicher Übungen zur Experimentalphysik (Mechanik) bearbeiten Studierende Videoanalyse-Aufgaben, in denen gegebene Experimentalvideos zur Beantwortung theoretischer und experimenteller Fragestellungen analysiert werden. Ferner führen Studierende bei der Bearbeitung von mobilen Videoanalyse-Aufgaben selbstständig Freihandexperimente mit einfachen Alltagsmaterialien durch, nehmen diese mit mobilen Endgeräten (Smartphone / Tablet-PC) auf und werten sie aus. Die neuen Aufgabenformate bieten dabei schon vor dem Laborpraktikum Interaktionsmöglichkeiten mit Realexperimenten und stellen eine Erweiterung zu den seit Jahren wohl etablierten traditionellen Aufgabenstellungen dar. Im Problemlösungsprozess der Aufgaben wird das selbstständige Erheben von Daten, der Umgang mit multiplen Repräsentationen und die Integration experimenteller und theoretischer Aktivitäten genutzt, um ein tieferes physikalisches Verständnis der im Video abgebildeten Prozesse zu fördern.

Im Rahmen dieser Arbeit wurden zahlreiche videobasierte Aufgabenstellungen konzipiert, deren Wirksamkeit in einer Feldstudie empirisch untersucht wurde. Dabei stützen sich die Hypothesen und Forschungsfragen der Untersuchung auf verschiedene Theorien und Theorierahmen aus der pädagogischen Psychologie (Cognitive Load Theory, Cognitive-Affective Theory of Learning with Media, Situiertes Lernen, Neugierde-Theorie), die positive Auswirkungen des Instruktionsmaterials auf kognitive (Leistung) und affektive Variablen (Motivation, Neugierde) erwarten lassen. Da aufgabenbasierte Interventionen in der Experimentalphysik bisher kaum in den Blick empirischer Forschung genommen wurden, mussten zunächst adäquate Testinstrumente konstruiert und psychometrisch validiert werden, um die Fragestellungen quantitativ beantworten zu können. An der **Pilotstudie** (Wintersemester 2012/2013 und Sommersemester 2013) nahmen 83 Studierende teil; die anschließende **Hauptstudie** (Wintersemester 2013/2014 und Wintersemester 2014/2015) wurde mit 165 Studierenden durchgeführt. In beiden Studien wurde die Effektivität von videobasierten

Aufgaben gegenüber inhaltsgleichen traditionellen Aufgaben in einem Rotationsdesign mit Versuchs- und Kontrollgruppe verglichen.

### Ausgewählte Ergebnisse

Es konnte zunächst festgestellt werden, dass sich sowohl die Bearbeitungszeit der Aufgaben als auch deren empfundene Schwierigkeit nicht signifikant zwischen den Gruppen unterschieden. Das Aufgabenformat wirkte sich auch nicht auf die wahrgenommene Beanspruchung durch das Studium im Allgemeinen aus. Die möglicherweise befürchteten Nachteile, dass videobasierte Aufgaben mehr Zeit in Anspruch nehmen würden als vergleichbare traditionelle Aufgaben oder dass die Lernenden durch Einführung eines neuen Mediums bzw. neuen Aufgabenformats überfordert wären, konnten demnach quantitativ nicht nachgewiesen werden. Weiterhin wurden explizite Vorteile durch das Bearbeiten von Videoanalyse-Aufgaben sowohl auf kognitiver als auch auf affektiver Ebene nachgewiesen:

Die **Repräsentationskompetenz**, d.h. die Fähigkeit, physikalische Darstellungsformen (Diagramme, Formeln, Vektoren) zu interpretieren und zueinander in Bezug zu setzen, war nach vier Übungswochen in der Interventionsgruppe signifikant größer als in der Kontrollgruppe (Signifikanzniveau  $p < 0.001$ , Effektstärke  $d = 0.72$ ). Unabhängig vom präsentierten Inhalt der gelösten Testaufgaben zeigten Studierende der Interventionsgruppe repräsentationsübergreifend bessere oder zumindest gleichwertige Leistungen im Vergleich zu Studierenden der Kontrollgruppe. Der mittelgroße bis große Effekt spricht für die praktische Bedeutsamkeit des Gruppenunterschieds und wird durch eine Mischungsanalyse, die variable Gewichtungen unterschiedlicher Typen von Testaufgaben berücksichtigt, methodisch abgesichert. Hierfür wurde eine neue analytische Behandlung für das Problem gemischter Aufgabenformate entwickelt. Auch bezüglich des **Konzeptverständnisses** trat hypothesenkonform ein signifikanter Leistungsunterschied zu Gunsten der Interventionsgruppe auf ( $p < 0.05$ ,  $d = 0.34$ ), der jedoch weniger stark ausgeprägt ist. In beiden Fällen konnte kein Einfluss der fachlichen Vorkenntnisse auf das Ergebnis festgestellt werden; die Förderung ist also weitgehend unabhängig von den Voraussetzungen der Lernenden.

Auf affektiver Ebene wurden weitere Effekte gefunden, die auf das Bearbeiten der videobasierten Aufgaben zurückgeführt werden können: Die Interventionsgruppe assoziierte die Übungsaufgaben mit einem größeren **Realitätsbezug** ( $p < 0.05$ ,  $d = 0.37$ ) und einem größeren **Experimentbezug** ( $p < 0.001$ ,  $d = 1.02$ ) als die Kontrollgruppe. Außerdem zeigte sie einen größeren Drang nach Wissen (**epistemische Neugierde**) bzgl. der Mechanik-Inhalte ( $p = 0.05$ ,  $d = 0.39$ ). Eine Förderung des Selbstkonzepts, des Interesses / Engagements oder der Autonomie bei der Aufgabenbearbeitung konnte nicht nachgewiesen werden.

Da die Instruktionsmaterialien unter Prämisse des fairen Vergleichs erstellt wurden, schöpfen die verwendeten Videoanalyse-Aufgaben noch nicht ihr volles Potential aus. Die Korrespondenz zu traditionellen Aufgaben stellt beispielsweise eine Restriktion bzgl. der Offenheit der Aufgabenstellungen und dem Grad der Kontextualisierung der abgebildeten Inhalte dar, sodass die erzielten Effekte unter konservativen Bedingungen zustande gekommen sind.

Ergänzend zu der Untersuchung von Lerneffekten videobasierter Aufgaben liegt ein wesentliches Ergebnis der Arbeit in der Konstruktion eines Testinstruments zur Erhebung repräsentationaler Fähigkeiten im Themenbereich der Kinematik. Durch klassische Test- und Itemanalysen auf Grundlage einer Eich-Stichprobe ( $N = 165$ ) wurde aus einer initialen Item-Menge eine hinreichend reliable Skala extrahiert (Reliabilität nach Cronbach:  $\alpha_C = 0.82$ ). Darüber hinaus konnte die Existenz eines Rasch-skalierbaren Kerns nachgewiesen werden, wodurch ein hohes statistisches Messniveau erreicht wird (spezifisch objektive Messung). Das Instrument beinhaltet eine Sicherheitsskala und liefert somit im Rahmen der vorliegenden Studie interessante Erkenntnisse über metakognitive Fähigkeiten der Studierenden und vorhandene Fehlkonzepte. Darüber hinaus tragen die mit dem Instrument gewonnenen Daten dazu bei, repräsentationale Fähigkeiten in Verbindung mit anderen Indikatoren der Lernleistung (z.B. Konzeptverständnis, Klausurergebnis) zu bringen.

# Abstract

## Starting Point

Video-based motion analysis (VBMA) is a simple and illustrative method to obtain and visualize experimental data of physical motion processes. For more than 20 years, this method has been used in physics education to foster learning of various mechanical concepts in both schools and universities. While many practical examples and conceptual contributions about applications of VBMA exist, there is virtually no empirical research about learning outcomes. Moreover, no systematic integration of student-centred video-based learning activities into university physics courses took place yet. This work explores the use of VBMA in first-year university physics courses. It is argued that the instructional strategy has the potential to support knowledge gains more effectively than traditional learning approaches without changing the basic structure of physics courses.

## Procedure and Methods

We define **two new types of video-based problems**, which are implemented into **weekly recitations of introductory physics**. First, students work with *VBMA-tasks* consisting of a pre-recorded video of an experiment combined with a pen-and-paper exercise. In self-study („homework“), students perform on-screen measurements of physical quantities and answer theoretical questions about the experiment (off-screen). Second, students perform hands-on experiments themselves and record them using a mobile device such as a smartphone or a tablet computer (*mobile VBMA-task*). Both types of video-based problems provide opportunities to gain experimental experiences in preparation for laboratory work and are used as an extension of traditional problems. During the problem-solving process, students gather their own data, use multiple representations (such as graphs, pictures and equations) and connect experimental results with theoretical calculations / assumptions in order to improve their conceptual understanding.

Within the framework of this thesis, several VBMA-tasks and mobile VBMA-tasks have been developed covering the most important concepts of university mechanics courses. The effectiveness of this instruction-based approach has been investigated in a field study. Based on the theoretical framework of educational psychology (cognitive load theory, cognitive-affective theory of learning with media, situated learning, curiosity), we hypothesize that (mobile) VBMA-tasks have positive impacts on both cognitive and affective variables. In order to investigate our hypotheses, we have developed and validated suitable measurement instruments. A **pilot study** with 83 students was conducted during winter term 2012/2013 and during summer term 2014; the subsequent **main study** (winter term 2013/2014 and winter term 2014/2015) involved 165 students. Both the pilot and main study investigated the effectiveness of video-based problems compared to traditional problems using a two-group pre-post design with rotation.

## Selected Results

First of all, results showed no differences between the VBMA and control groups for both time-on-task and perceived task difficulty. Moreover, the format of the problem had no impact on students' perceived stress. Therefore, quantitative evidence concerning potential disadvantages

of video-based problems — such as being more time-demanding than traditional problems or overextending students by introducing new media tools — were not found. Furthermore, results showed that working with (mobile) VBMA-tasks provides distinct advantages for supporting development of cognitive and affective learning:

The **Representational competence** of the students in the treatment group was significantly higher than in the control group ( $p < 0.001$ ,  $d = 0.72$ ) after four weeks of intervention. Representation competence refers to the skills of students to illustrate physical phenomena in more than one format, referred to as an external representation: graphs, equations, vectors, pictures. Independent from the specific physics content, students in the treatment group performed better than (or at least equal to) students in the control group for each type of representation. The medium to large effect size indicates practical relevance of the results, and was validated by a mixture analysis accounting for different types of item formats. For this purpose, a new method was developed dealing with the problem of combined test scores. In addition to the positive effect on representational competence, (mobile) VBMA-tasks had a significant positive impact on **conceptual understanding** ( $p < 0.05$ ,  $d = 0.34$ ). In both cases, the results were not influenced by students' prior knowledge. Students benefit from the intervention independent from their preconditions.

Concerning affective variables, we also found positive effects induced by (mobile) VBMA-tasks: the treatment group associated the tasks with a significantly higher **factual authenticity** ( $p < 0.05$ ,  $d = 0.37$ ), which means that the students perceived the tasks as being more strongly connected to real-world problems. In addition, students were more strongly connected to **disciplinary methods of experimentation** ( $p < 0.001$ ,  $d = 1.02$ ). Moreover, the **epistemic curiosity** regarding the contents of mechanics in the treatment group was significantly higher than in the control group ( $p = 0.05$ ,  $d = 0.39$ ). No effects concerning self-concept, interest / engagement or autonomy during problem solving were found.

Given that the instructions were designed to guarantee fairness among treatment groups, the full potential of (mobile) VBMA-tasks could not be exploited. For example, the (mobile) VBMA-tasks were restricted in terms of degree of openness and contextualization. Therefore, it is worth noting that the results were achieved under conservative conditions.

Apart from studying the effects of (mobile) VBMA tasks on learning, a substantial part of this work is concerned with the construction of an instrument to assess the representational competence of students within the conceptual framework of kinematics. We used a calibration sample for item analysis and were able to construct a reliable scale of questionnaire items (Cronbach's  $\alpha_c = 0.82$ ). Moreover, a subset of items fulfils the conditions of a Rasch scale. In the present study, the instrument reveals insights into the relationship between representational competence and relevant indicators of performance (such as conceptual understanding or exam performance). Moreover, the instrument comprises a confidence scale which allows exploration of metacognitive skills concerning representational competence as well as misconceptions.

# Inhaltsverzeichnis

Kurzzusammenfassung . . . . .	4
Abstract . . . . .	6
Symbol- und Abkürzungsverzeichnis . . . . .	12
Abbildungsverzeichnis . . . . .	14
Tabellenverzeichnis . . . . .	16
<b>1 Einleitung</b>	<b>19</b>
1.1 Ausgangspunkte . . . . .	20
1.2 Aufbau der Arbeit . . . . .	22
1.3 Projektverlaufsplan . . . . .	23
<b>I Videoanalyse: Theoretischer Hintergrund und Stand der Forschung</b>	<b>25</b>
<b>2 Videoanalyse als Physikmedium</b>	<b>27</b>
2.1 Messprinzip und Anwendungen in der Lehre . . . . .	27
2.1.1 Längen- und Zeitskalierung des Videos . . . . .	28
2.1.2 Messdatenerfassung . . . . .	28
2.1.3 Verarbeitung und Darstellung der Messdaten . . . . .	29
2.1.4 Messfehler und Vermeidung von Messfehlern . . . . .	29
2.1.5 Literaturrecherche zu Einsatzmöglichkeiten . . . . .	30
2.2 Videoanalyse mit mobilen Endgeräten . . . . .	34
2.2.1 Smartphones und Tablet-Computer als Experimentiermedien . . . . .	35
2.2.2 Applikationen zur mobilen Videoanalyse . . . . .	35
2.2.3 Anwendungsbeispiele . . . . .	36
<b>3 Theoretischer Hintergrund und empirischer Forschungsstand</b>	<b>39</b>
3.1 Physikdidaktischer und lernpsychologischer Hintergrund . . . . .	39
3.1.1 Visualisieren und qualitatives Begründen . . . . .	39
3.1.2 Kontextualisierung und situiertes Lernen . . . . .	40
3.1.3 Lernen mit multiplen Repräsentationen . . . . .	41
3.1.4 Theorie-Experiment-Wechselwirkung . . . . .	46
3.2 Studien zur Lernwirksamkeit der Videoanalyse . . . . .	46
3.2.1 Wirksamkeit bezüglich kognitiven Variablen . . . . .	48
3.2.2 Wirkung bzgl. affektiver Variablen, Sonstiges . . . . .	50
3.2.3 Zusammenfassung und Bewertung der Untersuchungslage . . . . .	51
<b>4 Konzeption videobasierter Aufgaben und erste Untersuchungsergebnisse</b>	<b>53</b>
4.1 Videoanalyse-Aufgaben . . . . .	53
4.2 Mobile Videoanalyse-Aufgaben . . . . .	57
4.3 Traditionelle Aufgaben . . . . .	59
4.4 Produktion von Videoanalyse-Aufgaben und Implementation in eine bestehende Übungsstruktur . . . . .	59

4.5	Pilotstudie im WS 12/13 und SS 13 . . . . .	60
4.5.1	Methoden und ausgewählte Ergebnisse . . . . .	60
4.5.2	Diskussion und Konsequenzen für die Hauptstudie . . . . .	60
<b>II Instrumentenentwicklung und Validierung</b>		<b>63</b>
<b>5</b>	<b>Konzeption der Fragebögen und Tests</b>	<b>65</b>
5.1	Leistungstest KiRC: Repräsentationskompetenz in der Kinematik . . . . .	65
5.1.1	Inhaltsdomäne, Entwicklungsbedarf und Vorarbeiten . . . . .	65
5.1.2	Entwicklungsschritte und Inhalte des Test . . . . .	67
5.1.3	Itemformate und Testsumme . . . . .	70
5.2	Konzepttest zu Bewegungen . . . . .	71
5.3	Erhebung von Antwortsicherheiten . . . . .	71
5.3.1	Mehrstufige Multiple-Choice-Items und Ziele von confidence rating . . . . .	72
5.3.2	Confidence rating in dieser Arbeit . . . . .	72
5.4	Fragebogen zur Erfassung verschiedener Motivationskomponenten . . . . .	73
5.4.1	Interesse / Engagement (IE) . . . . .	73
5.4.2	Akademisches Selbstkonzept (SK) . . . . .	74
5.4.3	Realitätsbezug / Authentizität der Aufgaben (RA) . . . . .	74
5.4.4	Disziplinäre Authentizität (dA) . . . . .	75
5.4.5	Autonomie (Atnm) . . . . .	75
5.5	Neugierdefragebogen . . . . .	76
5.6	Belastungsfragebogen . . . . .	76
5.7	Betreuerfragebogen . . . . .	77
<b>6</b>	<b>Testanalysen nach Klassischer Testtheorie (KTT)</b>	<b>79</b>
6.1	Methodische Grundlagen . . . . .	79
6.1.1	Prinzip der Skalenbildung . . . . .	79
6.1.2	Faktorenanalyse . . . . .	80
6.1.3	Psychometrische Kenngrößen der KTT . . . . .	81
6.1.4	Prüfung auf Normalverteilung und Ausreißerkorrektur . . . . .	84
6.2	Definition der Eich-Stichprobe und Strukturierung der Testanalysen . . . . .	84
6.3	Ergebnisse und Diskussion I: Ausreißeranalyse aller Items . . . . .	86
6.4	Ergebnisse und Diskussion II: Faktorenanalysen und psychometrische Kenngrößen	86
6.4.1	Belastungsfragebogen . . . . .	87
6.4.2	Neugierdefragebogen . . . . .	88
6.4.3	Motivationsfragebogen . . . . .	89
6.4.4	KiRC-Instrument . . . . .	92
6.4.5	Konzepttest . . . . .	96
6.4.6	Antwortsicherheiten . . . . .	97
6.4.7	Betreuerfragebogen . . . . .	98
6.5	Ergebnisse und Diskussion III: Normalverteilungsanalyse der Skalenmittelwerte	98
6.6	Zusammenfassung . . . . .	100
<b>7</b>	<b>Testanalysen nach Item-Response-Theorie (IRT)</b>	<b>103</b>
7.1	Methodische Grundlagen der IRT und des Rasch-Modells . . . . .	103
7.1.1	Das Rasch-Modell . . . . .	104
7.1.2	Prinzip der Parameterschätzung . . . . .	107
7.1.3	Modellgeltungstests nach Invarianzprinzipien . . . . .	107
7.1.4	Globale Modellgeltungstests . . . . .	108
7.1.5	Itemfit-Statistiken . . . . .	109
7.2	Rasch-Analyse des Leistungstests KiRC . . . . .	109
7.2.1	Spezifikationen der Testanalysen . . . . .	109

7.2.2	Ergebnisse I: Invarianzeigenschaften . . . . .	111
7.2.3	Ergebnisse II: Globale Modellpassung . . . . .	114
7.2.4	Ergebnisse III: Itemparameter, Item-Fit und ICCs . . . . .	114
7.2.5	Ergebnisse IV: Graphische Darstellung der Item- und Personenparameter . . . . .	116
7.3	Rasch-Analyse des Konzepttests . . . . .	117
7.4	Diskussion der Ergebnisse und Zusammenfassung . . . . .	117
<b>8</b>	<b>Analyse leistungsbezogener Antwortsicherheiten</b>	<b>119</b>
8.1	Methodische Grundlagen . . . . .	119
8.1.1	Berechnung und Vergleich absoluter Größen . . . . .	119
8.1.2	Zusammenhangsmaße auf Stichprobenebene . . . . .	120
8.1.3	Zusammenhangsmaße auf Itemebene . . . . .	120
8.2	Fragestellungen und Vorgehen . . . . .	121
8.3	Ergebnisse . . . . .	121
8.3.1	Ergebnisse auf Stichprobenebene . . . . .	121
8.3.2	Ergebnisse auf Ebene einzelner Items . . . . .	121
8.4	Diskussion und Interpretation der Ergebnisse . . . . .	122
<b>III</b>	<b>Hauptuntersuchung</b>	<b>125</b>
<b>9</b>	<b>Hypothesen und Forschungsfragen</b>	<b>127</b>
<b>10</b>	<b>Material und Methode</b>	<b>131</b>
10.1	Gesamtstichprobe . . . . .	131
10.2	Studiendesign . . . . .	132
10.2.1	Gruppeneinteilung und Kern-Stichprobe . . . . .	133
10.2.2	Begründung des Studiendesigns und kritische Stellungnahme . . . . .	134
10.3	Instruktionsmaterial . . . . .	135
10.3.1	Interventionsstärke (Aufgabendosis) . . . . .	135
10.3.2	Faire Vergleichbarkeit der Aufgaben . . . . .	136
10.4	Zielvariablen und Erhebungsinstrumente . . . . .	136
10.5	Statistische Methoden und Kenngrößen . . . . .	137
10.5.1	Kovarianzanalyse (ANCOVA) . . . . .	138
10.5.2	Effektstärken . . . . .	138
10.5.3	Ratekorrekturen (Formula Scoring) . . . . .	139
10.5.4	Normierter Lernzuwachs (Hake-Index) . . . . .	140
10.5.5	Gewichtete Testsummenbildung (Testmischungsanalyse) . . . . .	141
10.6	Strukturierung der Analysen . . . . .	143
<b>11</b>	<b>Ergebnisse</b>	<b>147</b>
11.1	Voranalysen und untergeordnete Fragestellungen . . . . .	148
11.1.1	Unterschiede zwischen Studienabbrechern und Weiterstudierenden . . . . .	148
11.1.2	Verhalten des Übungsleiters in den Gruppen . . . . .	148
11.1.3	Kenngrößen der Instrumente innerhalb der Kern-Stichprobe . . . . .	149
11.1.4	Ausreißeranalysen . . . . .	150
11.1.5	Festlegung der ANCOVA-Modelle . . . . .	150
11.2	Ergebnisse zur Hauptfragestellung 1: Wirksamkeit . . . . .	153
11.2.1	Repräsentationskompetenz . . . . .	154
11.2.2	Konzeptverständnis . . . . .	156
11.2.3	Motivation . . . . .	158
11.2.4	Neugierde . . . . .	164
11.3	Ergebnisse zur Fragestellung 2: Aufgabeninduzierte Belastung . . . . .	165
11.3.1	Aufgabenschwierigkeit (Cognitive Load) . . . . .	165

11.3.2	Beanspruchung . . . . .	166
11.3.3	Bearbeitungszeit . . . . .	167
11.4	Ergebnisse zu den methodischen Fragestellungen . . . . .	168
11.4.1	Manipulation Check . . . . .	168
11.4.2	Effektstärken- und Reliabilitätsanalyse von KiRC mit variabler Testsummengewichtung . . . . .	169
11.4.3	Einfluss einer Ratekorrektur auf die Ergebnisse der Vergleichsuntersuchung . . . . .	171
11.5	Ergebnisse zur Robustheitsfrage . . . . .	172
11.5.1	Einfluss des Geschlechts auf die Zielvariablen . . . . .	172
11.5.2	Einfluss der Vorleistungen auf die Zielvariablen . . . . .	174
11.5.3	Einflussfaktoren auf den Lernzuwachs . . . . .	174
11.6	Analyse der Klausurergebnisse . . . . .	175
11.7	Korrelationsanalysen . . . . .	177
<b>12</b>	<b>Diskussion</b>	<b>179</b>
12.1	Erster Interventionszeitraum . . . . .	180
12.1.1	Repräsentationskompetenz . . . . .	180
12.1.2	Konzeptverständnis (Kinematik) . . . . .	183
12.1.3	Motivation . . . . .	184
12.1.4	Neugierde . . . . .	186
12.1.5	Belastungsindikatoren . . . . .	187
12.2	Zweiter Interventionszeitraum . . . . .	188
12.2.1	Leistung (Konzeptverständnis) . . . . .	189
12.2.2	Motivation und Neugierde . . . . .	189
12.2.3	Belastungsindikatoren . . . . .	190
12.3	Klausurergebnis . . . . .	190
12.4	Ursache für Studienausstieg . . . . .	191
12.5	Einschränkungen und Generalisierbarkeit . . . . .	192
<b>Fazit und Ausblick</b>		<b>195</b>
<b>Anhänge</b>		<b>200</b>
<b>Textbezogene Anhänge</b>		<b>201</b>
A	Zeitschriftenliste der Literaturrecherche . . . . .	201
B	Testinstrumente . . . . .	202
C	Ergänzende Ergebnisberichte . . . . .	211
D	Berechnungen zur Testmischungsanalyse . . . . .	219
E	Instruktionsmaterialien . . . . .	220
<b>Literaturverzeichnis</b>		<b>237</b>
<b>Sonstige Anhänge</b>		<b>249</b>
Lebenslauf . . . . .		249
Publikationsliste . . . . .		250

# Symbol- und Abkürzungsverzeichnis

## Kenngrößen (klassische Testtheorie)

$n$	Anzahl Items
$D$	Diskriminationsindex
$\lambda$	Eigenwert
$\kappa_F$	Interrater-Reliabilität (Fleiss)
$j$	Item-Laufindex
$r_{jt}$	Item-Test-Korrelation
$\kappa$	Kommunalität
$k$	Kurtosis
$M$	Mittelwert
$\alpha_c$	Reliabilität (nach Cronbach)
$\alpha_{\setminus j}$	Reliabilität ohne Item $j$
$s$	Schiefe
$P$	Schwierigkeitsindex
$S$	Sicherheitsindex
$FS$	Stabilität einer Faktorenlösung
$SD$	Standardabweichung
$N$	Stichprobengröße
$T$	Testsumme

## Abkürzungen

AV	Abhängige Variable
FCI	Force Concept Inventory
FSI	Fragebogen Studieninteresse
HS	Hauptstudie
IG	Interventionsgruppe
KG	Kontrollgruppe
KV	Kovariate
MC	Multiple-Choice
MTF	Multiple-True-False
RF	Repräsentationsformen
SWS	Semesterwochenstunden
SS/WS	Sommer-/Wintersemester
TUG-K	Test of Understanding Graphs: Kinematics

## Statistische Kennzahlen

$df$	Anzahl der Freiheitsgrade
$R^2$	Bestimmtheitsmaß
$\chi^2$	Chi-Quadrat Statistik
$d$	Effektstärke
$r$	Korrelationskoeffizient
$CAQ$	Maß der Selbsteinschätzung
$\beta$	Regressionskoeffizient
$\alpha$	Signifikanzniveau
$t$	Teststatistik $t$ -Test
$p$	Überschreitungswahrscheinlichkeit

## Item-Response-Theorie

$\sigma$	Aufgabenschwierigkeit
$\chi^2, \xi^2$	Globale Prüfgrößen
$L$	Likelihood
$\mathcal{P}$	Lösungswahrscheinlichkeit
$\eta$	Personenfähigkeit

## Untersuchungsvariablen

CL	Aufgabenschwierigkeit
Atnm	Autonomie
PSQ	Beanspruchung
ToT	Bearbeitungszeit
dA	Disziplinäre Authentizität
Exp	Experimentbezug
IE	Interesse / Engagement
Kon	Konzeptverständnis
NZ	Neugierde (aktueller Zustand)
NP	Neugierde (Personeneigenschaft)
RA	Realitätsbezug
KiRC	Repräsentationskompetenz
SK	Selbstkonzept
$t_0, t_1, t_2$	Testzeitpunkte



# Abbildungsverzeichnis

Hinweis: Das Verzeichnis weist abgekürzte Bildunterschriften auf.

Abb. 1:	Projektverlauf . . . . .	24
Abb. 2:	Messdatenerfassung durch Videoanalyse und Messdaten-Darstellung . . .	28
Abb. 3:	Anzahl Publikationen zur Videoanalyse über die Zeit . . . . .	31
Abb. 4:	Beispiele zur Videoanalyse . . . . .	33
Abb. 5:	Beispiel zur mobilen Videoanalyse . . . . .	36
Abb. 6:	Beispiele für multiple Repräsentationsformen . . . . .	40
Abb. 7:	Beispiel für kohärente Repräsentationen . . . . .	42
Abb. 8:	Modell der CATLM . . . . .	43
Abb. 9:	V-Aufgabe 2.1 mit Lösungsansatz . . . . .	55
Abb. 10:	Stroboskopbild zur V-Aufgabe 4.2 und Lösungsansatz . . . . .	55
Abb. 11:	Stroboskopbild zur V-Aufgabe 5.2 und Lösungsansatz . . . . .	56
Abb. 12:	Bilderserie zur V-Aufgabe 7.2 und gemessene Kraft . . . . .	56
Abb. 13:	Zwei Beispielitems des Leistungstests KiRC. . . . .	69
Abb. 14:	Vorgehensweise bei Skalenbildung (schematisch). . . . .	80
Abb. 15:	Zusammenhang zwischen $P_j$ und $D_j$ (empirische Daten) . . . . .	83
Abb. 16:	Screeplot (Belastungsfragebogen) . . . . .	87
Abb. 17:	Änderungssensitivität KiRC . . . . .	95
Abb. 18:	Histogramme der Skalenmittelwerte (KTT) . . . . .	99
Abb. 19:	Aufgabencharakteristische Kurven im Rasch-Modell . . . . .	105
Abb. 20:	DIFplot . . . . .	111
Abb. 21:	Verteilung der $\xi^2$ -Werte nach Simulation Rasch-konformer Datensätze . .	114
Abb. 22:	Aufgabencharakteristische Kurven (KiRC) . . . . .	115
Abb. 23:	Raschparameter (KiRC) . . . . .	116
Abb. 24:	Ergebnisse Confidence-Rating . . . . .	122
Abb. 25:	Studienverlaufsplan . . . . .	132
Abb. 26:	Ergebnis des Gruppenvergleichs: KiRC . . . . .	154
Abb. 27:	Ergebnisse KiRC nach Inhalt und Repräsentationsform . . . . .	155
Abb. 28:	Ergebnis des Gruppenvergleichs: Konzepttest . . . . .	156
Abb. 29:	Item Kon3 und Kon4 aus dem Konzepttest . . . . .	158
Abb. 30:	Ergebnis des Gruppenvergleichs: Realitätsbezug . . . . .	159
Abb. 31:	Ergebnis des Gruppenvergleichs: Selbstkonzept . . . . .	160
Abb. 32:	Ergebnis des Gruppenvergleichs: Interesse / Engagement . . . . .	161
Abb. 33:	Ergebnis des Gruppenvergleichs: disziplinäre Authentizität . . . . .	162
Abb. 34:	Ergebnis des Gruppenvergleichs: Autonomie . . . . .	163
Abb. 35:	Ergebnis des Gruppenvergleichs: Neugierde (als aktueller Zustand) . . . .	164
Abb. 36:	Ergebnis des Gruppenvergleichs: PSQ und CL . . . . .	166
Abb. 37:	Streudiagramme der Beanspruchung . . . . .	167
Abb. 38:	Ergebnis des Gruppenvergleichs: Bearbeitungszeit . . . . .	167
Abb. 39:	Ergebnis des Gruppenvergleichs: Experimentbezug . . . . .	169
Abb. 40:	Testmischungsanalyse . . . . .	170
Abb. 41:	Mischungsanalyse ratekorrigierter Testsummen . . . . .	172

Abb. 42:	Interaktionseffekte zwischen abhängigen Variablen und Vorleistungen . .	174
Abb. A-1:	Screeplots . . . . .	211
Abb. A-2:	Modellprüfung, Personen- und Itemparameter aus Rasch-Analyse (Konzepttest) . . . . .	213
Abb. A-3:	Aufgabencharakteristische Kurven (Konzepttest) . . . . .	213
Abb. A-4:	Ergänzende Streudiagramme der Belastungsvariablen (CL und PSQ) . .	218

# Tabellenverzeichnis

Tab. 1:	Funktionalitäten von Videoanalyse-Programmen zur Realisierung des Messprinzips . . . . .	29
Tab. 2:	Videoanalyse-Experimente zu universitären Mechanik-Inhalten . . . . .	32
Tab. 3:	Empirische Studien zur Lernwirksamkeit der Videoanalyse . . . . .	47
Tab. 4:	Konstruktionsangaben zu den Items der Leistungstests . . . . .	68
Tab. 5:	Psychometrische Kenngrößen der KTT . . . . .	82
Tab. 6:	Zeitpunkte der eingesetzten Fragebögen in beiden Hauptstudien . . . . .	85
Tab. 7:	Ergebnisse der Ausreißeranalyse auf Itemebene . . . . .	86
Tab. 8:	Prüfung der Datensätze auf Eignung zur Hauptkomponentenanalyse . . . . .	86
Tab. 9:	Rotierte Faktorlösung (Belastung) . . . . .	87
Tab. 10:	Ergebnisse der Itemanalyse (Belastung) . . . . .	88
Tab. 11:	Ergebnisse der Faktoren- und Itemanalyse (Neugierde) . . . . .	89
Tab. 12:	Ausschnitt der rotierten Faktorlösung (Motivation) . . . . .	90
Tab. 13:	Skalenkennwerte der Motivationskomponenten . . . . .	92
Tab. 14:	Festlegung der Kern-Itemgruppen und Itemanalysen (KiRC) . . . . .	93
Tab. 15:	Skalenkennwerte (KiRC) . . . . .	94
Tab. 16:	Rotierte Faktorlösung der optimierten KiRC-Itemgruppe . . . . .	94
Tab. 17:	Ergebnisse der Itemanalyse (Konzepttest) . . . . .	96
Tab. 18:	Stichprobenübergreifender Vergleich der Schwierigkeitsindizes einzelner FCI-Items . . . . .	96
Tab. 19:	Faktoren- und Itemanalyse, sowie Skalenkennwerte (Antwort sicherheiten) . . . . .	97
Tab. 20:	Faktoren- und Itemanalyse, sowie Skalenkennwerte (Betreuer-Items) . . . . .	98
Tab. 21:	Ergebnisse der Normalverteilungsanalyse . . . . .	98
Tab. 22:	Zusammenfassung der Ergebnisse zur Skalenbildung . . . . .	101
Tab. 23:	Ergebnisse des Andersen LR-Tests auf Subgruppeninvarianz . . . . .	111
Tab. 24:	Ergebnisse des Martin-Löf-Tests auf Itemhomogenität . . . . .	112
Tab. 25:	Prüfung auf lokale stochastische Unabhängigkeit . . . . .	113
Tab. 26:	Itemparameter $\sigma$ und Fitstatistiken der Rasch-Itemgruppe . . . . .	114
Tab. 27:	Studienübergreifender Vergleich von Zusammenhängen zwischen Leistungs- ( $P$ ) und Sicherheitsindikatoren ( $S$ ) . . . . .	123
Tab. 28:	Gruppengrößen zu Beginn der Intervention ( $t_0$ ) . . . . .	133
Tab. 29:	Variablen und Instrumente der Hauptuntersuchung . . . . .	137
Tab. 30:	Strukturierung der Analysen . . . . .	145
Tab. 31:	Vergleich zwischen Studienaussteigern und Weiterstudierenden . . . . .	148
Tab. 32:	Vergleich der Einschätzungen des Übungsleiters zwischen den Gruppen . . . . .	149
Tab. 33:	Skalenkenngrößen innerhalb der Kern-Stichprobe zu verschiedenen Testzeitpunkten . . . . .	149
Tab. 34:	Korrelationen zwischen KV und AV . . . . .	151
Tab. 35:	Korrelationen der AV zwischen den Testzeitpunkten . . . . .	151
Tab. 36:	Homogenität der Untersuchungsgruppen . . . . .	152
Tab. 37:	Ergebnisse der Kovarianzanalysen zu den Hauptfragestellungen . . . . .	153
Tab. 38:	Deskriptiver Vergleich der Item-Schwierigkeiten (Konzepttest) . . . . .	157
Tab. 39:	Distraktorenanalyse ausgewählter Konzept-Items . . . . .	158

Tab. 40:	Ergebnisse der Kovarianzanalysen zur Belastung . . . . .	165
Tab. 41:	Ergebnisse der Kovarianzanalysen zum Manipulation Check . . . . .	168
Tab. 42:	Zusammensetzung der KiRC-Itemgruppen . . . . .	171
Tab. 43:	Ergebnisse zweifaktorieller Kovarianzanalysen . . . . .	173
Tab. 44:	Deskriptive Daten zu Gender-Effekten . . . . .	173
Tab. 45:	Ergebnisse zur Robustheit des Lernzuwachses . . . . .	175
Tab. 46:	Zusammenhang zwischen Klausur(aufgaben) und kognitiven Variablen . .	176
Tab. 47:	Deskriptive Statistiken der Klausuraufgaben und Gruppenvergleich . . .	177
Tab. 48:	Korrelationsmatrix zwischen den AV zu $t_1$ (obere Hälfte) und $t_2$ (untere Hälfte) . . . . .	177
Tab. 49:	Hypothesen und Ergebnisse im Überblick . . . . .	179
Tab. A-1:	Motivation-Items mit Quellenangaben . . . . .	202
Tab. A-2:	Neugierde-Items mit Quellenangaben . . . . .	203
Tab. A-3:	Belastungs-Items mit Quellenangaben . . . . .	204
Tab. A-4:	Betreuer-Items mit Quellenangaben . . . . .	204
Tab. A-5:	Ausgewählte Ergebnisse des Wald-Tests . . . . .	212
Tab. A-6:	Ergebnisse des Andersen LR-Tests und des Wald-Tests (Konzept-Items)	212
Tab. A-7:	Fit-Statistiken der Konzept-Items und Korrelationsmatrix . . . . .	213
Tab. A-8:	Test auf Homogenität der Regressionssteigungen . . . . .	215
Tab. A-9:	Kovarianzmodelle . . . . .	215
Tab. A-10:	Mittelwerte und Standardabweichungen der abh. Variablen . . . . .	216
Tab. A-11:	Deskriptive Statistiken der KiRC-Items auf Gruppenebene . . . . .	217
Tab. A-12:	Deskriptive Statistiken aus der Skalenbildung ausgeschlossener Motivations- Items . . . . .	218
Tab. A-13:	Geschlechtsbezogene deskriptive Daten ausgewählter Variablen . . . . .	218

# Kapitel 1

## Einleitung

Physik zählt zu den unbeliebtesten Fächern der Schule, was vor allem am hohen Abstraktionsniveau der Inhalte und an mangelnden Alltagsbezügen liegt (Holstermann & Bögeholz, 2007; Hoffmann et al., 1998). In der „synthetischen Wirklichkeit“ des Physikunterrichts sieht der Lernende „Dinge, die [er] sonst nirgendwo sieht“ und verwendet fachsprachliche „Begriffe, die [er] nirgendwo sonst benötigt“ (Müller, 2006). Trotzdem entscheiden sich jährlich ca. 15 000 Schülerinnen und Schüler für einen grundständigen Physikstudiengang<sup>1</sup>, wie Statistiken über Studierendenzahlen an deutschen Universitäten zeigen (Düchs & Ingold, 2015). Aber auch im Studium stellen stark theorielastige Inhalte eine Schwierigkeit dar, die Niederschlag in seit Jahren hohen Schwundquoten (Studienabbruch und -fachwechsel) von ca. 30% findet.<sup>2</sup> Bei Studierenden des Lehramts liegt die Schwundquote sogar deutlich über 30%, was mit Blick auf den derzeitigen Mangel an Physiklehrkräften besonders kritisch zu betrachten ist.

Vor dem Hintergrund dieser Problematik wurden vor allem in den USA vielfältige Innovationsansätze entwickelt, erprobt und untersucht, die darauf abzielen, das Verständnis von Studierenden zu verbessern und die Attraktivität von Physik-Studiengängen zu steigern (vgl. Meta-Analyse von Ruiz-Primo et al., 2011). Die Diversität der Ansätze reicht von dem Aufgreifen kognitiv anregender Elemente während der Vorlesung<sup>3</sup> über eine räumliche Umgestaltung des Hörsaals in lernerzentrierte, aktive Lernumgebungen bis hin zum Ersatz der Vorlesungen durch selbstständige Theorie- und Experimentierarbeit in Kleingruppen unter Expertenberatung. Zwar erzeugen diese Konzepte überwiegend positive Resultate gegenüber traditionellen Lehrveranstaltungen, doch ist der mit den Maßnahmen verbundene organisatorische und finanzielle Aufwand in vielen Fällen sehr hoch. Hinzu kommt, dass die meisten Wirksamkeitsstudien unter mangelhaften Untersuchungsmethoden leiden (Ruiz-Primo et al., 2011), sodass sich aus den Ergebnissen nur bedingt Implikationen für Innovationsansätze der Studieneingangsphase in Deutschland ableiten lassen.

Im Gegensatz dazu wird in dieser Arbeit die Konzeption und Implementation von experimentellen Aufgaben beschrieben, die ähnliche Interventionserfolge ohne eine grundlegende Umstrukturierung des dreigeteilten experimentalphysikalischen Grundlagenmoduls (Vorlesung, Übungen und Praktikum) anstreben. Die Aufgaben greifen auf die Videoanalyse von Bewegungen zurück und kommen in den vorlesungsbegleitenden Übungen zum Einsatz, die für das Lernen physikalischer Inhalte zentral sind (Finkelstein & Pollock, 2005). Die Wirksamkeit dieser Aufgaben bzgl. kognitiven und affektiven Variablen wird gegenüber seit Jahren etablierten traditionellen Aufgabenstellungen empirisch untersucht.

<sup>1</sup>Damit gemeint sind Studiengänge mit Hauptfach Physik: Physik Bachelor bzw. Diplom, Bachelor-Studiengänge mit Schwerpunkt Physik, Lehramt Physik (Staatsexamen bzw. Bachelor).

<sup>2</sup>Die Daten basieren auf den jährlichen Ermittlungen der Konferenz der Fachbereiche Physik ([www.kfp-physik.de](http://www.kfp-physik.de); 03/2016).

<sup>3</sup>Schlagwörter sind z.B. *Predict-Observe-Explain Aufgaben* oder *Peer-Instruction* (Crouch & Mazur, 2001).

## 1.1 Ausgangspunkte

### Hohe Leistungsansprüche und mangelnder Anwendungsbezug in Physikstudiengängen

Die oben genannte hohe Schwundquote bei Physikstudiengängen gab Anlass zur empirischen Forschung nach Ursachen für den Studienabbruch. Eine der wichtigsten Arbeiten in diesem Gebiet stammt von Albrecht (2011), der den Einfluss relevanter Merkmale auf den Studienerfolg und -abbruch modellgeleitet untersuchte. Den Ergebnissen zufolge liegt für die meisten Studenten<sup>4</sup> der zentrale Grund für einen Studienabbruch in dem hohen fachlichen Leistungsanspruch. Die Schwierigkeit des Fachinhalts stellt hohe zeitliche und kognitive Ansprüche an die Studierenden, infolgedessen eine große Belastung im Physikstudium wahrgenommen wird. Hinzu kommt die Neuorientierung als Universitätsstudent betreffend Änderungen der Lebensbedingungen (z.B. Wohnortwechsel), sozialer Integration in den universitären Lebensraum sowie erhöhter Verantwortung (z.B. Zeit-Management, Studienfinanzierung). Zwei weitere wichtige Befunde der Studie sind, dass ein Fachwechsel am häufigsten in physik-affine und praktisch-anwendungsorientierte Studiengänge stattfindet und dass die Entscheidung über eine Exmatrikulation häufig bereits im ersten Studienjahr gefällt wird.

### Übergangsproblematik zwischen Schule und Hochschule

Aus dieser Problematik ergibt sich die Forderung an die Physikdidaktik, geeignete Lehr-/Lernmethoden für diese noralgische Phase zwischen Schule und Hochschule zu finden, um Studierende beim Lernen physikalischer Inhalte zu unterstützen. Oben genannte Befunde zeigen, dass der Bedarf an lernförderlichen Maßnahmen während der Studieneingangsphase besonders groß ist. In dieser Phase bilden die experimentalphysikalischen Grundlagenveranstaltungen — Experimentalphysik 1: Mechanik und Thermodynamik sowie Experimentalphysik 2: Elektrodynamik und Optik — mit jeweils 4 SWS (Semesterwochenstunden) Vorlesung und 2 SWS Übung zusammen mit Laborpraktika die Hauptveranstaltungen des ersten Grundmoduls, das deutschlandweit eine vergleichbare Organisationsstruktur aufweist. Im Gegensatz zu dem Physikunterricht der Schule sind die Vorlesungen zwingend theorielastiger und unterliegen einem an straffen Zeitvorgaben gerichteten Ablaufplan. Außerdem nimmt das Experiment eine andere Rolle beim Lernen ein: Während in der Schule der Schwerpunkt auf dem Experiment und der induktiven Erarbeitung von Inhalten liegt, lehr-/lernsituationsabhängige Wechsel zwischen Theorie und Experiment möglich sind, und Lernende nach Möglichkeit eine aktive Rolle beim eigenständigen Experimentieren einnehmen (Schülerexperimente), wird das Experiment in der Vorlesung meist zu Veranschaulichungszwecken genutzt (Demonstrationsexperimente). Der straffe Zeitplan ermöglicht es kaum, die Experimente zur Erarbeitung von Gesetzmäßigkeiten oder Theorien zu nutzen; die Verarbeitungstiefe beim Lernenden ist gering. So verwundert es nicht, dass sich Studierende nach kurzer Zeit nur unzureichend an Demonstrationsexperimente erinnern können und keine praktisch bedeutsame Lernwirkung stattfindet (Crouch et al., 2004; Miller et al., 2013). Während die Lernaktivität in traditionellen Vorlesungen auf Zuhören und Beobachten begrenzt ist, findet die eigentliche Informationsverarbeitung hauptsächlich in den vorlesungsbegleitenden Übungen statt. Dort sind die Lernenden selbst aktiv und wenden die physikalischen Inhalte beim Aufgabenlösen an. Die Übungen tragen wesentlich zur Wissenskonstruktion bei und sind für das Verstehen von Physik essenziell (Finkelstein & Pollock, 2005). Deshalb ist es kritisch zu sehen, dass die Übungen in der Regel kaum experimentelle Anteile enthalten, sondern in den meisten Fällen an stark abstrakten, anwendungsfernen Inhalten ausgerichtet sind und zu sehr auf Mathematik fokussieren. Innovationsansätze, die diesen Problemen zu entgegnen versuchen, dürfen keinen Abstrich am fachlichen Anspruch in Kauf nehmen, um die Anschlussfähigkeit

<sup>4</sup>Auch wenn im Text an einigen Stellen der besseren Lesbarkeit wegen die männliche Form „Studenten“ bzw. „Schüler“ verwendet wird, ist stets auch die weibliche Form (Studentinnen bzw. Schülerinnen) inbegriffen.

zu theoretischen Lehrveranstaltungen zu gewährleisten.

### **Videobasierte Aufgaben — Ein Lösungsansatz**

Vor diesem Hintergrund wurde im Sommer 2012 das Projekt „*physics.move (Videobasierte Aufgaben in den klassischen Übungen zur Experimentalphysik 1)*“ am Fachbereich Physik der Technischen Universität Kaiserslautern initiiert. Das Projekt begegnet den Schwierigkeiten der Lernenden in der Studieneingangsphase (insbesondere im ersten Fachsemester) auf mehreren Ebenen: Die anschauliche Methode der Videoanalyse wird als unterstützendes Medium zum Lösen physikalischer Probleme eingeführt. Als Ergänzung zu bereits etablierten traditionellen Aufgabenstellungen unterstützen *Videoanalyse-Aufgaben* (V-Aufgaben) das Verständnis physikalischer Prinzipien durch eine Verbindung zwischen experimentellen und theoretischen Aktivitäten im Problemlösungsprozess. Die erhöhte Anschaulichkeit der Videoexperimente liefert einen leichteren Zugang zur Aufgabenstellung und fördert durch qualitative Aspekte die Begriffs- und Theoriebildung. Es kann davon ausgegangen werden, dass die Methode der Videoanalyse vielen Studierenden aus ihrer Schulzeit bekannt sein dürfte, denn sie nimmt seit über 20 Jahren Einzug in die Schulpraxis (Klingerl, 1991).

Des Weiteren nutzen die Studierenden zur Aufgabebearbeitung mobile Endgeräte und führen eigenständig Freihandexperimente durch, wodurch die Attraktivität des Studiengangs gesteigert wird (*mobile Videoanalyse-Aufgaben*, mV-Aufgaben). Diese Aufgaben eignen sich in besonderem Maße, schon frühzeitig experimentelle Anteile ins Physikstudium einzubringen. Dies ist deshalb von Interesse, weil sich experimentelle Kompetenzen erfahrungsgemäß über Jahre hinweg kontinuierlich aufbauen müssen und nicht ausschließlich in physikalischen Praktika und Abschlussarbeiten verortet sein sollten (Sacher, 2014). In der Tat ist ein Verständnis der Grundprinzipien physikalischer Arbeitsmethoden und des Experiments samt Anwendung, Einordnung und Beurteilung von Analysemethoden ein explizites Lernziel experimentalphysikalischer Grundlagenveranstaltungen (vgl. KFP, 2010). Beim Einsatz videobasierter Aufgaben wird nicht der Anspruch erhoben, physikalische Praktika abzulösen oder alle Fähigkeiten zu schulen, die ein forschender Physiker im Labor für wissenschaftliches Arbeiten benötigt. Vielmehr bereiten die Aufgaben dieses Ziel vor und haben das Potential, das Lernen und Verstehen von Physik mittels technischer Möglichkeiten zu unterstützen sowie Interesse an der Fachdisziplin durch abwechslungsreiche und authentische Problemstellungen zu wecken.

In Bezug auf die Typologie von Innovationen physikalischer Einführungsveranstaltungen von Ruiz-Primo et al. (2011) lässt sich der gewählte instruktionsbasierte Lehr-/Lernansatz in die Kategorie *konzeptorientierter, mediengestützter* Ansätze verorten. Diese verfolgen die Ziele, prozedurales Wissen über fachspezifische Methoden, Problemlösestrategien sowie Schlüsselkonzepte zu vermitteln und dieses auf realitätsnahe Problemstellungen anzuwenden. Die Lernmedien helfen bei der Visualisierung physikalischer Phänomene und bieten die Möglichkeit zur eigenständigen (experimentellen) Auseinandersetzung mit den Lerninhalten zur Förderung eines flexiblen Konzeptverständnisses. Die formulierten Ziele des Ansatzes werden im Rahmen einer empirischen Studie untersucht und in bestehende Forschungsergebnisse eingeordnet.

### **Forschungsinteresse**

Konkret wird die Wirksamkeit von videobasierten Aufgaben gegenüber traditionellen Aufgaben bzgl. verschiedenen kognitiven und affektiven Variablen in einem Versuchs-Kontrollgruppen Design (mit Rotation der Gruppen) mit drei Messzeitpunkten untersucht. Dabei finden Kontrollvariablen wie Vorleistungen in Mathematik, Physik oder die Abiturnote ebenso Berücksichtigung wie soziodemographische Angaben. Darüber hinaus wird die Beanspruchung der Studierenden durch das Studium im Allgemeinen, die empfundene Aufgabenschwierigkeit und die notwendige Zeit zum Bearbeiten der Übungsaufgaben erhoben.

Neben der Wirksamkeitsuntersuchung der neuen Aufgabenformate besteht ein wesentliches Ziel dieser Arbeit in der Konstruktion und Validierung geeigneter Instrumente, um den Interventionserfolg einschätzen zu können. Valide Testinstrumente zur Erhebung fachspezifischer Kompetenzen und motivationaler Faktoren liefern Erkenntnisse über den Einfluss der gemessenen Variablen auf den Studienerfolg (Verbleib im Studium, Klausurnote, Motivation) und liefern Informationen über den Zusammenhang zwischen kognitiven und affektiven Variablen. Diese Daten sind von allgemeinem Interesse der Forschung über die Eingangsphase von Physikstudiengängen, denn sie liefern Erkenntnisse über die Studienfachwahl, den Studienerfolg und -abbruch.

Darüber hinaus wird die Forschungslage bzgl. der Videoanalyse als Physikmedium mit Evidenzen angereichert. Obwohl die Videoanalyse seit über 20 Jahren von Physikdidaktikern als Lehr-/Lernmedium genutzt wird (vgl. z.B. Klingerl, 1991), gibt es nur wenige empirisch geprüfte belastbare Aussagen zur Lernwirkung. Dabei bildet die Videoanalyse eine zentrale Gelegenheit, verschiedene physikalische Darstellungsformen zueinander in Bezug zu setzen und zielgerichtet zum Problemlösen zu nutzen. Diese Fähigkeiten spielen in der Physik eine entscheidende Rolle und die damit verbundene latente Variable (Repräsentationskompetenz, vgl. De Cock, 2012) gehört zum Kern der hier untersuchten Forschungsfragen. Entgegen Ideologie und Bauchgefühl muss es Ziel einer evidenzbasierten, physikdidaktischen Forschung sein, Hypothesen mit klar definierten Verfahren zu prüfen, um Lehrkräften und Dozenten klare Leitlinien und Handlungsbotschaften mitzuteilen. Nach dem Vorbild der evidenzbasierten Medizin soll es auch Ziel der Physikdidaktik sein, wissenschaftliche Erkenntnisse durch ordentliche Vergleichsstudien zu generieren und Einzug in die „tägliche Behandlungspraxis“ finden zu lassen.

## 1.2 Aufbau der Arbeit

Die vorliegende Arbeit ist in drei Teile untergliedert, welche die drei inhaltlichen Leitlinien „Forschungsbasierte Entwicklung von Lehr-/Lern-Materialien“, „Konzeption und Validierung der Instrumente“ und „Wirksamkeitsuntersuchung“ aufbauend abbilden.

Im **ersten Teil** (Kap. 2–4) werden zunächst die Möglichkeiten der Videoanalyse als Lern- und Analysewerkzeug dargestellt. Der durch Anwendungsbeispiele vermittelte Eindruck des lernförderlichen Potentials der Videoanalyse wird auf eine lerntheoretische Grundlage gestellt. Obwohl die Wirksamkeit der Videoanalyse dadurch gut begründet werden kann, existieren hierzu nur wenige empirische Studien. Diese werden zusammengefasst, bezüglich der verwendeten Methodik kritisch bewertet und als Ausgangspunkt für eigene Fragestellungen genutzt. Schließlich wird gezeigt, wie eine Implementation der Videoanalyse in Form von Aufgaben im instruktionalen Rahmen der Übungen zur Experimentalphysik 1 (erstes Studiensemester) erfolgt. Dazu werden V- und mV-Aufgaben definiert und gegenüber traditionellen Aufgaben abgegrenzt. Der erste Teil schließt mit der Darstellung einer Pilotstudie ab, die zur erstmaligen Erprobung der Aufgaben und des Untersuchungsdesigns genutzt wurde.

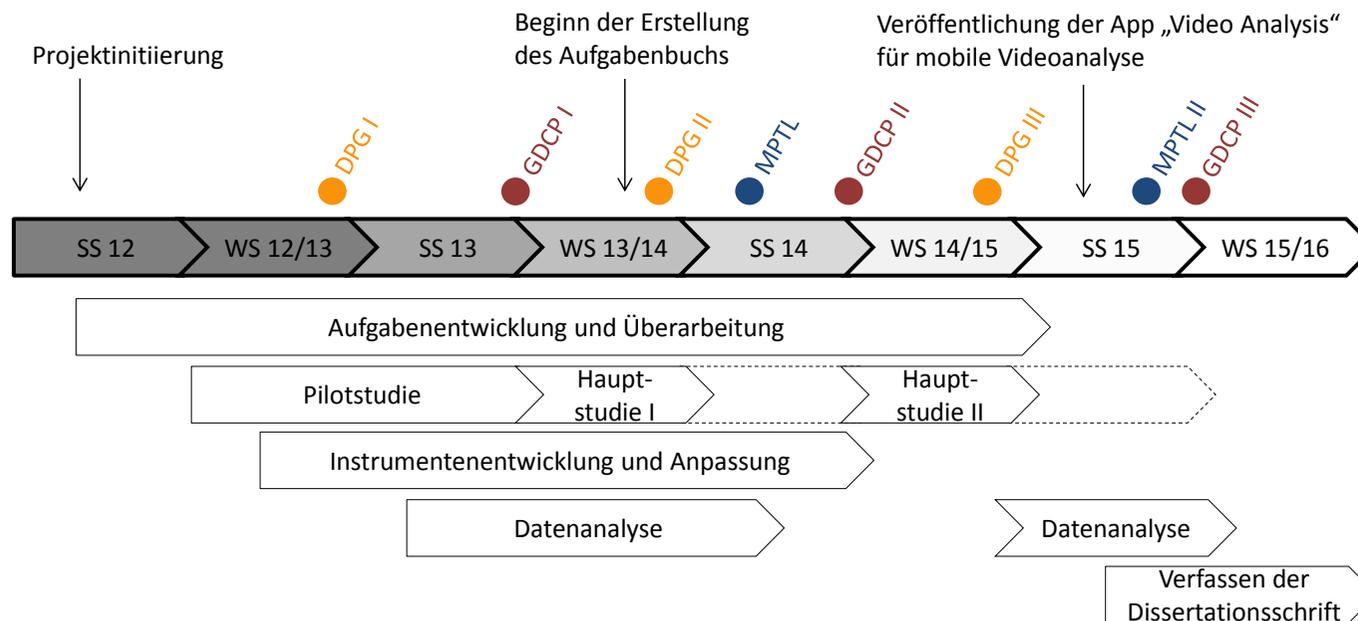
Der **zweite Teil** der Arbeit (Kap. 5–8) beinhaltet die theoriegeleitete Konzeption von Fragebögen und Tests zur Einschätzung des Interventionserfolgs bzgl. Motivation, Leistung, Neugierde und Belastung. Alle Instrumente werden auf Basis einer Eich-Stichprobe nach klassischen Verfahren der Testanalyse psychometrisch validiert. Die Leistungstests werden darüber hinaus einer Rasch-Analyse unterzogen, die stichprobenunabhängige Erkenntnisse über einzelne Testitems, Aussagen zur Passung des Tests auf die zugrundeliegende Stichprobe und schließlich Optimierungsmöglichkeiten der Instrumente liefert. Die Leistungstests wurden als *two-tier*-Tests konstruiert, d.h. dass Studierende zu ausgewählten Aufgaben die Zuversicht in ihre Lösung ausdrücken können. Die Auswertungen dieser Antwortsicherheiten in Bezug zur eigentlichen Leistungsfähigkeit erlaubt Aussagen über metakognitive Fähigkeiten von Studierenden und über das Auftreten von Fehlkonzepten.

Der **dritte Teil** der Arbeit (Kap. 9–12) umfasst die Hauptstudie und beginnt mit der Darstellung der Hypothesen und Forschungsfragen. Anschließend werden die Untersuchungs-

methoden ausführlich beschrieben. Dazu gehört die Charakterisierung der Stichprobe, des Instruktionmaterials, des Studienverlaufplans, der Instrumente und der genutzten quantitativen statistischen Verfahren. Es schließen die Ergebnisse und deren Diskussion an.

### 1.3 Projektverlaufsplan

Abbildung 1 zeigt eine Übersicht über den zeitlichen Verlauf des Projekts *physics.move* und hebt nationale und internationale Tagungen hervor, auf denen sowohl die konzeptionelle Idee des Projekts (gemeint ist die Implementation videobasierter Aufgabenformate in die Übungen zur Experimentalphysik 1) als auch die empirischen Befunde in Vorträgen dargestellt und diskutiert wurden. Die Pilotierungsphase verortet sich im Wintersemester 2012/2013 sowie im Sommersemester 2013. Die beiden Hauptstudien knüpften direkt an die Pilotierungsphase an. Im Sommersemester 2014 wurde — ebenso wie in den darauffolgenden Sommersemestern — keine Studie durchgeführt. Zwar wurden videobasierte Aufgaben in den Übungen eingesetzt, es fand aber aufgrund geringer Studierendenzahlen keine Datenerhebung statt. In dem Projektverlaufsplan sind zwei Meilensteine hervorzuheben: Die von der Physikdidaktik-Arbeitsgruppe der TU Kaiserslautern entwickelten und erprobten videobasierten Aufgaben werden in Form eines Aufgabenbuches veröffentlicht, wodurch der Nutzerkreis der Lernenden erweitert und die Implementation des Ansatzes in andere Physik-Fachbereiche vereinfacht wird. Die Entwicklung dieses Buches begann im Oktober 2013; der voraussichtliche Erscheinungstermin liegt im Sommer 2017. Wie in Kap. 2.1.5 beschrieben wird, wurde parallel zu dieser Arbeit eine Videoanalyse-Applikation für mobile Endgeräte entwickelt, die seit Sommer 2015 erhältlich ist.



**Abb. 1:** Zeitverlauf der Arbeitsphasen und Vorstellung des Projekts auf Tagungen: DPG = Deutsche Physikalische Gesellschaft (Frühjahrstagungen des Fachverbands Didaktik der Physik), GDCP = Gesellschaft der Didaktik für Chemie und Physik, MPTL = Multimedia Physics in Teaching and Learning (Tagung der European Physical Society).

Teil I

**Videoanalyse: Theoretischer  
Hintergrund und Stand der  
Forschung**



## Kapitel 2

# Videoanalyse als Physikmedium

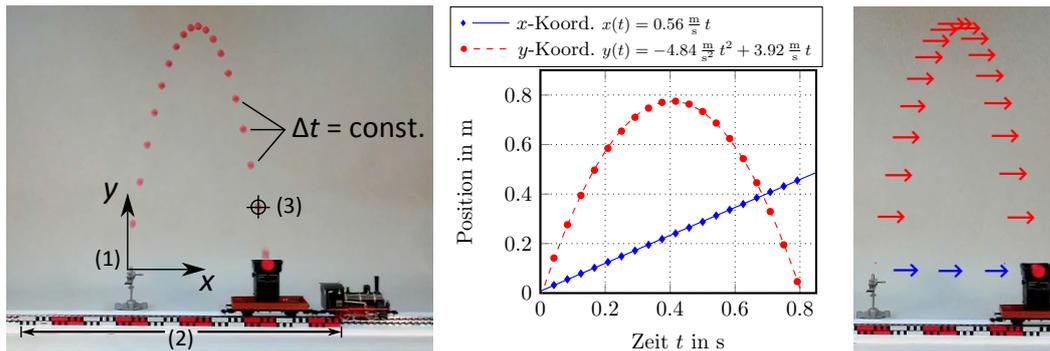
Dieses Kapitel beschreibt zunächst das messtechnische Prinzip der Videoanalyse und zeigt die vielfältigen Einsatzmöglichkeiten im Physikunterricht und -studium auf (Kap. 2.1). Die kontinuierlich steigende Anzahl nationaler und internationaler Publikationen in diesem Bereich lässt auf ein anhaltendes didaktisches Interesse an der Methode schließen, das zur Entwicklung zahlreicher Experimente führte. Mithilfe ausgewählter Beispiele wird demonstriert, dass ein wesentlicher Teil der Experimentalphysik 1-Lehrinhalte durch Videoexperimente aufgegriffen werden kann.

Die klassische Variante des Verfahrens erfordert mehrere Hard- und Softwarekomponenten, um Experimente aufzunehmen, auf einen Computer zu übertragen und schließlich auszuwerten. Mobile Endgeräte wie Smartphones oder Tablet-PCs hingegen ermöglichen die Durchführung dieser Schritte mithilfe eines einzigen Gerätes (*mobile Videoanalyse*). In Kap. 2.2 wird die Nutzung mobiler Endgeräte für die Videoanalyse motiviert, die daraus resultierenden didaktischen Möglichkeiten erarbeitet und durch selbst erstellte Beispiele gestützt.

Diese beiden Unterkapitel verdeutlichen, wie die Videoanalyse als Lernmedium und Analysewerkzeug von Bewegungen genutzt werden kann, und bilden die konzeptionelle Grundlage der vorliegenden Arbeit: Anknüpfend an die Messprinzipien sowie den experimentellen Möglichkeiten werden in den darauffolgenden Kapiteln die fachdidaktischen und lerntheoretischen Grundlagen aufbereitet (Kap. 3) und das Instruktionsmaterial definiert (Videoanalyse-Aufgaben, Kap. 4).

### 2.1 Messprinzip und Anwendungen in der Lehre

Der Videoanalyse von Bewegungen liegt die Idee zugrunde, dass aufgenommene Experimentiervideos die Ort- und Zeitinformationen von bewegten Objekten beinhalten, die sich bezüglich physikalischer Fragestellungen auswerten lassen. Unter Verwendung von geeigneter Computersoftware werden digitale Videos in Einzelbilder zerlegt, in denen die Ortsvektorkomponenten bezüglich eines zweidimensionalen Koordinatensystems gemessen werden können. Zur Illustration des Verfahrens dient folgendes Experiment: Eine Modelleisenbahn fährt mit konstanter Geschwindigkeit in horizontale ( $x$ -)Richtung und schießt während der Fahrt eine Kugel mithilfe eines Schussapparates, der sich mit der Modelleisenbahn bewegt, senkrecht nach oben ab (in  $y$ -Richtung). Die Kugel beschreibt eine Wurfparabel und landet wieder in dem Wagen. Abbildung 2 stellt einen Ausschnitt des Experimentiervideos (links) sowie die aus dem Video gemessenen Ort-Zeit-Daten der Kugel dar (mittig). Das Verständnis dieses „Wagen-Kugel-Experiments“ bereitet Schülerinnen und Schülern sowie Studierenden im ersten Studienjahr gleichermaßen Schwierigkeiten (Dilber et al., 2009) und dient in diesem wie auch im nächsten Kapitel in einer Variationsform der Veranschaulichung des Messprinzips und illustriert das verständnisfördernde Potential der Videoanalyse.



**Abb. 2:** Prinzip der Messdatenerfassung durch Videoanalyse (links). Dargestellt ist ein Screenshot eines Experimentiervideos mit Stroboskop-Aufnahme der Kugelbewegung. Zwischen den Kugel-Positionen liegen jeweils gleiche Zeitabstände  $\Delta t$ . Die Position der Kugel wird in Bezug zu einem zweidimensionalen Koordinatensystem (1) unter Berücksichtigung eines Längenmaßstabs (2) durch Objektmarkierung (3) in einzelnen Videobildern erfasst. Die Ort-Zeit-Daten sind komponentenweise mit den jeweiligen Ausgleichskurven dargestellt (mittig). Das rechte Teilbild zeigt die  $x$ -Komponenten der Geschwindigkeitsvektoren von Kugel und Wagen zu ausgewählten Zeitpunkten.

### 2.1.1 Längen- und Zeitskalierung des Videos

Die Skalierung eines Experimentiervideos besteht in der Regel aus drei Schritten: In einem ersten Schritt wird der zeitliche Abstand zwischen aufeinanderfolgenden Einzelbildern zur Zeitskalierung verwendet. Unter der Annahme, dass sich die Aufnahmezeit der Kamera (*Framerate*) während der Videoaufnahme nicht ändert, ist der zeitliche Abstand zwischen aufeinanderfolgenden Videobildern äquidistant und beträgt typischerweise 1 ms (High-Speed-Aufnahmen) bis 30 ms (handelsübliche Digitalkameras). In einem zweiten Schritt wird eine Referenzlänge in der Bewegungsebene des Objekts ausgemessen und zur Umrechnung von Bildschirmmaßeinheiten (Pixel) in reale Längen genutzt. Schließlich wird ein (zweidimensionales) Koordinatensystem bzgl. Lage und Orientierung festgelegt, in dem die Bewegung beschrieben wird. In Abb. 2 sind die genannten drei Skalierungselemente (Zeitabstände, Maßstab und Koordinatensystem) abgebildet. An die Skalierung schließt sich die Datenaufnahme durch die Identifikation des Objekts in einzelnen Videobildern an, wie im nächsten Abschnitt beschrieben wird.

### 2.1.2 Messdatenerfassung

Die Ortserfassung erfolgt durch das Fixieren eines Cursors an einer ausgezeichneten Stelle des Objekts (z.B. Flächenschwerpunkt) in einzelnen (typischerweise aufeinanderfolgenden) Videobildern. Es bietet sich an, den zu verfolgenden Punkt farblich hervorzuheben, beispielsweise durch Anbringen eines Markers. Einige Videoanalyse-Programme verfügen über eine automatische Objektverfolgung, die Farb- und Bewegungsanalysen zur Objekterkennung nutzen (Riera et al., 2003). Verglichen mit der manuellen Analyse verkürzt das automatische Verfahren die benötigte Zeit zur Datengewinnung und beugt einer Überlastung der Nutzer durch die repetitive und monotone Messwerterfassung vor. Dies stellt vor allem bei langen Experimentiervideos und High-Speed-Aufnahmen mit hohen Bildraten einen Vorteil gegenüber der manuellen Messwerterfassung dar. Es besteht ein anhaltendes Bemühen zur Steigerung der Zuverlässigkeit und Effizienz der automatischen Objekterkennung (Usman et al., 2015).

Neben Erfassung der Ort-Zeit-Daten können nach einer Skalierung andere Größen wie Längen, Flächen und Winkel direkt im Video gemessen werden.

### 2.1.3 Verarbeitung und Darstellung der Messdaten

Der Begriff Videoanalyse beschränkt sich im Verständnis der Anwender nicht auf bloße Erhebung von Messwerten aus einem Video, sondern umfasst auch deren Darstellung und physikalische Analyse. Die heutzutage erhältlichen Videoanalyse-Programme<sup>1</sup> verknüpfen das Messprinzip mit einer Programmumgebung, die verschiedene Funktionalitäten zur Weiterverarbeitung oder Darstellung der Daten anbietet, um das Lernen zu unterstützen. Beispielsweise bieten die meisten Programme an, Geschwindigkeiten und Beschleunigungen automatisch aus den Ort-Zeit-Daten zu berechnen (Differenzenquotienten). Integrierte Tabellenkalkulationen geben dem Lernenden die Möglichkeit, aus den Primärdaten weitere Größen mathematisch zu bestimmen (z.B. Energien, Kräfte). Die gemessenen oder mathematisch bestimmten (kurz: abgeleiteten) Größen lassen sich — je nach Darstellungsfunktionalitäten des Programms — in verschiedener Weise visualisieren. Beispielsweise wird durch die Einblendung der Geschwindigkeitsvektoren im Wagen-Kugel-Experiment eine Veranschaulichung des Superpositionsprinzips erreicht, wodurch der Ausgang des Experiments qualitativ verstanden werden kann: Kugel und Wagen bewegen sich zu jeder Zeit mit der gleichen horizontalen Geschwindigkeit, siehe Abb. 2 (rechts).

Durch die zeitsynchrone Verbindung verschiedener Repräsentationsformen (z.B. Vektorpfeile oder Diagramme) mit dem realen Bewegungsablauf ergeben sich wesentliche Mehrwerte zur Förderung des Verständnisses, die in Kap. 3 diskutiert werden. Die umfangreichen Darstellungsmöglichkeiten sind in Tab. 1 aufgelistet und werden z.B. in Suleder (2010) ausführlich beschrieben. Die Tabelle gibt dem Leser einen Überblick über die Skalierungs- und Analysefunktionen erhältlicher Programme, die nach einer intensiven Auseinandersetzung mit mehreren Videoanalyse-Programmen erfahrungsbasiert zusammengetragen wurden.

**Tab. 1:** Funktionalitäten von Videoanalyse-Programmen zur Realisierung des Messprinzips

Analyseschritt	Funktionalitäten
Videoskalierung	Definition von Koordinatenursprung und Achsenorientierung, Längen- und Zeitskalierung, Wahl eines Zeitnullpunktes, Einschränkung des Analysebereichs
Datenerfassung	Objektverfolgung (manuell / automatisch), Messpunkte modifizieren (löschen, neu setzen), Schrittweite festlegen (z.B. nur jedes $n$ -te Bild analysieren), Längen- und Winkelmessung
Datenverarbeitung	Automatische Berechnung von abgeleiteten Größen, manuelle Rechenoperationen (Tabellenkalkulation), Ausgleichsrechnungen und Statistiken
Darstellung	Stroboskopbilder, Serienbilder, Diagramme, symbolische Einblendungen (z.B. Positionsmarker, Vektoren)
Sicherung	Absichern von Projekten und Daten, Import und Export

### 2.1.4 Messfehler und Vermeidung von Messfehlern

Wie jedes physikalische Messverfahren ist die Videoanalyse messfehlerbehaftet. Die primäre statistische Messfehlerquelle ist auf die unsorgfältige Markierung des Objekts zurückzuführen (manuelle Analyse), die aber durch Verfügbarkeit automatischer Analysemethoden und dem Nutzen von Markern vernachlässigt werden kann (siehe Kap. 2.1.2). Die systematischen Fehlerquellen, die einen Einfluss auf die Genauigkeit der Ortsmessung nehmen, lassen sich nach Page et al. (2008) einerseits auf die Videoaufnahme (1–2), andererseits auf die Auswertungsmethode (3) zurückführen:

<sup>1</sup>Internationaler Standard ist die kostenlose Software „Tracker“, die in den meisten internationalen Publikationen als Referenz angegeben wird; im deutschsprachigen Raum sind „Coach“ und „measure Dynamics“ zu nennen. Auf die Unterschiede zwischen den Programmen wird in dieser Arbeit nicht eingegangen.

- (1) Die Bewegungsebene des Objekts und die Kameraebene sind nicht parallel.
- (2) Das Videobild wird durch das optische System der Kamera verzerrt / unscharf.
- (3) Die räumliche Auflösung des Pixelrasters ist endlich und beschränkt die Positionsbestimmung des Objekts im Video.

Experimentell können die Fehler dadurch reduziert werden, dass (1) ein ausreichend großer Abstand zwischen Kamera und Experiment eingehalten und die Kameraachse senkrecht zur Bewegungsebene ausgerichtet wird. Außerdem sollte (2) das Experimentierszenario ausreichend (und möglichst diffus) beleuchtet sein, um geringe Verschlusszeiten der Kamera zu ermöglichen. Page et al. (2008) gehen auf Kalibrierungsverfahren ein, um perspektivische Verzerrungen wegen des optischen Kamerasystems zu korrigieren und weisen experimentell nach, dass schon mit handelsüblichen Webcams Positionsbestimmungen mit einer relativen Genauigkeit unter 0.1% möglich sind.

Über die genannten Fehlerquellen hinaus führt Suleder (2010) noch weitere Punkte auf, die durch eine sorgfältige Videoaufnahme vermieden werden können (z.B. Vermeidung von Spiegelungen und Schatten im Bild). Wird eine ideale Videoaufnahme vorausgesetzt, kann eine Abschätzung des Ortsmessfehlers (3) aus der Messgenauigkeit am Bildschirm erfolgen: Bei der Skalierung mittels eines Referenzmaßstabs werden die Bildschirm-Längeneinheiten (Pixel) in reale Längen übersetzt (vgl. Kap. 2.1.1). Durch die Diskretisierung des Pixel-Rasters liegt der Messfehler also bei mindestens einem Pixel. Falls Bewegungsunschärfe auftritt, sollte die räumliche Ausdehnung des Objekts in jedem Frame vermessen und zur Angabe des Messfehlers verwendet werden (siehe Praxisbeispiel in Smith et al., 2010).

Für die Zeitmessung kann ein relativer Fehler von 1% angenommen werden (Suleder, 2010). Die Wahl der zeitlichen Auflösung (Framerate) wird primär dadurch bestimmt, auf welcher Zeitskala die wesentlichen Veränderungen im Experiment stattfinden. Werden aus den Ort-Zeit-Daten andere Größen abgeleitet, übertragen sich die Messfehler auf diese Größen. In diesem Zusammenhang wird empfohlen, große Messstrecken zur Bestimmung von möglichst genauen (Durchschnitts-)Geschwindigkeiten zu nutzen, da auf diese Weise kleinere relative Fehler bei der Bestimmung von Ort-Differenzen entstehen (Suleder, 2010, S. 18).

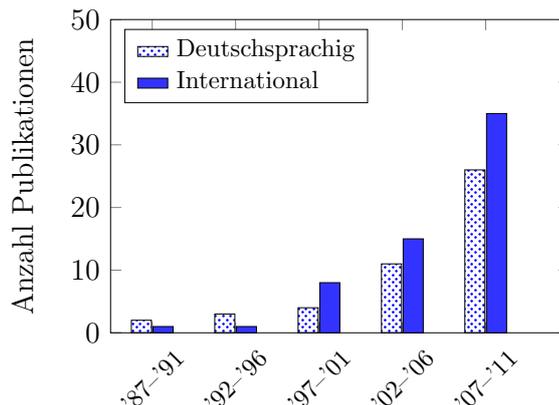
Wird das Experiment bei der Videoaufnahme derart in Szene gesetzt, dass die aufnahmebedingten Fehlerquellen reduziert werden, dann stellt die Videoanalyse ein berührungsfreies Messverfahren dar, das mit einfachen und kostengünstigen Mitteln bemerkenswert genaue Messungen erzielt.

### 2.1.5 Literaturrecherche zu Einsatzmöglichkeiten

In der nationalen und internationalen fachdidaktischen Literatur und in Lehrerzeitschriften finden sich zahlreiche Beispiele zum Einsatz der Videoanalyse. In diesem Abschnitt werden zunächst die Ergebnisse einer Literaturrecherche über den konzeptionellen Entwicklungsstand zusammengefasst. Im Anschluss wird aus dieser Bestandsaufnahme eine Auswahl getroffen, die die Bandbreite der Anwendungsmöglichkeiten aufzeigen soll. Mit Blick auf die inhaltliche Verortung dieser Arbeit werden zunächst Beispiele aus der Experimentalphysik 1 mit Einschränkung auf die *Mechanik* genannt und anschließend Berührungspunkte zu weiteren Themengebieten aufgezeigt.

#### Vorgehen und Ergebnisse der Literaturrecherche

In die Literaturrecherche wurden die referierten internationalen Zeitschriften *The Physics Teacher*, *Physics Education*, *European Journal of Physics* und *American Journal of Physics* im Zeitraum von 1987–2015 einbezogen. Auf den Online-Portalen dieser Zeitschriften wurde nach den Schlagwörtern „video analysis“, „video experiment“, „motion analysis“, „interactive digital video“, „motion tracking“ und Kombinationen daraus gesucht, und die gefundenen Beiträge anschließend auf Sekundärliteratur überprüft. Im deutschsprachigen



**Abb. 3:** Anzahl nationaler und internationaler Publikationen in ausgewählten Fachzeitschriften zu experimentellen Beispielen der Videoanalyse über die Zeit.

Raum wurden die Lehrerzeitschriften *Praxis der Naturwissenschaften — Physik in der Schule*, *Mathematisch Naturwissenschaftlicher Unterricht*, *Physik in unserer Zeit*, *Unterricht: Physik* sowie die Beiträge zur DPG-Frühjahrstagung zur Physikdidaktik (*PhyDid B*) von 1987–2011 berücksichtigt. Diese Zeitschriften wurden manuell nach den Stichwörtern „Bewegungsanalyse“, „Videoanalyse“ und „Videoexperiment“ durchsucht. Es wurden alle Beiträge berücksichtigt, in denen die Videoanalyse als zentrale Messmethode zur Datengewinnung in Experimenten genutzt wird. Eine vollständige Liste der Artikel befindet sich im digitalen Anhang.

In Abb. 3 ist die Anzahl der Veröffentlichungen über den Jahren von 1987–2011 dargestellt. Wie zu sehen ist, nimmt das fachdidaktische Interesse an dem Medium im Laufe der Zeit zu, was vor allem auf die voranschreitende Digitalisierung und auf fallende Preise für technische Ausstattung zurückzuführen ist. In dem frühen Entwicklungsstadium des Verfahrens wurde entweder auf die Auswertung von Standbildern zurückgegriffen (Klingerl, 1991) oder eigene Computer-Schnittstellen zur Aufbereitung der elektrischen Kamerasignale entwickelt, um die Ort-Zeit-Daten zu erzeugen (Dengler & Luchner, 1990). Die im Diagramm dargestellten Daten werden durch 65 weitere internationale Artikel im Zeitraum von 2012–2015 ergänzt, was durchschnittlich etwas mehr als einem Beitrag pro Monat entspricht. Neben dem anhaltenden Interesse an dem Verfahren ist ersichtlich, dass sich die Videoanalyse seit Beginn an im deutschsprachigen Raum etablierte, und dass die Anzahl deutscher Publikationen mit der Anzahl internationaler Publikationen vergleichbar ist. Da die Videoanalyse in einigen Bundesländern seit Kurzem auch explizit in den Lehrplänen der gymnasialen Oberstufe aufgeführt wird, ist auch weiterhin ein nachhaltiges Interesse an dem Medium zu erwarten.

Personen, die essenziell zur Entwicklung der Videoanalyse im deutschsprachigen Raum beitragen, sind Thomas Wilhelm (Habilitation über die Möglichkeiten der Videoanalyse im Physikunterricht; Wilhelm, 2011a), Michael Vollmer und Klaus-Peter Möllmann (zahlreiche Beiträge zu Themen, die weit über die Mechanik hinausgehen; insbesondere Analyse von Hochgeschwindigkeitsaufnahmen; z.B. Vollmer & Möllmann, 2012). Im internationalen Raum sind die Arbeitsgruppen um Brown (innovative Experimente für die Schule; z.B. Brown & Cox, 2009) sowie um de Ambrosis, Mascheretti, Malgieri und Onorato zu nennen (komplexere Beispiele aus der Mechanik mit universitärem Anforderungsniveau; z.B. Malgieri et al., 2014).

Zusammenfassend lässt sich eine reichhaltige Entwicklungslage von Videoanalyse-Experimenten feststellen. Die meisten gesichteten Beiträge beschreiben einzelne Experimente oder kurze Experimentierserien zu einem gemeinsamen Thema. Zwei Beiträge verweisen auf eine Sammlung von Experimentiervideos, die Lehrkräften als Handreichung dient (Abbott, 2009; Dziarstek & Hilscher, 1998). Abgesehen von einer systematischen Zusammenstellung von Experimenten zur Abdeckung der Mechanik-Inhalte aus der eigenen Gruppe (Gröber et al.,

2014) wurden keine Anhaltspunkte für eine systematische Implementationsstrategie des Mediums in universitäre Lehrveranstaltungen gefunden, insbesondere nicht im deutschsprachigen Raum.

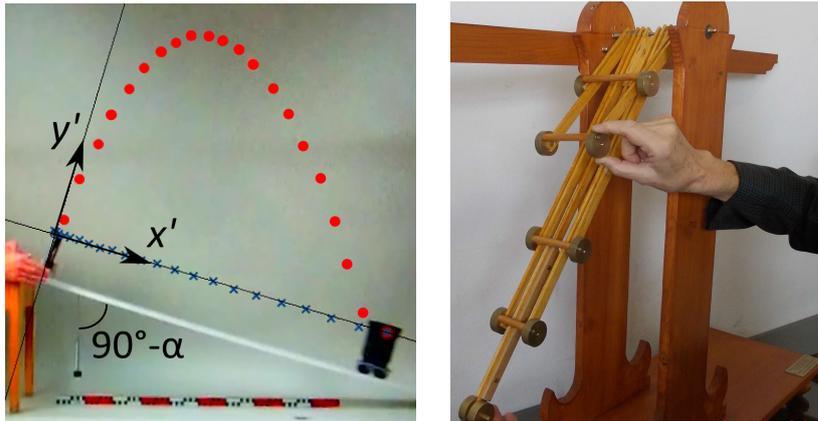
### **Videoanalyse in der Experimentalphysik 1 (Mechanik)**

In Tabelle 2 sind die Themengebiete der Experimentalphysik 1 aufgeführt, wobei die Unterteilung in Anlehnung an das Lehrbuch von Demtröder (2012) erfolgt. Auf die Kapitel Relativitätstheorie, Vakuum-Physik und Wärmelehre wurde verzichtet. Jedem Kapitel wurden einschlägige Beispiele bzw. Experimentierserien zugeordnet, um zu zeigen, dass die wesentlichen Lehrinhalte grundsätzlich durch Experimentiervideos abgedeckt werden können. Diese reichhaltige Grundlage spricht für eine systematische Materialentwicklung (Aufgabenstellungen, siehe Kap. 4). Bei der Zuordnung wurde jeweils ein Beispiel gewählt, das einen (akademisch motivierten) Laborversuch beschreibt, sowie ein Beispiel, das eine Anbindung an die Lebenswelt der Lernenden besitzt. Auf die Unterscheidung zwischen Experimenten mit und ohne Anwendungskontext wird in Kap. 3.1.2 eingegangen. Außerdem wurden vorhandene Experimentierserien zu bestimmten Themengebieten angegeben, die bauähnliche Versuche zu einem gemeinsamen Themengebiet beinhalten. Solche Serien wurden zum Teil in Staatsexamens- und Masterarbeiten konzipiert, die zeitlich parallel zur vorliegenden Arbeit in der Arbeitsgruppe angefertigt wurden. Inhaltlich sind die aufgeführten Beispiele in den meisten Fällen selbsterklärend und die Objekte, die videoanalytisch untersucht werden,

**Tab. 2:** Videoanalyse-Experimente zu universitären Mechanik-Inhalten

Inhalt <sup>a</sup> / Beispiel	Quelle
<u>Mechanik des Massenpunktes</u>	
(S) Drehimpuls des Massenpunktes	Spangenberg, 2015
(L) Atwoodsche Fallmaschine	Suleder, 2010
(K) Bungee-Springer	Heck et al., 2009
<u>Bewegte Bezugssysteme</u>	
(S) Trägheitskräfte	Oehme & Rieger, 1996
(L) Schuss von schiefer Ebene	Klein et al., 2015d
(K) Loopingfahrt	Michel & Wilhelm, 2010
<u>Systeme von Massenpunkte, Stöße</u>	
(S) Erhaltungssätze	Weidig, 2013
(L) Stöße zwischen Experimentierwagen	Wee, 2012
(K) Billard	Derby, 1999
<u>Dynamik starrer ausgedehnter Körper</u>	
(S) Winkelgrößen	Eadkhong et al., 2012
(L) Rollende Zylinder auf schiefer Ebene	Phommarach et al., 2012
(K) Karussell	Van Hook et al., 2007
<u>Reale feste und flüssige Körper, Gase</u>	
(S) Deformation von Körpern	Vollmer & Möllmann, 2011
(L) Block wird über beweglichen Wagen gezogen	Sullivan et al., 2006
(K) Aufstieg von Blasen in Sekt	Wilhelm, 2011b
<u>Strömende Flüssigkeiten und Gase</u>	
(L) Laminare und turbulente Strömungen	Kelley & Ouellette, 2011
(K) Rotierender Fußball (Magnus-Effekt)	Goff & Carré, 2009
<u>Mechanische Schwingungen und Wellen</u>	
(S) Schwingungen	Wehrbein, 2001
(L) Huygensches Pendel	Malgieri et al., 2014
(K) Schiffschaukel	Schüttler & Wilhelm, 2011

<sup>a</sup>Inhaltliche Gliederung nach Demtröder (2012). Angegeben sind Experimentierserien (S) sowie Versuchsbezeichnungen zu Labor- (L) und Kontextvideos (K).



**Abb. 4:** Zwei Beispiele für Videoexperimente zur Experimentalphysik 1 (Mechanik): Experimentelle Variation des Wagen-Kugel-Experiments aus Abb. 2 mit Positionsmarkierungen beider Objekte (links) und Apparatur zum Huygenschen Pendelversuch (rechts; Abb. entnommen aus Malgieri et al. 2014).

ergeben sich direkt aus der Versuchsbezeichnung. Zwei Experimente werden im Folgenden konkretisiert, denn sie werden im Laufe des nächsten Kapitels als Beispiele herangezogen, um die Möglichkeiten der Videoanalyse zur Lern- und Motivationsförderung zu veranschaulichen. Der Versuch „Schuss von der schiefen Ebene“ stellt eine Variation des Einstiegsbeispiels aus Abb. 2 dar. Die Modelleisenbahn bewegt sich antriebslos auf einer schiefen Ebene und schießt zu Beginn der Fahrt eine Kugel senkrecht zur Bewegungsrichtung — und damit senkrecht zur schiefen Ebene — ab (Klein et al., 2015d). Zum Erstaunen landet die Kugel, unabhängig vom Neigungswinkel der schiefen Ebene  $0 \leq \alpha < 90^\circ$ , wieder im Schussapparat (siehe Abb. 4, links). Mithilfe des Beispiels wird theoretisch und experimentell nachvollzogen, wie sich eine Rotation des Koordinatensystems auf Bewegungsdiagramme auswirkt.

Für den Versuch „Huygensches Pendel“ wurde das historische Gedankenexperiment von Huygens nachgebaut (Malgieri et al., 2014), bei dem mehrere Pendelkörper identischer Masse, aber unterschiedlicher Pendellänge durch eine Zwangsbedingung in der ersten Viertelschwingung gemeinsam und in der zweiten Viertelschwingung frei um einen gemeinsamen Aufhängepunkt schwingen, vgl. Abb. 4 (rechts). Dadurch wird erreicht, dass die äußeren Massen im Umkehrpunkt eine Höhe oberhalb ihrer Starthöhe erreichen, was dem Energieerhaltungssatz zu widersprechen scheint und deshalb die Neugierde der Beobachter weckt. Der Konflikt löst sich durch die videoanalytische Bestimmung der zeitlichen Energieverläufe aller Pendelmassen auf. Zusätzlich kann aus den Videodaten die Bewegung des Massenschwerpunkts auf einfache Weise veranschaulicht und mit der Theorie verglichen werden.

Beiden herausgestellten Beispielen ist gemeinsam, dass die Experimentiervideos als Impulsgeber kognitiver Konflikte dienen und die Auseinandersetzung mit dem Experiment und der Theorie anzuregen vermögen. In den Experimenten bewegen sich mehrere Objekte gleichzeitig. Dies führt zu einer Komplexität, die die Aufmerksamkeit bei einmaliger Betrachtung des Experiments (z.B. als Demonstration ohne Videoaufnahme) überfordert.

### Videoanalyse in anderen Themengebieten

Über die Mechanik hinaus findet die Videoanalyse in jüngster Zeit auch Einzug in andere Themengebiete, wie im Folgenden kurz skizziert wird.

- Belloni et al. (2013) nutzten einen videoanalytischen Zugang zu Experimenten in der *Astronomie*. Sie erstellten Zeitrafferaufnahmen des Nachthimmels, um die Länge eines siderischen Tages über die Messung von Zeiten und Winkelpositionen von Sternen zu

bestimmen. Ferner nutzten sie frei verfügbare Satellitenaufnahmen der Sonne, um deren Rotationsgeschwindigkeit durch die Vermessung von Sonnenflecken abzuschätzen.

- Smith et al. (2010) werteten die Bewegung eines stromdurchflossenen Leiters in einem Magnetfeld quantitativ aus und wiesen über die gemessenen Beschleunigungswerte des Leiters die Proportionalität zwischen Lorentzkraft und Stromstärke nach. Auch Molina-Bolívar und Abella-Palacios (2012) setzten die Videoanalyse in der *Elektrodynamik* ein: Die Autoren nutzten ein hohles, mit Magneten bestücktes Rohr, durch das sie nicht-magnetische Metallstücke fallen ließen (Wirbelstrombremse). Sie wiesen nach, dass die aufgrund der Lenzschen Regel erreichte stationäre Fallgeschwindigkeit der Metallstücke proportional zu der elektrischen Leitfähigkeit des Materials ist.
- Im Gegensatz zur Mechanik werden zur Untersuchung von Phänomenen in der *Optik* hauptsächlich zeitunabhängige Größen erfasst. Dabei wird auf die Möglichkeit zurückgegriffen, Winkel (z.B. zwischen zwei Strahlen) direkt im Video zu messen oder Konturen sowie Begrenzungslinien durch Positionsbestimmungen zu erfassen. Auf diese Weise könne Strahlengänge durch optische Geräte, Kaustiken, Schattenlinien, Beugungsmuster oder Emissionsspektren quantitativ ausgewertet werden (Brown & Cox, 2009; Gil et al., 2006; Rodrigues & Simeao-Carvalho, 2014).
- Darüber hinaus werden in manchen Publikationen einzelne, eher atypische Anwendungsbeispiele gegeben. So wertete Gallis (2010) eine Videosimulation zum Millikan-Experiment und Laws (1998) einen Cartoon-Kurzfilm aus. Wilhelm wies videoanalytisch nach, dass der Zerfall von Bierschaum durch eine Exponentialfunktion beschrieben werden kann und entwickelte ein Modellexperiment zur kinetischen Gastheorie, dass Holzkugeln in einer Rüttelkammer zum Nachweis der Maxwell-Boltzmann-Geschwindigkeitsverteilung nutzt (Geßner et al., 2009; Wilhelm, 2009).

Um mit dem Lernmedium einen Zugang zu weiteren physikalischen Größen zu erhalten, kann die Videoanalyse mit sensorbasierten Messsystemen kombiniert werden (Gröber et al., 2014). Über Sensoren gemessene Kräfte, Spannungen oder Ströme können auf Anzeigedisplays dargestellt und mit dem Experiment zusammen aufgenommen werden. Anwendungsbeispiele in der Experimentalphysik 1 sind die Analyse von Seil- und Lagerkräften, Zug- und Druckspannungen sowie die Messung von Temperaturen (Wärmelehre).

Wird der Versuch unternommen, die große Anzahl verfügbarer oder entwickelbarer Videoexperimente durch ein Kategoriensystem zu charakterisieren, dann stellt der Grad der Kontextualisierung neben der Komplexität der Experimente und der thematischen Verortung innerhalb der Physik ein wesentliches Unterscheidungsmerkmal dar. Dieser Grad reicht von modellhaften Laborversuchen (z.B. Wagen-Kugel-Experiment) bis zu authentischen Bewegungen aus Natur und Alltag (z.B. fahrende Autos, Sportler). Für Aufnahmen im Freien bietet sich die Verwendung mobiler Endgeräte wie Smartphones oder Tablet-Computer in besonderem Maße an: Durch die Verwendung spezieller Applikationen können die aufgenommenen Videos direkt am Endgerät analysiert werden, wie im nächsten Abschnitt dargestellt wird.

## 2.2 Videoanalyse mit mobilen Endgeräten

Mobile Endgeräte wie Smartphones oder Tablet-Computer eignen sich aufgrund der integrierten Digitalkamera zur Aufnahme von Bewegungen, die mit speziellen Applikationen direkt auf dem Gerät ausgewertet werden können. Dadurch werden Kompatibilitätsprobleme zwischen verschiedenen Hard- und Softwarekomponenten reduziert und es entsteht eine Zeitersparnis gegenüber der Videoanalyse mit dem Computer. Diesen Vorteilen stehen Nachteile wie eingeschränkte Funktionalitäten gegenüber. Aus didaktischer Perspektive bietet die mobile Videoanalyse damit einen einfachen Zugang zum selbstständigen Experimentieren und

erweitert die Auswertung von Videos um die Planung und Durchführung von Experimenten, die zwei wesentliche Komponenten des Experiments bilden (Klahr, 2000).

In diesem Abschnitt wird die Videoanalyse mit mobilen Endgeräten zunächst in den Kontext von Smartphone-Experimenten eingeordnet (Kap. 2.2.1) und auf das Angebot verfügbarer Applikationen zur mobilen Videoanalyse eingegangen (Kap. 2.2.2). Anschließend werden Möglichkeiten und Grenzen der mobilen Videoanalyse aufgezeigt (Kap. 2.2.3).

### 2.2.1 Smartphones und Tablet-Computer als Experimentiermedien

Im Allgemeinen bieten Smartphones und Tablet-Computer durch integrierte Sensoren diverse Möglichkeiten, physikalische Experimente durchzuführen, wodurch sie seit einigen Jahren in den Fokus der Physikdidaktik rücken (Kuhn, 2014). Durch geeignete Anwendungen („Apps“) können beispielsweise Beschleunigungs-, Lichtstärke- und Magnetfeldsensoren sowie das Mikrofon ausgelesen, die Daten gespeichert und dargestellt werden. Damit stellen die Geräte kleine mobile Messlabore dar, die bei den Lernenden weit verbreitet und in der heutigen Zeit allgegenwärtig sind. Zahlreiche Beiträge in Fachzeitschriften sowie regelmäßige Beiträge in spezialisierten Kolumnen beschreiben unterrichtspraktische Smartphone-Experimente für die Themenbereiche Mechanik, Optik, Akustik und viele mehr (Kuhn, 2014). Als wesentliche Vorteile dieses Mediums sind neben der hohen Mobilität die intuitive Handhabung, die aus messtechnischer Sicht einfache und schnelle Datenerhebung sowie die mit dem Medium verbundene Authentizität aufzuführen. Auf dem zuletzt genannten Aspekt fußt die Hypothese, dass bei den Lernenden die Motivation bei der Arbeit mit mobilen Endgeräten steigt. Diese Hypothese wurde bisher in den Themenbereichen Akustik (Sekundarstufe 1; Kuhn & Vogt, 2015) und Mechanik (Oszillationen; Sekundarstufe 2; Hochberg, 2016) untersucht.

Im Speziellen verfügen die meisten Smartphones und Tablet-Computer über eine Kamera zur Videoaufnahme, woraus sich die Möglichkeit zur *mobilen* Videoanalyse ergibt. Verglichen mit der Verwendung von Beschleunigungssensoren, die eine Alternative zur Analyse mechanischer Experimente darstellen, zeichnet sich die mobile Videoanalyse durch die hohe Anschaulichkeit des Messverfahrens und die berührungsfreie Messmöglichkeit aus. Für einen direkten Vergleich der mobilen Videoanalyse mit Beschleunigungssensoren ist zu berücksichtigen, dass die Lernenden mit den gemessenen Primärdaten vertrauter sind (Positionen an Stelle von Beschleunigungen; Beichner, 1994). Außerdem muss zur Messung der Beschleunigung mithilfe des integrierten Sensors das Gerät in den Versuchsaufbau integriert werden. Damit wird die Beschleunigung stets im mitbewegten System gemessen, was schon bei einfachen Experimenten bei den Lernenden zu Verwirrungen führen kann (und es besteht die Gefahr, das Messgerät zu beschädigen). Bei der Videoanalyse gestaltet sich allerdings die Erhebung von Messdaten zeitaufwendiger. Während Sensoren die Daten nach der Messung in einem verarbeitungsfähigen Format bereitstellen, muss nach der Durchführung und Aufnahme des Experiments erst eine Skalierung des Videos und eine Objektverfolgung durchgeführt werden. Dies geschieht unter Verwendung von Applikationen (Apps), die im nächsten Abschnitt beschrieben werden.

### 2.2.2 Applikationen zur mobilen Videoanalyse

Zur Auswertung aufgenommener oder abgespeicherter Experimentiervideos bedarf es einer App, mit der die Videos skaliert und die Bewegungsinformationen extrahiert und dargestellt werden können, vgl. Kap. 2.1. Im Sommer 2013 wurde eine Recherche über die verfügbaren Apps durchgeführt, um diese in Anlehnung an die gesammelten Erfahrungen computerbasierter Videoanalyse-Programme hinsichtlich der Funktionalitäten aus Tab. 1 zu testen (Klein et al., 2014b). Es wurden vier Applikationen gefunden und verglichen, die allesamt auf dem Apple-Betriebssystem (iOS) operieren — bis dato war keine App für den Android-Markt erhältlich. Zwischen den getesteten Apps ergaben sich vor allem hinsichtlich wichtiger Skalierungs- und Darstellungsfunktionen große Unterschiede, die insbesondere nicht an die Komplexität computerbasierter Programme heranreichen. Aus diesen Ergebnissen wurde



**Abb. 5:** Mobiles Analogie-Experiment zum Wagen-Kugel-Experiment aus Abb. 2: Ein Skateboardfahrer fährt mit konstanter Geschwindigkeit in  $x$ -Richtung und wirft einen Ball in  $y$ -Richtung (links bzw. mittig). Das Szenario wird mit einem Tablet-PC videografiert (rechts) und ausgewertet. Die Abbildung wurde aus Klein et al. (2013a) entnommen.

der Bedarf zur Entwicklung einer Android-basierten App abgeleitet, die die wesentlichen Funktionalitäten zur Bearbeitung videobasierter Aufgabenstellungen abbildet (vgl. Kap. 4). Die konzipierte App „*Video Analysis*“ ist seit Beginn 2015 erhältlich (Klein et al., 2015b) und wurde bereits an mehrfacher Stelle von Kollegen im Physikunterricht, in Lehrerfortbildungen (Tesch, 2015) und in der Lehre eingesetzt. In Klein et al. (2015b) werden die Funktionalitäten und einige Anwendungsmöglichkeiten der App skizziert. Mittlerweile sind auch kostenfreie Alternativen für Android-Geräte erhältlich, allerdings mit vergleichsweise weniger Optionen.<sup>2</sup>

### 2.2.3 Anwendungsbeispiele

Im Gegensatz zur klassischen Videoanalyse mit vorgegebenen Videos liegt der Fokus bei der mobilen Videoanalyse auf der Planung und Durchführung des Experiments. Durch die Möglichkeit, Messdaten ohne Videoexport unmittelbar auf dem mobilen Endgerät in kürzester Zeit auszuwerten, erhalten die Lernenden ein direktes Feedback über die Qualität ihrer Daten und können das Experiment ggf. wiederholen oder die Startbedingungen zur Aufnahme einer Messreihe verändern.

Aufgrund der Mobilität der Geräte bieten sich vor allem experimentelle Konstellationen an, die deutlich lebensnaher als Laborversuche sind. So kann das in Abb. 2 dargestellte Experiment beispielsweise dadurch realisiert werden, dass ein Skateboardfahrer während der Fahrt einen Ball senkrecht zur Bewegungsrichtung nach oben wirft (Klein et al., 2013a). Fährt die Person auf dem Skateboard mit konstanter Geschwindigkeit weiter, kann sie den Ball wieder auffangen (Abb. 5). Ein weiteres Experiment, das sich für den Einsatz in der Experimentalphysik 1 und (in geringer Komplexität) auch in der Schule eignet, ist das Fallen von Papier-Fallkegeln in Luft (Klein et al., 2014b). Durch die Luftreibung und die geringe Masse erreicht ein Papier-Fallkegel schon nach kurzer Fallstrecke ( $< 1$  m) eine konstante Endgeschwindigkeit, die videoexperimentell bestimmt werden kann. Werden mehrere flächengleiche Fallkegel übereinandergelegt, kann zudem die Abhängigkeit der Endgeschwindigkeit von der Masse des fallenden Körpers bestimmt und der Widerstandsbeiwert durch eine Ausgleichsrechnung ermittelt werden. Weitere Beispielexperimente werden in Kap. 4.2 genannt.

Technische Einschränkungen außer Acht lassend, können prinzipiell alle Experimente der „klassischen“ Videoanalyse (Videoaufnahme mit Digitalkamera, Auswertung am Computer) auch mit einem Smartphone / Tablet-Computer aufgenommen und ausgewertet werden. Die

<sup>2</sup>z.B. „VidAnalysis free“ (<https://play.google.com/store/apps/details?id=com.vidanalysis.free;03/2016>)

mobile Videoanalyse sollte sich aber durch die Möglichkeiten, die sich aus der Mobilität des Mediums ergeben, von der klassischen Videoanalyse abgrenzen; die Aufnahme von typischen Laborexperimenten würde den mobilen Charakter der Geräte konterkarieren. Außerdem sollte das Lernen mit der mobilen Videoanalyse nicht auf dem Nutzen elaborierter Auswertemethoden beruhen, die ohnehin besser mit dem Computer durchgeführt werden können. Aus technischer Sicht wird die Eignung von Experimenten durch die Hardware des Tablet-PCs eingeschränkt (z.B. Bildrate der Kamera, Rechenleistung des Tablet-PCs; vgl. Klein et al., 2014b). Es ist aber davon auszugehen, dass diese Einschränkungen mit voranschreitender Entwicklung an Bedeutung verlieren. Schon jetzt ermöglichen manche Smartphone-Kameras Hochgeschwindigkeitsaufnahmen mit bis zu 240 Bilder pro Sekunde in einer Full-HD Auflösung (1080p).



## Kapitel 3

# Theoretischer Hintergrund und empirischer Forschungsstand

Im letzten Kapitel wurde die Videoanalyse als physikalisches Mess- und Analysewerkzeug beschrieben und dessen Einsatzmöglichkeiten aufgezeigt. In diesem Kapitel wird aus fachdidaktischer und lernpsychologischer Perspektive begründet, weshalb sich die Videoanalyse für den Einsatz in der (universitären) Lehre eignet (Kap. 3.1). Die aufgeführten lernförderlichen Elemente sind zunächst nicht an ein Nutzungskonzept des Mediums (wie z.B. in Form von Aufgaben oder als Demonstrationsmedium) gebunden, werden aber an einigen Stellen in diesem Kapitel und systematisch in Kapitel 4 darauf bezogen.

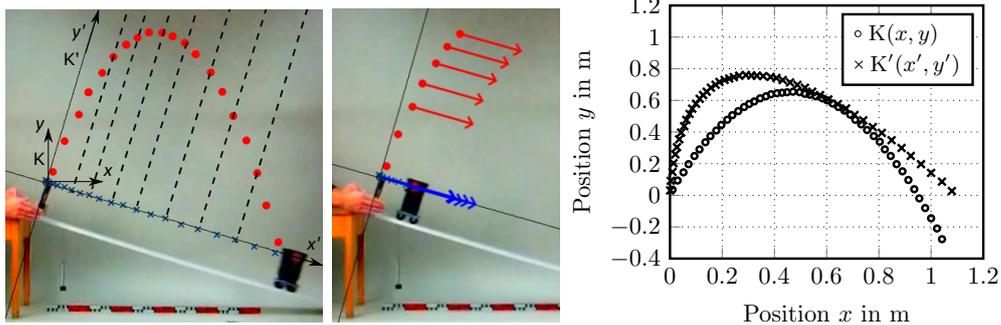
In Kapitel 3.2 werden die bisherigen empirischen Arbeiten zur Lernwirksamkeit der Videoanalyse zusammengefasst und Bezüge zu den dargestellten theoretischen Grundlagen aufgezeigt. Die Forschungslage wird abschließend bewertet, um Konsequenzen für die empirische Ausrichtung dieser Arbeit zu ziehen.

### 3.1 Physikdidaktischer und lernpsychologischer Hintergrund

#### 3.1.1 Visualisieren und qualitatives Begründen

Neben der hohen Anschaulichkeit des Messverfahrens liefert die visuelle Darstellung der Experimente einen direkten Zugang zu den abgebildeten Phänomenen, was besonders bei komplexeren Bewegungsabläufen mit mehreren Objekten vorteilhaft ist. Am Beispiel der beiden Wagen-Kugel-Experimente (Abb. 2 und 4) wird ersichtlich, dass eine einmalige Betrachtung der Experimente die Aufmerksamkeit der Beobachter überfordern würde. Durch die Möglichkeit, Videos zu jeder Zeit anzuhalten, schrittweise vor und zurück sowie verlangsam abzuspielen, kann eine intensive, *qualitative* Auseinandersetzung mit dem Experiment erfolgen, wie im Folgenden beschrieben wird.

Werden bei dem Wagen-Kugel-Experiment auf der schiefen Ebene die Positionen von Kugel und Wagen durch Anhalten des Videos zu jedem Zeitpunkt beobachtet und zueinander in Bezug gesetzt, liegt eine qualitative Begründung für den Ausgang des Experiments auf der Hand: Wird zu einem festen Zeitpunkt eine Gerade betrachtet, die senkrecht zur Fahrbahn verläuft und die die aktuelle Position des Wagens schneidet, so wird ersichtlich, dass diese Gerade auch die Position der Kugel schneidet. Mit anderen Worten befindet sich die Kugel zu jedem Zeitpunkt senkrecht über dem Wagen (senkrecht in Bezug auf die Unterlage, auf der sich der Wagen bewegt), siehe Abb. 6 links. Daraus kann geschlossen werden, dass in Richtung senkrecht zur gedachten Geraden die gleichen Kräfte auf Kugel und Wagen wirken, denn die beiden Körper erfahren die gleiche Beschleunigung. Zur Kontrolle können die Beschleunigungsvektoren ebenfalls visualisiert werden (Abb. 6 mittig). Auch wenn den



**Abb. 6:** Multiple Repräsentationsformen am Beispiel des Wagen-Kugel-Experiments auf der schiefen Ebene: Die Kugel befindet sich im Bezugssystem  $K'$  jederzeit senkrecht über dem Wagen (links), d.h. die  $x'$ -Komponenten der Beschleunigungsvektoren beider Objekte sind zu jeder Zeit längengleich (mittig). Das Diagramm (rechts) zeigt die Bahnkurve der Kugel in den relativ zueinander gedrehten Bezugssystemen  $K$  und  $K'$ .

Lernenden eine theoretische Beschreibung der abgebildeten Zusammenhänge zunächst schwer fallen mag, bietet die Möglichkeit der Beobachtung einen Einstieg in die Fragestellung und vermag Denkprozesse anzustoßen (Klein et al., 2015d). Die Nützlichkeit, beide Körper in einem rotierten Koordinatensystem (mit der  $x$ -Achse parallel zur schiefen Ebene) zu beschreiben, wird durch obige qualitative Betrachtungsweise erkennbar und ist ein wichtiger Schritt zur theoretischen Problemlösung.

Dieses Beispiel zeigt, dass die Möglichkeit zur qualitativ-inhaltlichen Auseinandersetzung mit realen Phänomenen durch die Videoanalyse prinzipiell gegeben ist. Unabhängig von quantitativen Auswertungsmöglichkeiten besitzt die Videoanalyse demnach das Potential zur Förderung des qualitativen Verständnisses (Seifert et al., 1997). Die Akzentuierung qualitativer Fragestellungen hängt von der Zielsetzung des Lehr-/Lernszenarios ab und ist besonders dann nützlich, wenn eine quantitative Behandlung von Phänomenen schwierig ist. Beispiele, bei denen vermehrt qualitative Betrachtungsweisen denn quantitative Analysen eine Rolle spielen, sind hochkomplexe physikalische Vorgänge, die mit Hochgeschwindigkeitskameras aufgenommen werden, z.B. das Platzen eines wassergefüllten Ballons. Den Lernenden liegt eine experimentelle Realität vor, die sie beobachten und nach qualitativ-inhaltlichen Gesichtspunkten analysieren können.

Neben der Möglichkeit qualitative Auseinandersetzungen mit dem Lerninhalt anzuregen, besitzt die Visualisierung von Phänomenen durch Videos auch Motivationspotential: Videos bewirken als realistische Abbilder eine starke Situierung und eine direkte Anbindung an einen konkreten Anwendungsbezug. Verglichen mit vereinfachten Darstellungen oder Schemazeichnungen bewirkt die Arbeit mit realistischen Abbildern (wie Videos) größere Lerneffekte, wie in einer Meta-Analyse nachgewiesen wurde (Höffler & Leutner, 2007). Der Realitätsbezug der Videos (und damit der Lernerfolg) ist vermutlich umso größer, je näher das abgebildete Phänomen an der Lebenswelt der Lernenden liegt. In der Physikdidaktik wird eine solche Situierung ausdrücklich gefordert, wie im nächsten Abschnitt dargestellt wird.

### 3.1.2 Kontextualisierung und situiertes Lernen

Ein ernüchterndes und gleichsam bedeutendes Ergebnis internationaler Leistungsvergleiche von Lernenden (z.B. PISA) waren Defizite bei der Anwendung naturwissenschaftlichen Wissens in neuen Kontexten (PISA, 2007). Der mangelnde Transfer erworbener Fähigkeiten auf konkrete Anwendungsprobleme wird auf zu stark kognitivistisch ausgerichtete Lehrmethoden zurückgeführt. Durch einen „Wissenstransport“ abstrahierter Inhalte, die zu stark an fachlichen Aspekten orientiert sind, wird unflexibles (sogenanntes „träges“) Wissen generiert, das nicht auf realitätsbezogene Aufgaben übertragen werden kann (Müller, 2006; Reinmann &

Mandl, 2006). Als Problemlösung wurden authentische Problemstellungen vorgeschlagen, die inhaltlich an die Lebenswelt der Lernenden angebunden sein sollen (*Kontextorientierung*). Vor diesem Hintergrund wurde der Begriff der geforderten „Authentizität“ vielfach diskutiert und verschieden interpretiert (Engeln, 2004; Radinsky et al., 2001). Neben dem Bezug des Lernmaterials zur realen Lebenswelt (Barab et al., 2000; Duit & Mikelskis-Seifert, 2007; Kuhn, 2010) wird darunter auch ein Bezug zu fachspezifischen Methoden verstanden (disziplinäre Authentizität; Enghag & Niedderer, 2007). Das Akzentuieren von Problemstellungen schlägt sich auch in der „neuen Aufgabenkultur der Physik“ nieder. Dabei sollen eben jene Aufgaben ins Zentrum des Lernens rücken, die Alltagserfahrungen thematisieren sowie den Umgang mit neuen Medien erfordern und sowohl selbstständige als auch kooperative Arbeitsweisen fördern (Häußler & Lind, 2000).

Hinter diesen kontextorientierten Lernansätzen steht die Rahmentheorie des *situierten Lernens* (Mandl & Kopp, 2005). Die Aneignung von Wissen findet demnach als individueller Prozess in bestimmten Handlungskontexten statt und wird in Anlehnung an konstruktivistische Positionen der pädagogischen Psychologie als aktiver Konstruktionsprozess angesehen. Die Selbsttätigkeit der Lernenden steht im Vordergrund. Durch eine Meta-Analyse konnte eine positive Wirkung kontextorientierter Unterrichtsansätze auf die Haltung von Schülerinnen und Schülern bzgl. des Fachs Physik nachgewiesen werden (Bennett et al., 2007). Konkrete instruktionsbasierte Beispiele, die sich in den Rahmen der Kontextorientierung einordnen lassen, sind Zeitungs- und Werbeaufgaben, deren nachhaltige Leistungs- und Motivationsförderung ebenfalls nachgewiesen wurde (Kuhn, 2010; Vogt, 2010).

### Bezug zur Videoanalyse

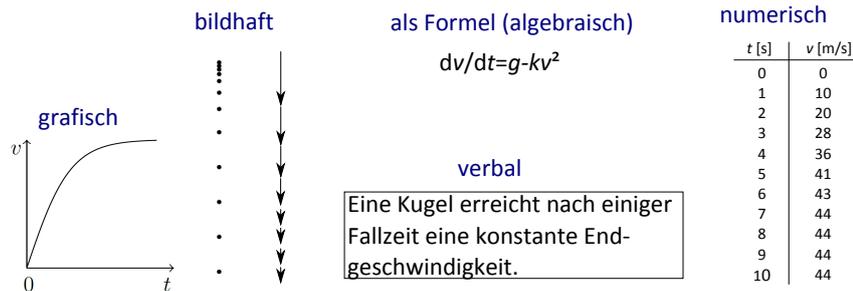
Durch die Videoanalyse kann eine Kontextorientierung auf mehreren Ebenen umgesetzt werden:

- *Authentische Phänomene (Realitätsbezug der Videos)*: Beispiele fallen vor allem in den Bereich „Physik und Sport“, Naturphänomene (z.B. Optik), Technik (z.B. Bremswege von Autos) und Biomechanik (z.B. Gehen, Laufen, Tanzen).<sup>1</sup> Den Beispielen ist gemeinsam, dass sie an Stelle von *Laborvideos* vorwiegend Aufnahmen in lebensweltlichen Szenarien verwenden. Paradebeispiele dafür sind die Arbeiten von Schüttler und Wilhelm (2011) sowie Palazzo und Schools (2008), die sich der „Physik im Freizeitpark“ widmen. Es wird in der fachdidaktischen Literatur an mehrfacher Stelle argumentiert, dass die Videoanalyse vor allem bei anwendungsbezogenen, authentischen Experimenten motivationsfördernd wirkt (Kuhn, 2010).
- *Authentizität des Mediums*: Durch die Verwendung eines Smartphones / Tablet-PCs als Analysewerkzeug entsteht eine weitere Anbindung an die Lebenswelt der Lernenden, denn sie sind mit dem Gerät vertraut (*materiale Situierung*, Kuhn & Vogt, 2015).
- *Authentizität der Analysemethode (disziplinäre Authentizität)*: Die fachspezifische Authentizität wird dadurch gefördert, dass die Videoanalyse Theoriebildung mit experimentellen Methoden verknüpft. Die Arbeitsweise der Physik zeichnet sich insbesondere durch das Zusammenspiel dieser beiden Komponenten aus (vgl. Kap. 3.1.4).

### 3.1.3 Lernen mit multiplen Repräsentationen

Der Begriff „Multiple Repräsentationen“ bezieht sich auf die unterschiedlichen (externen) Darstellungsarten (z.B. Sprache, Bilder, Diagramme, Formeln), mit denen ein physikalischer Prozess oder ein physikalischer Sachverhalt beschrieben werden kann. Sie sind Kommunikationsmittel, die eine wichtige Rolle beim Lernen physikalischer Konzepte spielen, indem sie einen Zugang zu Problemstellungen schaffen und den Wissenserwerb erleichtern können

<sup>1</sup>(Brown & Cox, 2009; Cluss et al., 2006; Gil et al., 2006; Goff & Carré, 2009; Gröber et al., 2014; Haugland, 2001; Hockicko et al., 2014; Rodrigues & Simeao-Carvalho, 2014; Wilhelm et al., 2003).



**Abb. 7:** Kohärente Repräsentationen eines physikalischen Bewegungsvorgangs: Geschwindigkeit-Zeit-Diagramm, Stroboskopbild (Punktspur) und vektorielle Darstellung der Beschleunigung, Differentialgleichung ( $g$  = Erdbeschleunigung,  $k$  = Konstante), verbale Beschreibung und Datentabelle. Repräsentationskompetenz wird als Fähigkeit verstanden, verschiedene Darstellungen einzeln zu interpretieren, sie miteinander zu verknüpfen, zwischen ihnen zu wechseln und sie zum Problemlösen zu nutzen.

(Rosengrant et al., 2006). Abbildung 7 verdeutlicht die vielfältigen Darstellungsmöglichkeiten an einem Beispiel aus der Kinematik, dem Fall eines Körpers unter dem Einfluss der (Newtonschen) Luftreibung.

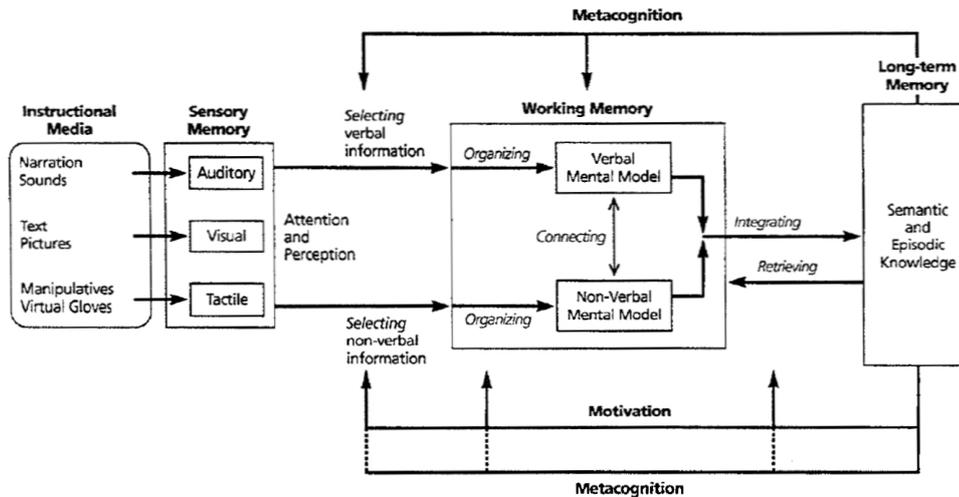
Physikdidaktische Forschungsergebnisse belegen, dass es Lernenden oftmals schwer fällt, eine Verbindung zwischen einem physikalischen Experiment (bzw. einer realen Abbildung davon) und einer abstrakteren Darstellungsform (z.B. Diagramme oder Formeln) herzustellen (Ainsworth, 2006; Van Heuvelen, 1991; Kohl & Finkelstein, 2005; Meltzer, 2005). Der kompetente Umgang mit Repräsentationsformen ist Teil fachlicher Expertise (*Repräsentationskompetenz*, siehe S. 44) und bedeutsam für das Verstehen von Physik. Der Integrationsprozess verschiedener externer Darstellungen in eine vernetzte mentale Struktur gilt dabei als notwendige Bedingung für den Erwerb von Repräsentationskompetenz. Ein lerntheoretischer Rahmen zur Erklärung dieses Prozesses bietet die Cognitive-Affective Theory of Learning with Media, deren Grundannahmen im folgenden Abschnitt erläutert werden. Anschließend wird die Bedeutung der Repräsentationskompetenz für das Lernen näher erläutert und ein Bezug zur Videoanalyse hergestellt.

### Cognitive-Affective Theory of Learning with Media (CATLM) und Cognitive Load Theory (CLT)

Erkenntnisse über die kognitiven Prozesse, die beim Arbeiten mit multiplen Repräsentationen ablaufen, liefert die Cognitive-Affective Theory of Learning with Media (CATLM; Moreno, 2005; Moreno & Mayer, 2007).<sup>2</sup>

Die erste zentrale Annahme der CATLM ist die Existenz zweier verschiedener Kanäle, innerhalb derer die Informationsverarbeitung stattfindet (visuell, auditiv). Die Modellstruktur der CATLM ist in Abb. 8 dargestellt. In verschiedenen Repräsentationsmodi/-codes (verbal, nonverbal) vorliegende Inhalte werden zunächst vom Lernenden über Sinneskanäle wahrgenommen. Verbale Lernmaterialien beziehen sich beispielsweise auf gesprochene oder gedruckte Sätze, nonverbale auf Animationen, Bilder, Videos oder Musik (vgl. duale Kodierungstheorie, Paivio, 1986). Nach der Wahrnehmung folgt die Differenzierung bezüglich relevanter Informationen (Selektion) und die Weiterleitung ins Arbeitsgedächtnis, wo die Informationen zu mentalen Modellen geformt (Organisation) werden. Im Arbeitsgedächtnis werden Verbindungen zwischen Repräsentationen hergestellt. Dabei unterscheiden Mayer und

<sup>2</sup>Die CATLM stellt eine Erweiterung der Cognitive Theory of Multimedia Learning dar (CTML; Mayer, 1999). Beide Theorien fußen auf der dualen Kodierungstheorie (Paivio, 1986), dem Modell des Arbeitsgedächtnisses von Baddeley und Hitch (1974), der Cognitive Load Theory (CLT; Sweller, 1988), und zeigen Berührungspunkte zu weiteren Theorien auf.



**Abb. 8:** Modell der Cognitive-Affective Theory of Learning with Media (CATLM). Abbildung entnommen aus Moreno (2005).

Anderson (1991) Verbindungen zwischen verschiedenen Repräsentationsformen *eines* Modus („*representational connections*“) von Verbindungen zwischen Repräsentationen *verschiedener* Modi („*referential connection*“). Representational connections, die für diese Arbeit eine bedeutende Rolle spielen (multiple Repräsentationsformen, siehe unten) meinen z.B. Verbindungen zwischen non-verbale Repräsentationsformen wie Diagrammen und schematischen Zeichnungen, während sich referential connections beispielsweise auf Verbindungen zwischen gesprochenem Text und Bild beziehen. Werden verschiedene Repräsentationsformen simultan dargestellt, wird verhindert, dass getrennte mentale Repräsentation für einen längeren Zeitraum im Arbeitsgedächtnis aufrechterhalten werden müssen, was zu einer Überlastung führen würde (siehe zweite Annahme des Modells, unten). Das Lernen wird demnach durch eine simultane statt einer sukzessiven Präsentation von Informationen unterstützt (Kontiguitätsprinzip, Mayer & Moreno, 2003). Im Modell der CATLM werden die Informationen aus dem Arbeitsgedächtnis schließlich in das Langzeitgedächtnis integriert (Integration). Das Langzeitgedächtnis ist in eine semantische und episodische Komponente unterteilt, die sich auf Faktenwissen bzw. persönlichen Erinnerungen beziehen.

Zudem berücksichtigt die Theorie motivationale und metakognitive Faktoren, die im Rahmen einer affektiven Mediation auf den Lernprozess einwirken bzw. das Lernen regulieren. Lernende, die über ihr eigenes Wissen und ihre eigenen Fähigkeiten reflektieren können, lernen zielgerichteter. Außerdem wird der sensorische Kanal erweitert: Neben auditiven und visuellen Informationen werden taktile Reize aufgeführt. Diese spielen im Rahmen der Videoanalyse zwar eine untergeordnete Rolle, können aber positive Lernwirkungen anderer Arbeiten in diesem Gebiet retrospektiv erklären (vgl. Kap. 3.2). Die Möglichkeit zur Reflexion des eigenen Lernens wird durch die Videoanalyse explizit aufgegriffen, wie weiter unten ausgeführt wird.

Eine zweite zentrale Annahme der CATLM ist die begrenzte Kapazität des Arbeitsgedächtnisses zur Informationsverarbeitung, was in Einklang mit Annahmen der *Cognitive Load Theory* steht. Der *Cognitive Load (CL)*<sup>3</sup> beschreibt die Belastung des Arbeitsgedächtnisses während des Lernens und setzt sich additiv aus drei Komponenten zusammen. Der *intrinsic CL* bezieht sich auf die Wesensart des Lernmaterials, wird durch das Vorwissen des Lernenden und die Komplexität des Lerninhalts bestimmt und ist unvermeidbar. Der *extraneous CL* bezieht sich auf die Präsentation der Lerninhalte, wird also durch die Gestaltung des Lernmaterials beeinflusst, und sollte so gering wie möglich sein. Der *germane CL* bezieht sich auf die notwendige Anstrengung zur Konstruktion und Automatisierung von Lern-

<sup>3</sup>Da es sich um einen Eigenbegriff aus der Kognitionspsychologie handelt, wird von einer Übersetzung abgesehen.

Schemata, ist lernrelevant und sollte möglichst hoch sein, um bessere Verständnisleistungen zu ermöglichen.

Schließlich betrifft die letzte Annahme die *aktive Rolle* bei der Informationsverarbeitung, die auch in den Theorien anderer Kognitionspsychologen Niederschlag findet (Ausubel, 1968; Bruner, 1961). Demnach bedarf es einer aktiven Beschäftigung mit dem Lerngegenstand, um kohärente mentale Repräsentationen des Lerngegenstandes bilden zu können und Neues mit Bekanntem zu vernetzen. Diese Annahme wird im Modell der CATLM (Abb. 8) an zwei Stellen abgebildet: Erstens moderiert die kognitive Aktivierung die Verknüpfung von verbalen und non-verbalen mentalen Modellen des Lerngegenstands, d.h. es werden verschiedene repräsentationale Abbilder in Bezug zueinander gesetzt. Zweitens setzt die Verbindung zwischen Arbeits- und Langzeitgedächtnis aktive Integrations- und Abrufprozesse voraus, die nur durch eine aktive Auseinandersetzung mit dem Lerngegenstand erfolgen. Sie ermöglichen Anknüpfungspunkte zu bereits vorhandenem Wissen und das Anwenden eingeübter Prozeduren.

Bevor ein konkreter Bezug zwischen CATLM und der Videoanalyse als Lernmedium hergestellt wird, wird die Rolle von multiplen Repräsentationen für das Lernen aus physikdidaktischer Position dargestellt.

### Repräsentationskompetenz als Komponente fachlicher Expertise

Unter Repräsentationskompetenz werden die Fähigkeiten verstanden, externe Darstellungsformen zu interpretieren, selbstständig zu erzeugen und zwischen verschiedenen Darstellungen zu wechseln (De Cock, 2012). In diesem Zusammenhang werden unter dem englischen Begriff „*representational fluency*“ (Hill et al., 2015) die Fähigkeiten zusammengefasst (1) mit einzelnen Repräsentationsform umzugehen und zwischen ihnen zu wechseln, (2) Repräsentationen zielgerecht zu nutzen und (3) bisher unbekannte Repräsentationsformen schnell zu erlernen (vgl. auch Bieda & Nathan, 2009; diSessa, 2004; Kohl & Finkelstein, 2005). Das flexible Nutzen unterschiedlicher Repräsentationen ermöglicht die Bildung eines tieferen Verständnisses. Dies führt dazu, dass auf unterschiedliche repräsentationale Zugänge zur Lösung eines physikalischen Problems zurückgegriffen werden kann („*representational flexibility*“; De Cock, 2012). Ist ein Lernender im Stande, mehrere Repräsentationsformen beim Lösen physikalischer Probleme zu nutzen, ist er flexibler im Aneignen neuer Inhalte und im Problemlösen. Folglich ist Repräsentationskompetenz eine Komponente fachlicher Expertise und steht in Verbindung zum konzeptionellen Verständnis. Folgerichtig führen Etkina et al. (2006b) die Fähigkeit, physikalische Vorgänge in verschiedenen Darstellungsformen zu beschreiben, als erste von sieben fachdisziplinären Fähigkeiten auf, die in der Hochschule ausgebildet werden sollten.

Empirische Studien belegen, dass sich Novizen und Experten darin unterscheiden, wie sie Repräsentationen nutzen und interpretieren (Hsu et al., 2004; Kohl & Finkelstein, 2008). Während Experten Repräsentationen systematisch und zielgerecht einsetzen, tendieren Novizen dazu, zwischen verschiedenen Darstellungsformen unmotiviert und unsystematisch zu wechseln. Fehlende Routinen erschweren Integrationsprozesse von Informationen des Langzeitgedächtnisses ins Arbeitsgedächtnis und führen gemäß der CATLM zu einem erhöhten Cognitive Load, da einzelne Repräsentationen länger im Arbeitsgedächtnis verweilen. Entsprechend benötigen Experten weniger Zeit, um relevante Informationen auf ein Problem zu beziehen, was empirisch nachgewiesen wurde (Etkina et al., 2006b; Kohl & Finkelstein, 2008). In Kapitel 5.1 wird ein Überblick über den Stand der Forschung zur Repräsentationskompetenz, bezogen auf das Inhaltsgebiet Kinematik, gegeben. Es sei vorweggenommen, dass die repräsentationalen Fähigkeiten von Studierenden zu Beginn des Physik-Studiums unzureichend ausgeprägt sind (McDermott et al., 1987), weshalb die explizite Berücksichtigung von Repräsentationskompetenz in Einführungsveranstaltungen zunehmend gefordert wird (Nieminen et al., 2012).

### Bezug zur Videoanalyse

Im Folgenden wird ein expliziter Bezug zwischen den Unterstützungsmaßnahmen und Möglichkeiten der Videoanalyse und den Prinzipien der CATLM hergestellt. Die englischen Schlagwörter neben dem Text entstammen aus der Originalliteratur von Mayer und dort zitierter (älterer) Literatur, weshalb von einer Übersetzung abgesehen wird, und beziehen sich auf die Darstellung der CATLM in den vorangegangenen Abschnitten (Mayer & Anderson, 1991; Mayer, 1999, 2009).

- 1) Textbasierte Instruktionen (Aufgabenstellungen) und zugehörige Experimentalvideos werden dem Lernenden gemeinsam präsentiert und stehen während dem Lernprozess simultan zur Verfügung. Die Problemstellung wird visualisiert. *dual coding, referential connection*
- 2) Die Interaktivität des Mediums ermöglicht eine intensive Auseinandersetzung mit dem Lerngegenstand. Beispielsweise können in dem Wagen-Kugel-Experiment Position und Orientierung des Bezugssystems verändert werden, um die Auswirkung auf die Bewegungsgraphen unmittelbar zu studieren (Abb. 6 rechts). Die aktive Rolle der Studierenden wird einerseits durch die videoexperimentellen Tätigkeiten, andererseits durch das selbstständige Experimentieren im Rahmen von mobilen Videoanalysen gefördert. *active learning\**
- 3) Durch die Videoanalyse stehen dem Lernenden simultan mehrere Repräsentationsformen zur Verfügung, um den Bewegungsvorgang in verschiedenen Abstraktionsstufen zu beschreiben. Videoanalyse-Programme ermöglichen es beispielsweise, Diagramme zeitsynchron zum Ablauf der Bewegung aufzubauen oder vektorielle Größen über das Videobild einzublenden (vgl. Abb. 2 rechts bzw. Abb. 6 mittig). Durch diese Multikodierung wird eine kognitive Verbindung zwischen verschiedenen Repräsentationsformen gefördert. *representational connection, contiguity*
- 4) Einzelne visuelle Informationen müssen nicht im Arbeitsgedächtnis aufrechterhalten werden, denn durch Videoanalyse ist es möglich, mehrere Repräsentationsformen nebeneinander zu positionieren und zeitsynchron wiederzugeben sowie Prozesse dynamisch vor- bzw. zurückspulen und mehrfach zu betrachten. *cogn. load, synchronicity*
- 5) Die optionale Bereitstellung von Unterstützungsmaßnahmen auf verschiedenem Abstraktionsniveau reduziert die kognitive Belastung. Liegt der instruktionale Fokus beispielsweise auf der Interpretation von Messdaten, stellt die händische Erzeugung von Diagrammen einen extraneous Cognitive Load dar, der durch die automatische Diagrammerzeugung vermieden wird. Die Sinnhaftigkeit einer automatischen (statt einer händischen) Diagrammdarstellung hängt von der Expertise der Lernenden und der Etappe im Lernprozess ab. *cogn. load, scaffolding\*\**
- 6) Durch die Darstellung mehrerer Messreihen in einem Diagramm können die Bewegungen verschiedener Objekte analysiert und miteinander verglichen werden. Im Gegensatz zur sequentiellen Analyse der Objekte wird dadurch der Cognitive Load reduziert. *cogn. load, contiguity, synchronicity*

\*vgl. Ausubel (1968); Bruner (1961); \*\*vgl. Stone (1998).

Die genannten Aspekte können sich gemäß der CATLM positiv auf den Lernerfolg auswirken, den Umgang mit Repräsentationen fördern und dienen als Grundlage zur Formulierung der Hypothesen in dieser Arbeit (siehe Kapitel 9).

### 3.1.4 Theorie-Experiment-Wechselwirkung

Da es sich bei den Experimentiervideos im Regelfall um reale Aufnahmen eines Experiments handelt, kann ein Bezug der Messdaten zu einer theoretischen Beschreibung des Experiments hergestellt und diskutiert werden. Die Zielsetzung in den meisten Arbeiten zur Videoanalyse ist der Nachweis physikalischer Gesetzmäßigkeiten (z.B. Wurfparabeln), die Überprüfung von Modellgrenzen (z.B. Gültigkeit der Energieerhaltung) oder die Bestimmung von Parametern durch die Anpassung analytischer Funktionen an die Messdaten (z.B. Widerstandsbeiwerte oder Reibungskoeffizienten). Die Versuchsanordnungen werden unter Berücksichtigung des theoretischen Hintergrunds entsprechend präpariert.

Dieses *Wechselspiel zwischen Theorie und Experiment* hat ein hohes physikdidaktisches Potential zur Steigerung des konzeptionellen Verständnisses der Lernenden (vgl. z.B. Trumper, 2003) und spiegelt einen wichtigen Aspekt fachdisziplinärer Arbeitsweisen wider (Etkina et al., 2006a). Die inhaltliche Ausarbeitung oder die methodische Manipulation dieses Aspekts ist nicht Inhalt dieser Arbeit, wird aber in anderen Forschungsarbeiten verfolgt (Gröber et al., 2014). Vielmehr wird das Wechselspiel zwischen Theorie und Experiment als Prinzip eingesetzt, Videoanalyse-Aufgaben zu konstruieren und ein Motivationspotential zu generieren: Wenn es Lernenden gelingt, ihre theoretischen Überlegungen oder Berechnungen mit den experimentellen Daten in Einklang zu bringen, erhalten sie ein positives Feedback, was Thornton (1987) auch im Zusammenhang mit anderen Medien (MBLs<sup>4</sup>) als pädagogischen Mehrwert nennt. Dies kann sich positiv auf die Motivation — insbesondere auf das akademische Selbstkonzept (vgl. Kap. 5.4) — auswirken. Durch das Medium erhalten die Lernenden eine Kontrollmöglichkeit ihrer theoretischen Berechnungen und können ihr qualitatives Verständnis prüfen. Die Möglichkeit zur Reflexion der Problemlösung zeigt einen Berührungspunkt zur Theorie der Metakognition auf.

## 3.2 Studien zur Lernwirksamkeit der Videoanalyse

Im Rahmen der Arbeit wurden existierende Studien zur Wirksamkeit der Videoanalyse systematisch recherchiert. Auswahlkriterien der Publikationen sind:

- Der Artikel wurde im Zeitraum von 1985 – 2015 publiziert.
- Die Ergebnisse der Studie wurden englischsprachig oder deutschsprachig kommuniziert. Eine Liste durchsuchter Quellen (insgesamt 34 ausgewählte Fachzeitschriften / Tagungsbände) befindet sich im Anhang (S. 201).
- Die Videoanalyse von Bewegungen war wesentlicher Bestandteil des Lehr-/Lernszenarios.
- Im Artikel werden empirische Daten zum Lernerfolg (Leistung, Motivation o.ä.) berichtet.

Die Auswahlkriterien wurden von 23 Artikeln erfüllt. Beiträge, die ausschließlich anekdotische Erfahrungen mit dem Einsatz der Videoanalyse (ohne empirischen Beleg) berichten, blieben unberücksichtigt. Tab. 3 stellt die Ergebnisse der Literaturrecherche dar und führt neben der inhaltlichen Verortung auch Angaben zu den Materialien und Methoden der Studien auf. Methodische Kriterien, an denen die Qualität der Studien beurteilt werden kann, betreffen das Studiendesign, die Güte der verwendeten Instrumente sowie die Materialgleichheit zwischen Vergleichsgruppen. Zur Einschätzung der praktischen Relevanz und Generalisierbarkeit der Ergebnisse werden weiterhin Stichprobengröße, Interventionsdauer und die Verwendung inferenzstatistischer Analysemethoden aufgeführt. Im Folgenden werden die Arbeiten entsprechend den untersuchten Fragestellungen gegliedert, zusammengefasst und die Untersuchungslage abschließend bewertet.

<sup>4</sup>Microcomputer-Based Laboratories (MBL) bezeichnet computergestütztes Messen, Auslesen und Darstellen physikalischer Größen mithilfe von Sensoren für Lernzwecke (typischerweise in Laborpraktika).

**Tab. 3:** Empirische Studien zur Lernwirksamkeit der Videoanalyse

Referenz	Untersuchungsinhalt	Interventionsdauer	Stichprobe ( $N$ )	Design <sup>a</sup>	TG <sup>b</sup>	MG <sup>c</sup>	Hyp <sup>d</sup>
<b>Untersuchung kognitiver Variablen</b>							
<i>Lernwirksamkeit von Echtzeit-Diagrammdarstellung</i>							
Beichner (1990)	Kinematik: Echtzeit-Diagramme	2 Stunden	SuS (237)	4 × 2	●	●	- (!)
Brungardt und Zollman (1995)	Kinematik: Echtzeit-Diagramme	2 Wochen	SuS (30)	2 × 1	●	●	- (!)
Struck und Yerrick (2010)	Graphenverständnis	1 Woche	SuS (39)	2 × 3	○	●	+ (!)
<i>Lernwirkung bzgl. Repräsentationen</i>							
Beichner (1996)	Kinematik: Graphenverständnis	verschieden	SuS (368)	7 × 1	●	○	+ (!)
Boyd und Rubin (1996)	Stroboskopbilder, Diagramm	1 Aufgabe	SuS (8)	(Q)	-	-	+
Pappas et al. (2002)	Tabelle, Graphen, Bezugssysteme	2 Lerneinheiten	LA-Stud (5)	1 × 2	○	-	+
Koleza und Pappas (2008)	Graphen, Bezugssysteme (POE)	1 Woche	LA-Stud (48)	(Q)	-	-	+
Kanim und Subero (2010)	Stroboskopbilder, Vektoren	k.A.	Stud (223)	1 × 1	●	-	-
<i>Lernwirkung bzgl. des physikalischen Konzeptverständnisses</i>							
Wilhelm et al. (2007)	Mechanik, projektorientiert	1 Exp.	NF-Stud (38)	1 × 2	●	-	+ (!!)
Hockicko (2012)	Rollende Kugeln	k.A.	NF-Stud (162)	2 × 2	○	○	+ (!)
Trudel und Metiouri (2012)	Bewegungen mit $v = const.$	5×5 Stunden	SuS (32)	1 × 6	○	-	+ (!)
Zajkov und Mitrevski (2012)	Freier Fall, senkrechter Wurf	1 Aufgabe	LA-Stud (10)	1 × 3	○	-	+
Trudel und Metiouri (2013)	Wurfbewegungen	4 × 75 min	SuS (45)	1 × 6	○	-	+ (!)
Hockicko et al. (2014)	Bremswege	1 Aufgabe	NF-Stud (515)	2 × 2	○	○	+ (!)
Malgieri et al. (2014)	Huygens-Pendel	30 – 120 min	LA-Stud (29)	1 × 1	○	-	+
De Ambrosis et al. (2015)	Rotation, Translation	3×2 Stunden	LA-Stud (20)	1 × 2	○	-	+
Wee et al. (2015)	Freier Fall, senkrechter Wurf	2×70 min	SuS (123)	1 × 2	○	-	+ (!!)
<b>Untersuchung affektiver Variablen und Sonstiges</b>							
Escalada et al. (1996)	Motivation / Neugierde	k.A.	SuS (k.A.)	(Q)	-	-	+
Escalada und Zollman (1997)	Einflussfaktor PC-Vorkenntnisse	3–5 Aufgaben	LA-Stud (100)	2 × 2	●	●	+ (!!)
Rodrigues et al. (2001)	Neugierde	2 Aufgaben	Stud (207)	(Q)	-	-	+
Palazzo und Schools (2008)	Motivation (Kontexte)	k.A.	SuS (268)	1 × 1	○	-	+
Wee (2012)	Einstellung ggü. Medium	140 min	SuS (64)	(Q)	-	-	+
Wee et al. (2012)	Akzeptanz und Wahrnehmung	k.A.	SuS (k.A.)	(Q)	-	-	+

*Bemerkungen:* SuS = Schülerinnen und Schüler, LA = Lehramt, NF = Nebenfach, Stud = Studierende, VA = Videoanalyse, MBL = Microcomputer-Based Laboratories (in diesem Fall Bewegungssensoren, vgl. Fußnote 4 auf S. 46), POE = siehe Fußnote 5 auf S. 49.

<sup>a</sup>Gruppen × Testzeitpunkte, (Q) Qualitative Studie. <sup>b</sup>Informationen über Testgüte (= TG) vorhanden, z.B. Bericht von Kennwerten oder Einsatz standardisierter Tests (nur bei nicht-qualitativen Studien). <sup>c</sup>Materialgleichheit (= MG), nur bei Vergleichsstudien. <sup>d</sup>Hypothese (= Hyp) zur positiven Lernwirkung der Videoanalyse wurde nachgewiesen (+). Verwendung von Signifikanztests zur statistischen Absicherung der Ergebnisse (!) bzw. Bericht von Effektstärken (!!).

### 3.2.1 Wirksamkeit bezüglich kognitiven Variablen

Während in den 1990er Jahren überwiegend das Verständnis von Repräsentationen und die Wirksamkeit von Echtzeit-Diagrammdarstellungen untersucht wurden, rückte in diesem Jahrzehnt die Untersuchung des physikalischen Konzeptverständnisses anhand einzelner Beispiele oder Themen in den Vordergrund. In Anlehnung an diese Chronologie der Forschungstradition werden die Arbeiten wie folgt unterteilt: (1) Lernwirksamkeit von Echtzeit-Diagrammdarstellungen, (2) Lernwirkung bzgl. Repräsentationen und (3) Lernwirkung bzgl. des physikalischen Konzeptverständnisses.

#### (1) Lernwirksamkeit von Echtzeit-Diagrammdarstellung (EZDD)

Beichner (1990) untersuchte, wie sich die Echtzeit-Darstellung von Diagrammen (EZDD) im Rahmen von Videoanalysen auf das Lernen auswirkt. Er knüpfte damit an eine Studie von Brassel (1987) an, in der das Lernen mit MBLs (siehe Fußnote 4, S. 46) untersucht wurde. Brassel nutzte MBLs, um kinematische Daten händisch erzeugter Bewegungen in Echtzeit in einem Diagramm darzustellen. Dieses experimentelle Szenario spricht gemäß der CATLM den taktilen und den visuellen Kanal an (vgl. Abb. 8), wodurch es zur Bewegungsempfindung kommt (sog. „kinästhetisches Feedback“). Lernende, die mit kinästhetischem Feedback und EZDD arbeiteten, zeigten einen wesentlich größeren Lernerfolg als solche, bei denen die händisch erzeugte Bewegung erst 20–30 Sekunden später im Diagramm angezeigt wurde. Beichner verzichtete auf das kinästhetische Feedback, um allein die Auswirkung von EZDD auf den Lernerfolg zu messen. In einem Design mit zwei Testzeitpunkten wurden Gruppen, die eine Wurfbewegung videoanalytisch untersuchten (EZDD) mit anderen Gruppen verglichen, die Diagramme zur Wurfbewegung selbst zeichneten und zeitversetzt zur realen Bewegung betrachteten. Es konnte kein Einfluss von EZDD auf den Lernerfolg (die Fähigkeit, Graphen zu interpretieren) festgestellt werden.

Brungardt und Zollman (1995) untersuchten die Auswirkung von EZDD auf das Graphenverständnis über einen Instruktionszeitraum mehrerer Wochen. Die beiden Untersuchungsgruppen arbeiteten mit zeitsynchroner bzw. zeitverzögerter Diagrammdarstellung einfacher eindimensionaler Bewegungsvorgänge (z.B. freier Fall). Auch in dieser Studie wurde kein signifikanter Gruppenunterschied festgestellt, aber über eine Reihe positiver Nebeneffekte durch das Arbeiten mit EZDD berichtet (z.B. führte EZDD zu weniger Verwechslungen zwischen den Größen Geschwindigkeit und Beschleunigung). Das Ausbleiben von Effekten wird vor allem auf lernhinderliche Defizite verfügbarer Videoanalyse-Programme zurückgeführt, die ökonomisches Arbeiten mit EZDD nur unzureichend ermöglichten.

Fünfzehn Jahre später schloss eine weitere Arbeit an die Wirksamkeitsfrage von EZDD an. Struck und Yerrick (2010) verglichen die Wirkung des Arbeitens mit MBLs und mit Videoanalyse direkt miteinander. Eine Untersuchungsgruppe erforschte eindimensionale Bewegungsvorgänge zunächst mit digitaler Videoanalyse (keine EZDD) und anschließend mit MBLs (kinästhetisches Feedback und EZDD). Eine zweite Gruppe nutzte die beiden Methoden in umgekehrter Reihenfolge. Zu Beginn, in der Mitte (beim Wechsel der Methode) und zum Ende der Intervention, die ungefähr eine Woche dauerte, lösten die Lernenden stets die gleichen Aufgaben, die darin bestanden, Kinematikgraphen zu verbal beschriebenen Bewegungsvorgängen selbstständig zu zeichnen. Die Ergebnisse zeigen, dass EZDD und kinästhetisches Feedback (durch MBLs) zu einer vergleichbaren Steigerung des Graphenverständnisses führte wie die zeitverzögerte Darstellung ohne kinästhetisches Feedback (mithilfe der Videoanalyse). Im Gegensatz zu den Studien von Beichner (1990) und Brungardt und Zollman (1995) wurden die Experimentiervideos selbst erstellt und es lag insgesamt eine höhere Lerner-Aktivität durch eigenständiges Experimentieren vor. Bis auf wenige Einzelfälle führten beide Methoden ungeachtet der eingesetzten Reihenfolge zu einer vergleichbaren Lernwirkung.

## (2) Lernwirkung bzgl. Repräsentationen

Eine der ersten Vergleichsstudien über die Wirkung der Videoanalyse auf das Graphenverständnis stammt ebenfalls von Beichner. Er setzte einen standardisierten Test zur Erfassung des Graphenverständnisses in sieben Gruppen ein, nachdem sie über Mechanik unterrichtet wurden (Beichner, 1996). Der Unterricht unterschied sich bzgl. der Nutzungsweise und des zeitlichen Umfangs der Videoanalyse. Die Ergebnisse zeigen keine Unterschiede zwischen Lernenden, die nie mit Videoanalyse arbeiteten und Lernenden, in denen die Videoanalyse nur in Form von Demonstrationen genutzt wurden. Das Graphenverständnis ist umso größer, je höher der Anteil der Videoanalyse an dem Curriculum ist und je aktiver sich die Rolle der Lernenden gestaltet. Die höchste Leistung erreichten die Lernenden, die mehrere Monate selbstständig Videos aufnahmen und auswerteten. Da Lernende verschiedener Klassen verglichen wurden und keine Vortestergebnisse vorliegen, sind die Ergebnisse vorsichtig zu interpretieren. Nichtsdestotrotz ist eine hohe Leistung bei hoher Lerner-Aktivität theoriekonform und steht in Einklang mit den Ergebnissen von Struck und Yerrick (2010).

Kanim und Subero (2010) setzten die Videoanalyse im Experimentalphysik-Praktikum ein, um Stroboskopaufnahmen zweidimensionaler Bewegungen (z.B. Pendel) zu erzeugen und Vektoren im Videobild darzustellen. Sie untersuchten den Einfluss dieser Intervention auf das Verständnis des Richtungscharakters physikalischer Größen und des Zusammenhangs zwischen Kraft und Beschleunigung. Die Variablen wurden vor der Intervention mit offenen Fragestellungen und nach der Intervention mit standardisierten Items erhoben. Zwar wurde eine Leistungssteigerung nachgewiesen, der erzielte Erfolg blieb aber hinter den Erwartungen zurück. Im deskriptiven Vergleich mit den Ergebnissen von Studentengruppen anderer Universitäten nach verschiedenen Interventionen (Vorlesungen, Spezialtutorien) schnitt die Untersuchungsgruppe schlechter ab. Die Autoren führten den Misserfolg auf die mangelnde methodische Anbindung des Praktikumversuchs an andere Lehrveranstaltungen zurück.

In zwei Untersuchungen wurde das Verständnis über den Zusammenhang zwischen Bewegungsdiagrammen und Bezugssystemen sowie das Verständnis von Tabellen untersucht (Koleza & Pappas, 2008; Pappas et al., 2002). In einer Vorstudie erhielten Pappas et al. (2002) zunächst Anhaltspunkte dafür, dass die Videoanalyse zu einer besseren Interpretationsfähigkeit von Tabellen und Diagrammen führt. Die Ergebnisse stützen sich auf einen Prä-Post-Vergleich von Studierenden, die drei Lerneinheiten zur Videoanalyse von Standardbewegungen (schiefer Wurf, eindimensionale Bewegungen) absolvierten.

Für eine qualitative Folgestudie konstruierten Koleza und Pappas (2008) acht Lerneinheiten zu verschiedenen eindimensionalen Bewegungen aus dem Alltag. Den Lernenden wurden sogenannte POE-tasks (*Predict-Observe-Explain*-Aufgaben)<sup>5</sup> gestellt, die sie mithilfe der Videoanalyse in Gruppenarbeit beantworteten. Im Anschluss an die Intervention wurden einige Lernende interviewt. Es zeigte sich eine Verbesserung des konzeptionellen Verständnisses der Geschwindigkeit und der Rolle des Bezugssystems bei der Beschreibung von Bewegungen. Die Autoren stellen insbesondere heraus, dass die interaktiven und dynamischen Möglichkeiten des Mediums erst durch die Integration in Lernszenarien — wie in diesem Fall POE-tasks — wirksam seien, was in Einklang mit der CATLM steht (*active learning*).

## (3) Lernwirkung bzgl. des physikalischen Konzeptverständnisses

Malgieri et al. (2014) konstruierten das Huygens-Pendel (vgl. Abb. 4, rechts) und untersuchten das Verständnis der Studierenden bzgl. physikalischer Erhaltungssätze und der Schwerpunktbewegung. Nach einer theoretischen Auseinandersetzung mit dem Experiment wurde die Videoanalyse genutzt, um die Bewegung des Massenmittelpunkts und den zeitlichen Verlauf der Gesamtenergie des Systems zu untersuchen. Mit Multiple-Choice-Aufgaben und offenen Fragen wurde das Verständnis zu mehreren Messzeitpunkten während der Arbeitsphasen erhoben. Die Ergebnisse zeigen, dass die Lernenden im Voranschreiten des Unterrichts ein

<sup>5</sup> In POE-tasks äußern die Lernenden zunächst eine begründete Vermutung über den Ausgang eines Experiments oder eine spezifische Fragestellung (Vorhersage), beobachten dann das Experiment und erklären es schließlich.

besseres Verständnis des Huygens-Pendels entwickelten, was auf die Darstellung kinematischer Größen und Energien sowie die experimentelle Prüfung der theoretischen Überlegungen zurückgeführt werden konnte. Positive Nebeneffekte betreffen die Förderung kooperativer Arbeitsweisen und die Freude an den analytischen und experimentellen Betrachtungen des Systems.

Die meisten anderen Arbeiten zum physikalischen Konzeptverständnis stellen ebenfalls einen Lernfortschritt über die Zeit fest (vgl. Tab. 3). Aus methodischer Sicht positiv hervorzuheben sind die beiden Arbeiten von Hockicko, in denen der Lernerfolg durch Videoanalyse jeweils mit dem einer Kontrollgruppe verglichen wurde. Die Ergebnisse zeigen, dass die Arbeit mit der Videoanalyse zu einem besseren Verständnis der Physik rollender Kugeln unterschiedlicher Trägheitsmomente auf einer schiefen Ebene (Hockicko, 2012) und zur besseren Einschätzung der Länge von Bremswegen führt (Hockicko et al., 2014).

Im Gegensatz zu anderen Arbeiten konzipierten Wilhelm et al. (2007) eine komplette Lehrveranstaltung (Experimentalphysik für Nebenfach-Studierende) nach konstruktivistischen Prinzipien um: Anstelle einer klassischen Vorlesung arbeiteten Studierende in Kleingruppen projektorientiert an vernetzten Themenbereichen zur Mechanik. Der Kurs schloss mit einem Videoanalyseprojekt ab, bei dem verschiedene Bewegungsvorgänge (z.B. Schlag eines Golfballs, Person auf Kinderrutsche, Newtonsche Wiege) untersucht wurden. Die Ergebnisse des Prä- und Post-Tests (FCI) zeigen einen großen Lernzuwachs dieses Ansatzes insgesamt.

### 3.2.2 Wirkung bzgl. affektiver Variablen, Sonstiges

Neben der kognitiven Wirkung der Videoanalyse wurden Motivation und Neugierde bei der Arbeit mit der Videoanalyse sowie die Einstellung gegenüber dem Lernmedium untersucht. In diesem Zusammenhang wurde auch der Einfluss von Vorkenntnissen und Vorerfahrungen mit dem Computer auf die Akzeptanz der Methode und auf die Lernmotivation diskutiert.

Escalada et al. (1996) befragten die Lernenden nach einer Videoanalyse-Lerneinheit über ihre Einstellung gegenüber dem Medium. Aus einer qualitativen Analyse der gegebenen Antworten ging hervor, dass die Durchführung der Videoanalyse wenig Vorerfahrung mit Computern bedarf. Außerdem standen die Lernenden dem anschaulichen Messprinzip, den Visualisierungsmöglichkeiten und der Interaktivität des Mediums positiv gegenüber. Ihre Neugierde wurde besonders dadurch geweckt, dass sie alltägliche Phänomene untersuchten. Geäußerte Kritik galt der höher empfundenen Bearbeitungszeit im Vergleich zur gewohnten, traditionellen Arbeitsweise ohne interaktive Videos.

Von den positiven Ergebnissen dieser qualitativen Studie inspiriert, führten die Autoren im darauffolgenden Jahr eine quantitative Untersuchung durch (Escalada & Zollman, 1997). Die Studierenden, die keine Erfahrung mit der Videoanalyse besaßen, konnten freiwillig an einem Videoanalyseprojekt teilnehmen, das als Alternative parallel zu einem traditionellen Seminar angeboten wurde. Die Autoren nutzten eine Klausur am Ende des Seminars als Indikator der Leistung und erhoben während dem Projekt die Sicherheit im Umgang und die Angst beim Arbeiten mit dem Computer sowie Einstellungen gegenüber dem Computer als Arbeitsmittel. Außerdem wurde zum Abschluss die Lerneffektivität durch den Einsatz der Medien und die notwendige Anstrengung zur Computernutzung beurteilt. Es konnte nachgewiesen werden, dass die Computer-Vorerfahrung weder Einfluss auf die Wahrnehmung der Videoanalyse-Lerneinheiten, noch auf die erzielte Leistung hatte, und dass sich die Lernenden nach dem Projekt sicherer im Umgang mit dem Computer fühlten. Die Ergebnisse aus der Vorstudie bzgl. der Motivationsförderung wurden durch qualitative Aussagen bestätigt. Das Klausurergebnis unterschied sich zum Ende des Seminars nicht zwischen Versuchs- und Kontrollgruppe, was die Autoren auf den geringen Anteil der Videoanalyse bezogen auf die Gesamtdauer der Veranstaltung zurückführten.

Palazzo und Schools (2008) setzten Videos mit einer starken Kontextanbindung (z.B. Looping im Freizeitpark, Kollisionsversuche mit Autos) im Unterricht ein. Die Auswertung eines Abschlussfragebogens ergab, dass die Schüler einen deutlichen Anwendungsbezug der Schulphysik erkannten, aber gemischter Meinung bzgl. der Lernwirksamkeit waren. Die

Lehrkräfte hingegen vermuteten einheitlich, dass durch das Verfahren eine Leistungssteigerung erzielt werden könne und stimmten darin überein, dass sich der zeitliche Aufwand der Vorbereitung außerhalb und der Durchführung innerhalb des Unterrichts lohne. Auch die Arbeiten der Arbeitsgruppe von Wee (Wee et al., 2012; Wee, 2012) sowie die Untersuchung von Rodrigues et al. (2001) legen nahe, dass sowohl Lernende als auch Lehrende instruktionale Mehrwerte der Videoanalyse gegenüber traditionellen Methoden erkennen. Neben erhöhter Motivation durch aktives Lernen und der Möglichkeit zur Analyse lebensnaher Bewegungen sahen die Lernenden durch die Videoanalyse das Potential, physikalische Arbeitsweisen durch Experimentieren und Modellieren aufzugreifen und fühlten sich näher an der Fachdisziplin. Lehrende dokumentierten, dass die Analyse und Darstellung von Daten zu angeregten Diskussionen und zum Überdenken bisheriger Vorstellungen führten.

### 3.2.3 Zusammenfassung und Bewertung der Untersuchungslage

Es gibt empirische Hinweise darauf, dass der Einsatz der Videoanalyse das Verständnis von Repräsentationen, insbesondere von Diagrammen, fördert (Beichner, 1996; Boyd & Rubin, 1996; Koleza & Pappas, 2008; Pappas et al., 2002; Struck & Yerrick, 2010). In Einklang mit Annahmen der CATLM zeigen die Ergebnisse, dass der größte Lernerfolg dann erzielt wird, wenn die Lerner-Aktivität am größten ist, d.h. wenn Experimentiervideos eigenständig aufgenommen, analysiert und die Ergebnisse diskutiert werden (Beichner, 1996; Koleza & Pappas, 2008). Über die Rolle der Echtzeit-Diagrammdarstellung (EZDD) für das Verständnis von Diagrammen liegen unterschiedliche Ergebnisse vor. Frühe Arbeiten kamen zu dem Schluss, dass EZDD nicht für die Lernförderung verantwortlich ist (Beichner, 1990; Brungardt & Zollman, 1995). Da die thematisierten Experimente aber von sehr geringer Komplexität waren (eindimensionale Bewegungen und Würfe), kann die untergeordnete Rolle von EZDD für das Lernen nicht verallgemeinert werden. Im Gegenteil führen Malgieri et al. (2014) korrekte physikalische Betrachtungen des Huygens-Pendels auf die zeitsynchrone Darstellung von Energie und Bewegung sowie auf Veranschaulichung der Schwerpunktbewegung zurück. Es ist anzunehmen, dass EZDD gerade bei komplexeren Sachverhalten zu einem besseren Verständnis der physikalischen Sachverhalte beitragen kann. Außerdem sind die von den Versuchsleitern bemängelten Defizite älterer Videoanalyse-Programme heutzutage überwunden, sodass sich lernförderliche Möglichkeiten ergeben, die zu den Zeitpunkten früherer Studien (wie Beichner, 1990 sowie Brungardt & Zollman, 1995) nicht vorhanden waren (beispielsweise eine Umsetzung des Kontiguitätsprinzips der CATLM ohne großen technischen Aufwand).

Ferner wurden positive Auswirkungen der Videoanalyse auf die Interpretation von Datentabellen und auf das Verständnis von Bezugssystemen festgestellt (Koleza & Pappas, 2008; Pappas et al., 2002), die auf den aktiven Umgang mit den jeweiligen Elementen zurückgeführt wurden (beispielsweise gezieltes Manipulieren von Bezugssystemen). Effekte bzgl. des Vektorverständnisses blieben gemäß der Studie von Kanim und Subero (2010) zwar hinter den Erwartungen zurück, aber kritisch betrachtet erscheint es schwierig, aus den berichteten Daten verallgemeinerbare Aussagen zur Lernwirkung zu treffen: Zum einen beruht die Quantifizierung des Lernerfolgs auf stark unterschiedlichen Tests vor und nach der Intervention und einem Vergleich mit den Post-Daten von Studierenden anderer Universitäten. Zum anderen wurden weder die Vorleistungen der Vergleichsgruppen berücksichtigt noch Aussagen zu der Lernmethode in der Vergleichsgruppe getroffen. Diese mangelnden methodischen Kontrollen zeichnen sich bei den meisten Arbeiten zur Lernwirksamkeit der Videoanalyse bzgl. des konzeptionellen Verständnisses ab (vgl. Tab. 3):

- Nur zwei von neun Untersuchungen bezogen Kontrollgruppen in die Untersuchung ein.
- In den beiden Vergleichsstudien wird keine Aussage über die Material- bzw. Inhaltsgleichheit der Untersuchungsgruppen getroffen.
- Nur eine Arbeit berichtet Testkennwerte bzw. nutzt einen standardisierten Test zur Leistungsmessung.

- In der einzigen Arbeit, in der standardisierte Instrumente vor und nach der Intervention genutzt werden, sind keine spezifischen Aussagen zur Lernwirksamkeit der Videoanalyse möglich, da die Videoanalyse nur einen geringen Teil der Intervention bildet (Wilhelm et al., 2007).
- Nur zwei von neun Arbeiten berichten Effektstärkemaße, um die praktische Relevanz des Lernerfolgs beurteilen zu können.

Auch die Ergebnisse zur Motivationswirkung beruhen nicht auf standardisierten Instrumenten, sondern begründen sich auf den Vergleich deskriptiver Statistiken oder auf qualitativen Daten. Ungeachtet dessen legen die berichteten Ergebnisse eine hohe Motivationswirkung nahe, die nicht nur von Versuchsleitern bzw. Lehrkräften vermutet, sondern auch von den Lernenden wahrgenommen wird (Escalada et al., 1996; Rodrigues et al., 2001; Wee, 2012; Wee et al., 2012). Die Befürchtung, dass Lernende durch den Einsatz der Methode überfordert seien, stellte sich als unbegründet heraus (Escalada & Zollman, 1997). Im Gegenteil fanden auch wenig computer-erfahrene Lerner einen schnellen Einstieg in das Verfahren, lobten dessen hohe Anschaulichkeit und verbesserten ihre Einstellung gegenüber dem Lernen mit dem Computer durch das Arbeiten mit der Videoanalyse. Negative Äußerungen bezogen sich auf den erhöhten Zeitaufwand durch das Arbeiten mit der Videoanalyse.

### Fazit

In Anbetracht der Vielzahl an Publikationen zu experimentellen Beispielen (vgl. 2.1.5) und der gut begründbaren Lernwirksamkeit der Videoanalyse (Kap. 3.1), fällt die empirische Forschungslage überraschend mäßig aus. Gerade im deutschsprachigen Raum wurde keine einzige Arbeit gefunden, die die Lernwirkung der Videoanalyse explizit untersuchte. Die Inhalte, die Gegenstand der empirischen Untersuchung waren, sind in den meisten Fällen Einführungsbeispiele aus der Kinematik. Die Analyse mehrerer Objekte oder von komplizierten Bewegungen bilden die Ausnahme. In fast allen Studien wurde das Potential der Videoanalyse, einen anschaulichen Zugang zu unanschaulichen Phänomenen zu bieten, demnach nicht ausgeschöpft. Außerdem beschränkte sich der Einsatzzeitraum des Verfahrens meist auf wenige Stunden. Dies sind Faktoren, die bei der Interpretation der Ergebnisse berücksichtigt werden müssen und die den Bedarf an weiterer empirischer Untersuchung aufzeigen.

Vergleichsstudien, die einem hohen methodischen Standard entsprechen (standardisierte Instrumente, Prä-Post-Design mit Kontroll- und Vergleichsgruppe, Materialgleichheit zwischen den Untersuchungsgruppen) betreffen entweder die kognitionspsychologische Fragestellung der Echtzeit-Diagrammdarstellung oder den Einfluss von PC-Vorkenntnissen auf die Lernwirkung (siehe Tab. 3). Zusammengefasst erlaubt es die empirische Forschungslage nicht, belastbare Aussagen über die Lernwirkung der Videoanalyse im Vergleich zu anderen Lehr-/Lernmethoden treffen zu können. Insbesondere fand bisher keine systematische Implementation der Videoanalyse in reguläre universitäre Lehrveranstaltungen außerhalb von Praktika statt und die Anwendung der Videoanalyse in den Studien beschränkte sich in fast allen Fällen auf Standardbeispiele der Kinematik. Diese Sachlage verwundert, wurden doch in Kap. 2.1.5 die vielfältigen Möglichkeiten dargelegt, insbesondere physikalisch anspruchsvollere Experimente in das Curriculum universitärer Lehrveranstaltungen einzubinden.

Das im folgenden Kapitel beschriebene Projekt *physics.move* setzt die Videoanalyse als essenzielles Element von vorlesungsbegleitenden, verpflichtenden Übungsaufgaben ein. Damit wird die wesentliche Erkenntnis aus der empirischen Forschungslage aufgegriffen, die Videoanalyse in ein instruktionsbasiertes Nutzungskonzept mit hoher Lerner-Aktivität einzubinden. Die Wirksamkeitsuntersuchung der Videoanalyse auf kognitive, metakognitive und affektive Variablen unter Verwendung psychometrisch validierter Testinstrumente in einer Vergleichsstudie über einen längeren Zeitraum (ein Semester) soll einen Teil der empirischen Lücke, die der breiten Entwicklungsgrundlage der Videoanalyse nicht gerecht wird, schließen.

## Kapitel 4

# Konzeption videobasierter Aufgaben und erste Untersuchungsergebnisse

Während die letzten beiden Kapitel die Grundlage, den Entwicklungsstand sowie den lerntheoretischen Hintergrund der Videoanalyse darstellten, wird in diesem Kapitel ein instruktionales Nutzungskonzept des Mediums für eine systematische Implementation in der Hochschullehre erarbeitet. In Kap. 4.1 und 4.2 werden videobasierte Aufgaben definiert und die konzeptionellen Eigenschaften dieses Aufgabenformats durch Beispiele illustriert. Mit Blick auf den Einsatz in vorlesungsbegleitenden Übungskursen zur Experimentalphysik 1 (Mechanik) wird ein Bezug zu und eine Abgrenzung gegenüber traditionellen Aufgaben hergestellt (Kap. 4.3). Für eine praktische Implementation in bestehende Übungsstrukturen wird im Sinne einer komplementären Vernetzung der videobasierten und traditionellen Aufgaben argumentiert.

Anschließend werden der derzeitige Entwicklungsstand sowie die Implementationsstrategie in die Übungsstruktur dargelegt (Kap. 4.4). Die videobasierten Aufgaben wurden im Wintersemester 2012/2013 zum ersten Mal in regulären Übungen zur Experimentalphysik 1 am Fachbereich Physik der TU Kaiserslautern eingeführt. Die Gelegenheit wurde genutzt, um das Studiendesign der geplanten Wirksamkeitsuntersuchung zu testen und erste Daten zur Lern- und Motivationswirkung zu erheben. Die wesentlichen Erkenntnisse dieser Pilotstudie werden in Kap. 4.5 dargestellt.

### 4.1 Videoanalyse-Aufgaben

Videoanalyse-Aufgaben (V-Aufgaben) sind auf einen physikalischen Inhalt fokussierte Übungsaufgaben, die aus einem vorgegebenen Video eines Realexperiments sowie einer Problemstellung bestehen. V-Aufgaben umfassen zusammenhängende theoretische und videobasierte experimentelle Teilaufgaben (Gröber et al., 2014) und sollen außerhalb von Lehrveranstaltungen, d.h. im Selbststudium, bearbeitet werden. Beispiele für theoretische Teilaufgaben sind die Herleitung von Formeln oder die Berechnung physikalischer Größen. Beispiele für videobasierte experimentelle Aufgabenanteile sind das Messen von Zeiten, Längen und Winkeln im Video, das Durchführen einer Videoanalyse (d.h. Bestimmung von zweidimensionalen Ort-Zeit-Koordinaten eines Objekts) oder die Verarbeitung von Messdaten (vgl. Klein et al., 2015a). Eine genaue Operatorensprache differenziert zwischen theoretischen und experimentellen Teilaufgaben. Beispiele für Operatoren, die zu experimentellen Tätigkeiten auffordern, sind „Messen“, „Aufnehmen“, „Auswerten“ oder „Experimentell bestimmen, nachweisen“. Im Folgenden sind zwei Beispiele für Videoanalyse-Aufgaben gegeben, die sich auf die beiden Wagen-Kugel-Experimente aus Kap. 2 und 3 beziehen (vgl. Abb. 2 und 4).

**I Wagen-Kugel-Experiment***(V-Aufgabe)*

Im Video wird eine Kugel mithilfe eines Schussapparats von einer fahrenden Modelleisenbahn abgeschossen.

- Skizzieren Sie qualitativ die Bahnkurve der Kugel aus Sicht eines mit der Modelleisenbahn mitbewegten Beobachters bzw. eines nebenstehenden Beobachters und kontrollieren Sie Ihre Skizze durch Messung beider Bahnkurven.
- Bestimmen Sie experimentell die Geschwindigkeitskomponenten der Kugel beim Abschuss (Zeitpunkt  $t = 0$ ) aus Sicht des nebenstehenden Beobachters und berechnen Sie daraus die theoretische Flugzeit und Flugweite der Kugel. Kontrollieren Sie die Ergebnisse im Video.
- Welche Voraussetzungen sind im Experiment erfüllt, dass die Kugel wieder in den Schussapparat zurückfällt? Stellen Sie die Geschwindigkeitsvektoren von Wagen und Kugel komponentenweise dar und interpretieren Sie die Darstellung.

**II Wagen-Kugel-Experiment auf der schiefen Ebene***(V-Aufgabe)*

Das Video zeigt, wie ein zunächst ruhender Wagen einen Hang mit Neigungswinkel  $\alpha$  hinunterrollt und bei Bewegungsbeginn ( $t = 0$ ) eine Kugel senkrecht zum Hang abschießt.

- Nehmen Sie die Bewegung der Kugel und des Wagens in einem mit der  $x$ -Achse entlang des Hangs orientierten Koordinatensystem auf und stellen Sie  $y(x)$ ,  $x(t)$  und  $y(t)$  von Kugel und Wagen jeweils in einem gemeinsamen Diagramm dar. Bezeichnen Sie die abgebildeten Bewegungsformen!
- Weisen Sie experimentell nach, dass die Beschleunigung des Wagens entlang der  $x$ -Achse konstant ist und bestimmen Sie diese. Überprüfen Sie den Wert mithilfe des Hangwinkels  $\alpha$ .
- Argumentieren Sie zunächst qualitativ und zeigen Sie dann quantitativ, dass es sich im Video nicht um einen Glücksschuss handelt, sondern dass die Kugel für beliebige Winkel  $0 \leq \alpha < 90^\circ$  wieder im Wagen landet. Begründen Sie, warum es für das Gelingen des Experiments notwendig ist, dass die Kugel senkrecht zur Unterlage abgeschossen wird, nicht jedoch, dass der Wagen zum Abschusszeitpunkt ruht.

Beide Aufgaben zeichnen sich durch einen knappen Einleitungstext aus, der die im Video dargestellte Situation kurz beschreibt und ggf. Symbole zur einheitlichen Bezeichnung physikalischer Größen einführt. Da die Anfangsbedingungen der Experimente und die geometrische Anordnungen der Objekte im Video eindeutig dargestellt sind, wird auf deren Versprachlichung verzichtet. Zusätzlich zur Aufgabenstellung erhalten die Studierenden ein Bild des Versuchsaufbaus, das ggf. Informationen enthält, die nicht im Video abgebildet werden (können), z.B. Aufnahmezeitpunkt der Kamera, Materialkonstanten oder Massen. In der ersten Teilaufgabe werden Diagramme genutzt, um die Bewegungen der Objekte zu veranschaulichen und einen Zugang zur komponentenweisen Betrachtung der Bewegungen zu erhalten. Die videoexperimentellen Tätigkeiten umfassen die Skalierung des Videos, die Aufnahme der Ortsvektoren sowie deren Darstellung. Durch das qualitative Skizzieren bzw. das Benennen der Bewegungsformen wird ein Theoriebezug hergestellt (Wechsel der Koordinatensysteme, Superpositionsprinzip). Die zweite Teilaufgabe besteht jeweils darin, aus dem Video gemessene Daten mit theoretischen Berechnungen zu vergleichen. Bei Aufgabe I stehen die Formeln der Wurfgesetze im Vordergrund, bei Aufgabe II der Zusammenhang zwischen Hangabtriebsbeschleunigung und Neigungswinkel. Die jeweils letzte Teilaufgabe testet das qualitative Verständnis des Wagen-Kugel-Experiments im Kern ab (Superpositionsprinzip). Aufgabe I (c) fordert darüber hinaus zur Interpretation einer Repräsentation auf, mit deren Hilfe den Lernenden der Zugang zum Verständnis des Experiments erleichtert wird. In Aufgabe II (c) hingegen werden qualitative und quantitative Argumente kombiniert, um ein umfassendes Verständnis des Problems anzuregen. Im Sinne der Problemdiskussion in Kap. 3.1.1 könnte die qualitative Betrachtungsweise von Teilaufgabe (c) alternativ als erste Teilaufgabe gestellt werden.

Da V-Aufgaben einen Teil des Instruktionsmaterials der Vergleichsstudie bilden (siehe Kap.

4.2), sind weitere V-Aufgaben im Anhang dargestellt (ab. S. 220). Um die konzeptionellen Eigenschaften der V-Aufgaben weiter zu verdeutlichen, werden im Folgenden einige der im Anhang dargestellten Aufgaben kurz beschrieben und durch Abbildungen (9–12) illustriert:

- Die Aufgabe „*Bewegung mit zeitweise konstanter Beschleunigung*“ (Aufgabe 2.1 in der Zählung des Materials) besitzt einführenden Charakter: Ein Wagen bewegt sich auf einer schiefen Luftkissenbahn zunächst hangabwärts, stößt elastisch mit einer Feder und kehrt infolgedessen seine Bewegungsrichtung um. Diese eindimensionale Bewegung soll in einem Diagramm dargestellt und das Ergebnis videoanalytisch überprüft werden. Neben einer qualitativen Deutung des Vorzeichens der Beschleunigung wird diese quantitativ bestimmt und durch eine Kontrollmessung verifiziert. Inhaltlich ist diese Aufgabe in der Kinematik verortet und fördert die Interpretation von Bewegungsdiagrammen sowie die Verknüpfung zwischen Diagramm und realer Bewegung.

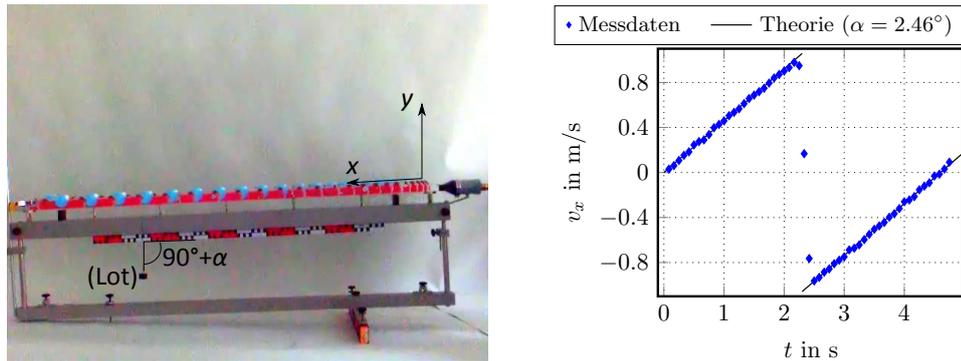


Abb. 9: Stroboskopbild zur V-Aufgabe 2.1 und Auswertung

- Das Video zur Aufgabe 4.2 „*Geschossgeschwindigkeit*“ zeigt den Schuss einer Luftgewehrkugel in eine Pendelmass, die anschließend zu oszillieren beginnt („Ballistisches Pendel“). Aus der Geschwindigkeit  $v'$  der Pendelmass unmittelbar nach dem inelastischen Stoß mit der Gewehrkugel, die entweder direkt gemessen oder aus dem maximalen Auslenkwinkel unter Anwendung des Energieerhaltungssatzes hervorgeht, kann bei bekannten Massen auf die Geschossgeschwindigkeit  $v$  geschlossen werden. Die Aufgabenstellung lässt offen, welche Größen zur Bestimmung der Geschwindigkeit  $v$  gemessen werden müssen, wodurch eine kognitive Auseinandersetzung mit dem Experiment angeregt wird.

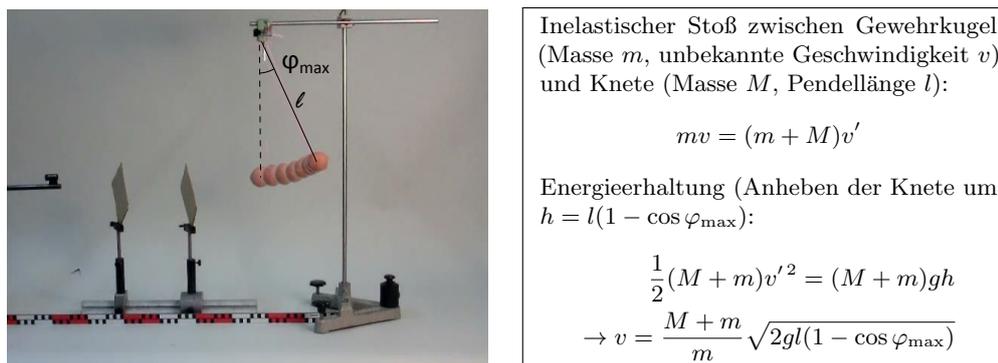
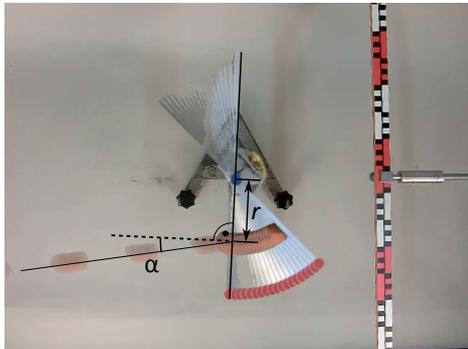


Abb. 10: Stroboskopbild zur V-Aufgabe 4.2 und Lösungsansatz

- Aufgabe 5.2 „*Schuss auf Platte*“<sup>1</sup> erfordert ähnliche Denkprinzipien wie Aufgabe 4.2: Aus der Rotationsgeschwindigkeit einer Platte unmittelbar nach dem inelastischen

<sup>1</sup>Diese Aufgabe ist in Gröber et al. (2014) veröffentlicht und wird dort ausführlich charakterisiert.

Stoß mit einer Knetmasse wird auf die Geschwindigkeit der Knetmasse vor dem Stoß geschlossen, die es im Anschluss durch eine direkte Messung zu überprüfen gilt. Schwierigkeiten der Aufgabe bestehen in der richtigen Anwendung der Erhaltungssätze, in der Verknüpfung des Drehimpulses des Systems nach dem Stoß mit dem linearen Impuls der Knetmasse vor dem Stoß und darin, das zusätzliche Trägheitsmoment der Knetmasse sowie deren Einfallswinkel  $\alpha$  zur Oberflächennormalen der Platte zu berücksichtigen.



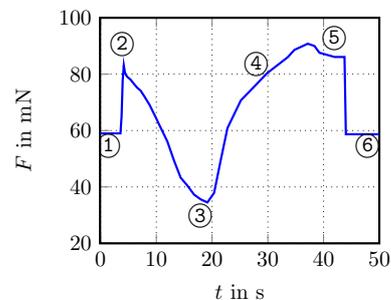
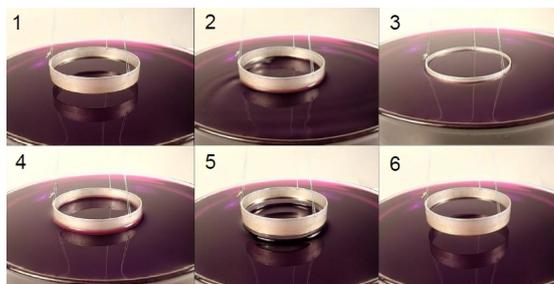
Die Bestimmungsgleichung für die Geschwindigkeit  $v$  der Knetkugel ergibt sich aus der Drehimpulserhaltung nach dem inelastischen Stoß zwischen der Knetkugel (Punktmasse  $m$ , Einfallswinkel  $\alpha$ ) und der Scheibe (Trägheitsmoment  $\Theta$ ) zu

$$mrv \cos \alpha = (\Theta + mr^2)\omega,$$

wobei  $r$  der Abstand zur Drehachse bezeichnet. Der Wert für  $\omega$  wird videoanalytisch bestimmt.

**Abb. 11:** Stroboskopbild zur V-Aufgabe 5.2 und Lösungsansatz

- In Aufgabe 7.2 „Oberflächenspannung — Ringmethode“<sup>2</sup> wird ein Experiment aufgegriffen, das typischerweise auch in der Vorlesung demonstriert wird: Ein Metallring wird in ein Wasserbad eingetaucht und anschließend herausgezogen, während mit einem Kraftmesser zeitsynchron die am Ring angreifende resultierende Kraft gemessen wird. Beim Herausziehen benetzt das Wasser zunächst die Seitenflächen des Rings und es bildet sich eine Lamelle an der Ringunterkante aus, die schließlich abreißt. Der Fokus der Aufgabe liegt auf dem genauen Beobachten des experimentellen Vorgangs, um den Kraftverlauf physikalisch begründet in Phasen zu gliedern: Der freie Ring (1) wird in das Wasserbad geführt (2) und taucht fast vollständig ein (3). Er wird wieder herausgezogen, wobei es zur Benetzung der Seitenflächen (4) und zur Ausbildung einer Lamelle (5) kommt, die schließlich abreißt (6).



**Abb. 12:** Bilderserie zur V-Aufgabe 7.2 und gemessene Kraft

Wie aus obiger Definition und aus den Beispielen ersichtlich wird, induzieren V-Aufgaben ein Wechselspiel zwischen Theorie und Experiment, welches den Lernenden durch die Aufgabenbearbeitung führt und ihn beim Bilden physikalischer Konzepte unterstützen soll. Die Aufgaben fördern weiter den Umgang mit multiplen Repräsentationsformen zum Vernetzen physikalischer Konzepte und gewähren dem Lernenden die Freiheit, weit offener mit dem Material zu arbeiten, als es in der Aufgabe geschrieben steht. Abhängig von der Muße der Lernenden besteht stets die Möglichkeit, Bewegungsvektoren darzustellen und als Problemlösungshilfe zu nutzen. Ebenso kann mithilfe der Videoanalyse erforscht werden, wie sich die Lage und die Orientierung des Koordinatensystems auf die Bewegungsbeschreibung auswirkt und es

<sup>2</sup>Das Video zum Versuch entstand im Rahmen einer dem Projekt zugeordneten Master-Arbeit (Marxen, 2013).

können alternative Lösungsansätze gesucht werden.<sup>3</sup> Wie das letzte aufgeführte Beispiel zeigt, bieten videobasierte Aufgaben eine gute Gelegenheit, in der Vorlesung thematisierte Demonstrationsexperimente in der Übung zu vertiefen. Dadurch entsteht eine lehrveranstaltungsübergreifende Verknüpfung des Lerninhalts. Es ist zu vermuten, dass dadurch die Wirksamkeit von Demo-Experimenten gesteigert werden kann, die ohne Verankerung nur unzureichend im Stande sind, das Lernen zu fördern (Crouch et al., 2004).

Zu einzelnen Videoexperimenten lassen sich spezifische Mehrwerte angeben, die durch den Einsatz der Videoanalyse gegenüber traditionellen Aufgaben erzielt werden, z.B. das Arbeiten mit Diagrammen oder die explizite Darstellung von Vektoren. Eine derartige Ausarbeitung zu 28 Experimenten der Mechanik (Experimentalphysik 1) findet sich in Gröber et al. (2014, Tab. A1).

## 4.2 Mobile Videoanalyse-Aufgaben

Mobile Videoanalyse-Aufgaben (mV-Aufgaben) unterscheiden sich von V-Aufgaben prinzipiell dadurch, dass die Experimentiervideos nicht vorgegeben sind. Stattdessen instruieren mV-Aufgaben die Planung, Durchführung und Videographie von Freihandexperimenten geringer Komplexität unter Verwendung von Alltagsmaterialien. Das Attribut „mobil“ deutet an, dass die Lernenden zur Videoaufnahme und Auswertung mobile Endgeräte wie Tablet-Computer oder Smartphones nutzen können (vgl. Kap. 2.2). Der Einleitungstext der Aufgabe führt zur Problemstellung hin und definiert gemeinsam mit Hinweisen im Aufgabenstamm den Offenheitsgrad der Aufgaben. Die begrenzte Bearbeitungszeit wöchentlicher Übungsaufgaben macht es notwendig, den Offenheitsgrad gering zu halten und den Lernenden Führung bzgl. der Experimentdurchführung zu geben — vor allem sollten die Ziele der Aufgabe eindeutig sein und wenig Interpretationsspielraum zulassen. Mobile Videoanalyse-Aufgaben werden durch Operatoren gekennzeichnet, die sich über die Auswertung des Videos hinaus auf die Planung, Durchführung und Aufzeichnung des Experiments beziehen (z.B. „Experiment planen, konzipieren, aufbauen, durchführen, modifizieren“ oder Anweisungen wie „Videographieren, Dokumentieren“). Im Folgenden wird das Wagen-Kugel-Experiment als mV-Aufgabe formuliert, um die unterschiedliche Intention zwischen V- und mV-Aufgaben herauszustellen.

### III Wagen-Kugel-Experiment mit Alltagsmaterialien (mV-Aufgabe)

Während sich eine Person mit konstanter Geschwindigkeit in horizontale ( $x$ -)Richtung bewegt, wirft sie einen Gegenstand (z.B. einen Ball) senkrecht nach oben ( $y$ -Richtung), führt die Bewegung fort und fängt den Gegenstand (bei „richtiger“ Durchführung) mühelos wieder auf.

- a) Videographieren Sie das Experiment aus Sicht eines außen stehenden Beobachters und erstellen Sie die Ort-Zeit-Diagramme der bewegten Person und des geworfenen Objekts. Weisen Sie experimentell nach, dass sich die Person tatsächlich mit konstanter Geschwindigkeit bewegt (Hinweis: Die Person könnte mit einem Skateboard auf einem ebenen Untergrund fahren).
- b) Erklären Sie qualitativ, warum und unter welchen Bedingungen der Gegenstand wieder aufgefangen werden kann. Nennen Sie dazu möglichst viele Parameter, die bei dem Experiment eine Rolle spielen und stützen Sie Ihre Argumentation durch aus dem Video generierte Darstellungen.
- c) Skizzieren Sie die Bahnkurve des Balls aus Sicht der bewegten Person und überprüfen Sie Ihre Skizze durch eine geeignete Aufnahme des Experiments.

<sup>3</sup>Aufgrund der vielen Möglichkeiten, die V-Aufgaben bieten, ist es für einen fairen Vergleich der Lernwirksamkeit gegenüber traditionellen Aufgaben wichtig, die Bearbeitungszeit der Aufgaben instruktional vergleichbar zu halten und diese zu erheben (siehe Kap. 10.4).

Die mV-Aufgabe greift die wesentlichen Elemente des Wagen-Kugel-Experiments auf (z.B. Superpositionsprinzip, Wechsel des Bezugssystems) und versetzt die Studierenden in eine noch aktivere Rolle während des Problemlösungsprozesses. Sie werden in das Experiment involviert, nehmen eigene experimentelle Daten selbst auf und werten diese aus. Dadurch besteht die Möglichkeit, dass die Studierenden Autonomie erleben, die motivationssteigernd wirkt (Klein et al., 2013a). Zunächst sollen weitere Beispiele die wesentlichen Eigenschaften der mV-Aufgaben verdeutlichen, bevor diese zusammengefasst werden. Ausformulierte Aufgabenstellungen befinden sich im Anhang (ab S. 220):

- Die Aufgabe „*Massenänderung während der Fahrt*“ (Aufgabe 4.2 in Zählung des Instruktionsmaterials) thematisiert die Auswirkung einer Massenzunahme bzw. -abnahme auf die Geschwindigkeit eines bewegten Objekts. Die Studierenden stellen z.B. Szenarien nach, in denen fahrende Experimentierwagen mit Massen bestückt werden, die zuvor eine Geschwindigkeitskomponente in Fahrtrichtung besitzen. Durch eine Serie einfacher Freihandexperimente werden die theoretischen Grundlagen der Impulserhaltung bei inelastischen Stößen, das zweite Newtonsche Axiom und die Rolle von Reibungskräften durch stark konzeptionell ausgerichtete Fragestellungen untersucht.
- Bei Aufgabe 5.1 „*Rollende Zylinder*“<sup>4</sup> wird das Beschleunigungsverhalten von Getränkedosen unterschiedlichen Trägheitsmoments auf einer schiefen Ebene in einer Experimentierserie quantitativ ausgewertet. Die Zusammenhänge zwischen den verschiedenen Energieformen und der resultierenden Beschleunigung werden für eine qualitative Begründung der experimentellen Befunde herangezogen.
- Bei Aufgabe 8.2 „*Fallbewegungen unter Reibungseinfluss*“ formulieren die Studierenden aufgrund halb-quantitativer Abschätzungen die Hypothese, dass ein in Luft fallender Papierkegel der Newtonschen — und nicht der Stokesschen — Reibungskraft unterliegt und testen diese experimentell. Ferner bestimmen sie den Widerstandsbeiwert.
- In der mV-Aufgabe „*Fadenpendel*“ (Aufgabe 9.2) basteln die Studierenden mit einfachen Mitteln ein schwingungsfähiges Fadenpendel (mit vernachlässigbarer Dämpfung) und variieren dessen Länge, um einen Bezug zur Kreisfrequenz herzustellen. Ferner wird das Pendel in einem nächsten Schritt derart modifiziert, dass eine signifikante Dämpfung messbar wird.

Der Fokus von mobilen Videoanalyse-Aufgaben liegt auf (1) der selbstständigen Variation experimenteller Parameter zur Hypothesenprüfung (Variablenkontrolle), (2) dem konfirmatorischen Prüfen theoretischer Vorhersagen durch lokale empirische Befunde, (3) der Präparation experimenteller Szenarien, (4) dem qualitativen Beobachten und Erklären von Phänomenen und (5) dem Kreieren von Hypothesen (Klein et al., 2015a). Aus den Aufgaben wird deutlich, dass zur Versuchsdurchführung in den meisten Fällen Alltagsmaterialien benötigt werden, deren Beschaffung kein Hindernis darstellen sollte (Papier, Dosen, Schnur, etc.). In Ausnahmefällen werden Materialien bereitgestellt (z.B. Experimentierwagen). Die vorgegebenen Experimentiervideos der V-Aufgaben hingegen enthalten Versuchsmaterialien, die nicht ohne großen Aufwand zugänglich wären (vgl. z.B. Aufgabe „*Geschossgeschwindigkeit*“) und grenzen sich inhaltlich von mV-Aufgaben dadurch ab, indem sie die Auswertemethodik akzentuieren und einen höheren mathematisch-formalen Anspruch besitzen. Die Bearbeitung von mV-Aufgaben fördert die Kooperation mit Kommilitonen, denn die Durchführung der Experimente (und deren Videographie) ist in Einzelarbeit mühsam. Außerdem regen sie die Kreativität an, indem sie den Lernenden Handlungsspielraum gewähren. Durch freies Experimentieren können Fragestellungen entstehen, die über die Intention der Aufgabe hinausgehen, über die diskutiert und denen nachgegangen werden kann.

<sup>4</sup>Die beiden Aufgaben 5.1 und 8.2 sind in Klein et al. (2015a) veröffentlicht.

### 4.3 Traditionelle Aufgaben

Unter traditionellen Aufgaben (T-Aufgaben) werden seit Jahren etablierte Physikaufgaben verstanden, die keinen experimentellen Anteil besitzen und hauptsächlich textbasiert sind. Beispiele sind im gedruckten Anhang in Form von korrespondierenden Aufgaben zu den V- und mV-Aufgaben sowie im digitalen Anhang (T-Aufgaben ohne korrespondierende videobasierte Aufgabe) zu finden. Typische Operatoren, die T-Aufgaben kennzeichnen, sind „Herleiten“, „Berechnen“, „Bestimmen“, „Abschätzen“ oder „Nähern“. Der Fokus traditioneller Aufgaben liegt größtenteils auf der physikalisch-mathematischen Beschreibung von Vorgängen, Phänomenen oder Systemen. Physikalische Ansätze und mathematische Fertigkeiten werden zur Bestimmung eines Ergebnisses (Formel / Zahlenwert) genutzt, das nicht weiter durch experimentelle Befunde verifiziert wird. T-Aufgaben haben im Curriculum durchaus ihre Berechtigung, denn sie formen das mathematisch-physikalische Begriffsverständnis, üben die präzise, mathematisch korrekte Beschreibung von Situationen ein und vermitteln mathematische Fertigkeiten seitens der Theoriebildung. Viele wichtige Kontrollmechanismen, z.B. die Analyse von Größenordnungen, Betrachtung von Grenz- und Spezialfällen oder die Prüfung physikalischer Einheiten erfordern eine theoretische Auseinandersetzung mit der physikalischen Problemstellung. Obwohl multiple Repräsentationsformen zum Verstehen physikalischer Zusammenhänge hilfreich sein können, kann der Fokus auf eine einzige Repräsentationsform (algebraische Ausdrücke) auch als Stärke traditioneller Aufgaben betrachtet werden, da kein Repräsentationswechsel erforderlich ist und somit die kognitive Belastung geringer ist. Das Lösen traditioneller Aufgaben ist nicht an Zusatzmaterial (Hardware / Software) geknüpft, in das sich Lernende erst einarbeiten müssen. Trotz diesen Stärken können T-Aufgaben beim Lernenden jedoch ein falsches Bild disziplinären Problemlösens erzeugen, da die experimentelle Komponente gänzlich ausgeklammert wird. V- und mV-Aufgaben hingegen akzentuieren das Wechselspiel zwischen Theorie und Experiment, um den dualen Problemlösungsprozess abzubilden (vgl. Klahr, 2000) und eine Kontrollmöglichkeit der Ergebnisse zu schaffen, die über mathematische Manipulationen hinausgeht.

Es sei darauf hingewiesen, dass V- und mV-Aufgaben keinen Ersatz für T-Aufgaben darstellen sollen oder können, sondern eine sinnvolle Ergänzung sind, die weiteren Zielen der Experimentalphysik-Lehrveranstaltung gerecht werden. Die beiden videobasierten Aufgabenformate können ohnehin nicht auf theoretische Aufgabenanteile verzichten, da eine Analyse des Experiments stets eine Auseinandersetzung mit dem theoretischen Hintergrund voraussetzt. In Klein et al. (2015a) werden V-, mV- und T-Aufgaben, die jeweils den gleichen Inhalt unter Nutzen ihrer jeweiligen Vorteile aufgreifen, beschrieben und bezüglich ihrer Lernziele gegenübergestellt.

### 4.4 Produktion von Videoanalyse-Aufgaben und Implementation in eine bestehende Übungsstruktur

Im Rahmen dieser Arbeit wurden gemeinsam mit Kollegen der Arbeitsgruppe ca. 40 V-Aufgaben und ca. 10 mV-Aufgaben produziert, die Ende des Jahres 2016 in Form eines Aufgabenbuches<sup>5</sup> erscheinen. Die Produktion einer V-Aufgabe umfasst die Erstellung des Aufgabentextes, die Aufnahme des Experiments sowie die Ausarbeitung einer Musterlösung. Als Ausgangspunkt und Orientierung zur Erstellung der videobasierten Aufgaben diente zunächst ein über Jahre im Physik-Fachbereich der TU Kaiserslautern angesammelter Pool (traditioneller) Übungsaufgaben, bevor gänzlich neue Aufgaben entwickelt wurden. Da in der Studie die Wirkung der neuen Aufgabenformate gegenüber traditionellen Aufgaben verglichen wird, wurde zu jeder V-Aufgabe bzw. mV-Aufgabe eine korrespondierende T-Aufgabe erstellt oder die bereits vorhandene T-Aufgabe überarbeitet, sodass die Aufgabenpaare bzgl. Inhalt,

<sup>5</sup>Arbeitstitel „*Smarte Aufgaben und Experimente zur Mechanik und Thermodynamik*“, Springer-Verlag (Heidelberg)

Schwierigkeitsgrad und Umfang vergleichbar sind (für Details siehe Instruktionsmaterial der Studie, Kap. 10.3).

Die videobasierten Aufgabenformate wurden im Rahmen der Studie derart in die Übungen implementiert, dass sie mindestens 50% aller Übungsaufgaben bildeten. Dabei führten in den ersten Wochen zunächst V-Aufgaben an die Auswertemethodik heran, bevor mV-Aufgaben das eigenständige Experimentieren erforderten. Die verwendeten Videos wurden zusammen mit einem Bild des Versuchsaufbaus über eine Online-Plattform zum Download angeboten. Hinweise, die bei der Einführung der neuen Aufgabenformate in eine bestehende Übungsstruktur beachtet werden sollen, sind in Klein et al. (2015a) zusammengefasst.

## 4.5 Pilotstudie im WS 12/13 und SS 13

Die Implementation der V-Aufgaben in die experimentalphysikalischen Übungsgruppen an der TU Kaiserslautern startete im Wintersemester 2012/2013 (WS 12/13), mV-Aufgaben wurden erstmals im darauffolgenden Sommersemester 2013 (SS 13) eingesetzt. Der Einsatz der Aufgaben in diesen Semestern diente grundsätzlich dazu, das konzipierte Instruktionsmaterial im Feld zu testen und Probleme bei der Implementation zu identifizieren, um diese für die folgende Hauptstudie beheben zu können. Darüber hinaus wurde die Gelegenheit genutzt, eine Begleitstudie durchzuführen, um aus ersten Daten Hinweise auf den Interventionserfolg zu erhalten, das konzipierte Studiendesign zu realisieren und Eindrücke über die allgemeine Akzeptanz der Aufgabenformate bei den Studierenden zu gewinnen. In diesem Abschnitt werden lediglich die wesentlichen Ergebnisse dieser Pilotierungsphase zusammengefasst, sowie die Konsequenzen für die Hauptuntersuchung und den weiteren Verlauf der Arbeit berichtet. Eine detaillierte Darstellung und Diskussion findet sich in der Vorveröffentlichung der Pilotstudie (Klein et al., 2015c sowie Klein et al., 2013b).

### 4.5.1 Methoden und ausgewählte Ergebnisse

Die Studierenden wurden in zwei Gruppen aufgeteilt, die Übungsblätter mit T-, V-, und mV-Aufgaben (Interventionsgruppe, IG) bzw. ausschließlich T-Aufgaben bearbeiteten (Kontrollgruppe, KG). Die Testzeitpunkte, auf die sich untenstehende Ergebnisse beziehen, lagen vor Bearbeitung des ersten und nach Bearbeitung des vierten Übungsblattes. Es wurde zu jedem Zeitpunkt Motivation (Realitätsbezug, Interesse / Engagement, Selbstkonzept), Leistung und Aufgabenschwierigkeit (Cognitive Load) mit standardisierten Instrumenten erhoben.<sup>6</sup>

Im WS 12/13 wurden keine signifikanten Unterschiede bzgl. Leistung, Motivation und Cognitive Load zwischen KG ( $N = 26$ ) und IG ( $N = 40$ ) festgestellt. Im SS 13 wurde ein großer Leistungszuwachs über die Zeit diagnostiziert ( $F(1, 15) = 10.86$ ,  $p < 0.01$ ,  $d = 1.57$ ), aber es ergaben sich keine Unterschiede zwischen KG ( $N = 7$ ) und IG ( $N = 9$ ). Die Intervention konnte das Selbstkonzept der IG gegenüber der KG signifikant stabilisieren ( $F(1, 15) = 5.98$ ,  $p < 0.05$ ,  $d = 1.09$ ), und auch bezüglich des Realitätsbezugs der Übungsaufgaben zeichnete sich eine Tendenz zu Gunsten der IG ab, die aber nicht statistisch signifikant wurde ( $p = 0.10$ ).

### 4.5.2 Diskussion und Konsequenzen für die Hauptstudie

Die Ergebnisse zeigten, dass das Arbeiten mit V-Aufgaben zu mindestens gleichwertigen Ergebnissen bzgl. Motivation und Leistung führte wie das Arbeiten mit T-Aufgaben.

- Positiv zu bewerten ist, dass keine Unterschiede bzgl. der empfundenen Aufgabenschwierigkeit (Cognitive Load) in beiden Gruppen gemessen wurden. Dies bedeutet, dass

<sup>6</sup>Auf eine Darstellung der Instrumente bzw. Untersuchungsmethoden wird an dieser Stelle verzichtet, um eine Redundanz zu Teil II bzw. III dieser Arbeit zu vermeiden.

die (kognitive) Belastung der Studierenden durch videobasierte Aufgaben vergleichbar mit der Belastung durch T-Aufgaben ist. Einzelne Studierende berichteten allerdings davon, deutlich mehr Zeit zum Lösen der V-Aufgaben als zum Lösen der T-Aufgaben aufwenden zu müssen. Um diese Eindrücke zu quantifizieren wurde entschieden, die *Bearbeitungszeit der Aufgaben* sowie die *allgemeine Beanspruchung durch das Studium* in der Hauptstudie zu erheben.

- In der Pilotstudie kamen zwei nahezu identische Motivationsfragebögen — adaptiert aus der Arbeit von Kuhn (2010) — zum Einsatz, die sich einmal auf die Vorlesung und einmal auf die Übung bezogen. Infolge der hohen Korrelation der mittleren Beurteilungen zwischen den beiden Bögen ( $r > 0.85$ ) wurde auf den Vorlesungsbogen in der Hauptstudie verzichtet.
- Die Stichprobengröße im SS 13 betrug lediglich  $N = 16$ . Es wurde entschieden, auf Begleitstudien im Sommersemester zu verzichten, da die geringe Stichprobengröße den Untersuchungsaufwand nicht rechtfertigt. Stattdessen wurden in den darauffolgenden Sommersemestern (SS 14, SS 15) videobasierte Aufgaben ohne Begleitstudie eingesetzt.
- Die Interpretation der Ergebnisse bzgl. den kognitiven Variablen muss berücksichtigen, dass die eingesetzten Testinstrumente zum Zeitpunkt der Pilotstudie noch nicht auf die eigentlichen Zielvariablen der Intervention abgestimmt waren. So wurde hauptsächlich auf den FCI zurückgegriffen, der das konzeptionelle Verständnis von Kräften erfasst (Hestenes, Wells & Swackhamer, 1992). Wie in Kap. 3.1.3 und 4.1–4.3 begründet wurde, stellen die konzipierten Übungsaufgaben aber vielmehr eine zentrale instruktionale Gelegenheit dar, die repräsentationalen Fähigkeiten von Studierenden zu fördern. Damit waren die eingesetzten Tests nicht im Stande, mögliche Lernwirkungen der Intervention messbar zu machen. Der große Fortschritt von Pilot- zur Hauptstudie bestand in der Entwicklung und Anpassung eines Tests zur Repräsentationskompetenz und in der Anpassung weiterer Testinstrumente, die zur Beantwortung der theoriebasierten Hypothesen und Forschungsfragen erforderlich sind:
  - 1) Der entwickelte Repräsentationstest war im zweiten Teil der Pilotstudie (SS 13) bereits in einer vorläufigen Version verfügbar und wurde eingesetzt, um die Schwierigkeit der konzipierten Testaufgaben einzuschätzen. Er wurde bis zur finalen Version (Abschluss der Hauptstudie) stets weiterentwickelt und validiert und wird in Kap. 5.1 präsentiert.
  - 2) Neben dem Repräsentationstest wurden einzelne FCI-Items aus inhaltlichen Gründen in einen separaten Kurztest aufgenommen (vgl. Kap. 5.2).
  - 3) Es wurde entschieden, Antwortsicherheiten (*confidence rating*) zu ausgewählten Leistungs-Aufgaben zu erheben (vgl. Kap. 5.3). Damit sind Rückschlüsse auf metakognitive Fähigkeiten der Lernenden möglich und es können Lernschwierigkeiten identifiziert werden. Ohne Mehraufwand können die Daten der Hauptstudie dazu beitragen, dieses neue physikdidaktische Forschungsfeld mit Erkenntnissen zu bereichern (siehe Kap. 8).
  - 4) Der Motivationsfragebogen von Kuhn (2010) wurde um mehrere Subskalen ergänzt, um bisher unberücksichtigte Facetten des Authentizitätsbegriffs aufzugreifen (siehe Kap. 5.4).
  - 5) Ein Neugierde-Fragebogen mit zwei Komponenten wurde aus der Literatur adaptiert (vgl. Kap. 5.5).
  - 6) Das Betreuerverhalten (Engagement, Motivation) wurde von den Studierenden beurteilt (vgl. Betreuerfragebogen, Kap. 5.7).

Im zweiten Teil dieser Arbeit wird die Instrumentenentwicklung und Validierung ausführlich dargestellt, die zur Erhebung der Zielvariablen in der Hauptuntersuchung (dritter Teil der

Arbeit) dienen. Teil II der Arbeit bildet somit einerseits die Grundlage zur Beantwortung der Wirksamkeitsfragen, ist aber andererseits inhaltlich von studienübergreifender Bedeutung: Es werden (1) Testinstrumente für vergleichbare Untersuchungen oder Folgestudien mit ähnlicher Forschungsausrichtung mit ausführlich dokumentierten Messeigenschaften bereitgestellt, (2) Leitfäden und Auswertungsmethoden zusammengefasst, die die klassische Testtheorie, die Item-Response-Theorie sowie die Analyse der Sicherheitsangaben zu Leistungstests betreffen sowie (3) Ergebnisse über metakognitive Fähigkeiten beim Umgang mit multiplen Repräsentationsformen generiert, die dieses junge Forschungsgebiet anreichern. In Kap. 3.2 wurde kritisiert, dass die meisten Testinstrumente und Leistungsindikatoren bisheriger Studien über die Wirksamkeit der Videoanalyse dem methodischen Präzisionsanspruch nicht genügen. Dieser Befund trifft besorgniserregender Weise auch für Vergleichsstudien über Innovationsansätze in Physik-Lehrveranstaltungen zu (Ruiz-Primo et al., 2011). Der hohe Stellenwert der Instrumentenentwicklung und Validierung in dieser Arbeit soll diesem Trend entgegenwirken.

## Teil II

# Instrumentenentwicklung und Validierung



## Kapitel 5

# Konzeption der Fragebögen und Tests

In diesem Kapitel wird die theoriegeleitete Konzeptualisierung der Untersuchungsvariablen dargestellt. Dazu zählen im affektiven Bereich die *empfundene Belastung*, die *Neugierde* und die *Motivation*, die nachweislich im Zusammenhang mit der Leistungsfähigkeit stehen (Hattie, 2009). Zur Erfassung des Lernfortschritts im kognitiven Bereich wurde das KiRC-Instrument entworfen, das den fachlichen Kerninhalt der Untersuchung — Kinematik und Repräsentationen — mithilfe mehrerer Aufgaben abbildet und die Variable *Repräsentationskompetenz* operationalisiert. Die genannten Variablen stehen seit einigen Jahren im Interesse der physikdidaktischen Lehr-/Lernforschung und der psychologischen Diagnostik, weshalb bei der Testkonstruktion auf eine Reihe von Vorarbeiten zurückgegriffen werden konnte.

Dieses Kapitel ist wie folgt gegliedert: Zunächst wird die Entwicklung des KiRC-Instruments sowie dessen inhaltlicher und struktureller Aufbau beschrieben (Kap. 5.1). Anschließend folgt die Darstellung eines zweiten, kürzeren Leistungstests zur Erfassung des Konzeptverständnisses zu Bewegungen (Kap. 5.2). Beide Leistungsinstrumente beinhalten eine Sicherheitsskala, deren Zweck in Kap. 5.3 erläutert wird. Es schließen sich Beschreibungen der Fragebögen zur Erfassung der affektiven Variablen (Motivation, Neugierde), der aufgabeninduzierten Belastung und des Betreuerverhaltens an (Kap. 5.4–5.7).

### 5.1 Leistungstest KiRC: Repräsentationskompetenz in der Kinematik

Das Ziel bei der Entwicklung des KiRC-Instruments bestand darin, über einen standardisierten Test zur Erfassung der *Repräsentationskompetenz* zu verfügen, der in der Studieneingangsphase ökonomisch einsetzbar ist und zuverlässige Ergebnisse liefert. Unter Repräsentationskompetenz wird die Fähigkeit verstanden, einzelne Darstellungsformen zu interpretieren, selbstständig zu erzeugen und zwischen verschiedenen Darstellungen zu wechseln (vgl. Kap. 3.1.3).

#### 5.1.1 Inhaltsdomäne, Entwicklungsbedarf und Vorarbeiten

Um Repräsentationskompetenz als latente Variable mithilfe von Testaufgaben valide messen zu können, muss ausgeschlossen werden, dass das fachliche (deklarative) Wissen über den inhaltlichen Gegenstand der Aufgaben einen entscheidenden Einfluss auf die Testergebnisse nimmt (sog. Konfundierung). Die grundlegende Schwierigkeit bei der Trennung deklarativen Wissens von repräsentationalen Fähigkeiten besteht darin, dass eine kognitive Interaktion

mit externen Repräsentationen stets in Bezug zu physikalischem Wissen geschieht. Eine näherungsweise konfundierungsfreie Messung wird dadurch erreicht, dass sich die im Test abgebildeten Inhalte auf *ein* Themengebiet beschränken, das den Lernenden möglichst gut vertraut ist. Ansonsten stellen Variationen im Fachwissen einen zu großen unkontrollierbaren Faktor bei der Testbearbeitung dar.

Bisher entwickelte Testinstrumente bilden die Erfassung von Repräsentationskompetenz nur unzureichend ab. Nieminen et al. (2010) stellen heraus, dass bestehende Konzepttests keine hinreichende Evaluation des Umgangs mit verschiedenen Repräsentationsformen ermöglichen. Zwar spiele diese Komponente disziplinärer Kompetenz bei vielen Testaufgaben implizit eine Rolle, eine systematische Untersuchung sei mithilfe bisher verfügbarer Instrumente aber nicht gegeben. Vielmehr konfundiere Repräsentationskompetenz als eine Art *versteckte Variable* innerhalb bestehender Tests mit der zu messen intendierten Variable (z.B. konzeptionelles Verständnis), was zu einer unerwünschten Beeinflussung der Testsummen führt. Nieminen und Kollegen entwickelten vor diesem Hintergrund den R-FCI, eine repräsentationsbezogene Variante des FCI, der ausgewählte FCI-Items in dreifacher Variation (symbolisch, graphisch, sprachlich) anbietet. Die Einschränkung auf *ein* Themengebiet (in diesem Fall Kräfte) und die systematische Variation der Repräsentationsform stellten eine wichtige Entwicklung zur gezielten Untersuchung repräsentationaler Fähigkeiten dar. Jedoch ist das repräsentationale Anforderungsniveau des Tests für Studierende zu niedrig (z.B. Vergleich von Balkendiagrammen mit sprachlichen Relationsaussagen).

Das KiRC-Instrument greift die Idee der Einschränkung auf ein Themengebiet auf, stellt aber höhere repräsentationale Anforderungen (siehe Kap. 5.1.2). Um den Einfluss des deklarativen Wissens weiter einzuschränken, wurde das Inhaltsgebiet Kinematik gewählt. Die Annahme ist gerechtfertigt, dass der fachliche Inhalt der Kinematik den Physik-Studierenden keine allzu großen Schwierigkeiten bereitet, sodass die Variation in den Testantworten weitestgehend auf die unterschiedlichen repräsentationalen Anforderungen der Items zurückgeführt werden kann. Abgesehen von dieser methodischen Begründung bietet sich diese Domäne für die Testentwicklung aus fachdidaktischer und testtheoretischer Perspektive in besonderer Weise an:

- Die Kinematik ist reich an Repräsentationen (vgl. Abb. 7), und an damit verbundenen Lernschwierigkeiten (Beichner, 1994; McDermott et al., 1987; Meltzer, 2005; Thornton & Sokoloff, 1998).
- Die Kinematik wird als physikalischer Grundstock der gesamten Mechanik und der darauf aufbauenden Fachsystematik betrachtet. In der Kinematik werden konzeptionelle und mathematische Denkweisen entwickelt, die Anschluss in vielen anderen physikalischen Fachgebieten finden. Beispielsweise wurde gezeigt, dass sich Fehlvorstellungen im Zusammenhang mit Rotationsgrößen auf Verständnisprobleme der Kinematik zurückführen lassen (Mashood & Singh, 2012).
- Aus praktischer Sicht ist ein Test über Repräsentationen im Themengebiet Kinematik für eine sehr große Zielgruppe relevant (z.B. Studierende anderer Studiengänge, Schülerinnen und Schüler).

Die Untersuchung des Verständnisses multipler Repräsentationen rückte in den vergangenen Jahren fortwährend in den Fokus physikdidaktischer Forschung. Auf Grundlage einer systematischen Literaturrecherche teilte Kohl (2007) die Arbeiten zu Repräsentationen in fünf verschiedene Kategorien ein: Arbeiten, die sich mit dem Verständnis (1) nur einer oder (2) multipler Repräsentationsformen beschäftigen<sup>1</sup>, (3) Arbeiten, die den Zusammenhang

<sup>1</sup>Ein Beispiel für (1) ist die Arbeit von Beichner (1994), in der ein Test zur Erfassung (ausschließlich) graphischer Fähigkeiten im Bereich Kinematik entwickelt wurde (TUG-K, Test of Understanding Graphs in Kinematics). Unter (2) zählen beispielsweise Konsistenzuntersuchungen beim Nutzen kohärenter Repräsentationen (wie in Abb. 7 dargestellt) oder Fragestellungen zur Auswahl von Repräsentationen, wenn dem Lernenden mehrere angeboten werden (Scheid, 2013).

zwischen Repräsentationskompetenz und Problemlösen untersuchen, (4) Arbeiten zum Einfluss dargebotener Kontexte auf das Testergebnis (z.B. physikalisch, mathematisch, neutral) und schließlich (5) Arbeiten, die sich mit Fragestellungen zur Metakognition beschäftigen. Wie im nächsten Abschnitt gezeigt wird, sind die Ergebnisse der Untersuchungen in die Entwicklung des KiRC-Instruments eingeflossen. Es wurden (1) die Erkenntnisse zu Lernschwierigkeiten mit einzelnen Repräsentationen beachtet sowie Aufgaben konstruiert, die (2) einen Repräsentationswechsel erfordern und die (3) nicht durch bloßes Erinnern gelöst werden können, sondern die Anwendung prozeduralen Wissens erfordern. Ferner wurde (4) eine Kontextsensitivität des Tests durch das Vermeiden von Kontexten weitestgehend ausgeschlossen und es wurden (5) Antwortsicherheiten zur ökonomischen Erfassung metakognitiver Fähigkeiten eingebunden (vgl. Kapitel 5.3).

### 5.1.2 Entwicklungsschritte und Inhalte des Test

Zunächst wurden die zentralen Inhalte und Konzepte der Kinematik im Zusammenhang mit verschiedenen Repräsentationsformen aus der Literatur aufgearbeitet, mit denen die Studierenden Schwierigkeiten zeigen. Die als wesentlich identifizierten Repräsentationsformen umfassen Diagramme, formal-algebraische Ausdrücke und bildhafte Darstellungen. Inhaltlich stellen sich die Zusammenhang zwischen den kinematischen Größen (Position, Geschwindigkeit und Beschleunigung), der Umgang mit Koordinatensystemen, die Interpretation von Bahnkurven sowie die Verwandtschaft zur Rotationskinematik als zentrale Elemente dar, die in den Items aufgegriffen wurden. Auch paradigmatische Anwendungsfälle (freier Fall, schiefer Wurf) wurden berücksichtigt. Ein initialer Pool mit 30 Aufgaben, inspiriert durch Vorarbeiten aus der Literatur, wurde in einer kleinen Stichprobe ( $N = 30$ ) eingesetzt, um den kognitiven Anspruch der Aufgaben anhand der Lösungswahrscheinlichkeiten zu beurteilen. Außerdem sollten die Studierenden einige ihrer Antworten schriftlich begründen. Die Begründungen von falschen Antworten waren eine gute Quelle zur Konstruktion plausibler Distraktoren, wurden aber nicht im Sinne einer Leistungsmessung ausgewertet. Einzelnen Studierenden wurden darüber hinaus ausgewählte Aufgaben in einem Interview vorgelegt und sie wurden gebeten, die Aufgaben durch „lautes Denken“ zu lösen. Diese Daten dienten ebenso als Quelle zur Konstruktion plausibler Antwortalternativen. Bevor das KiRC-Instrument erstmalig in der Hauptstudie eingesetzt wurde, gaben Experten Feedback über die fachliche Korrektheit, den Anspruch und die Relevanz einzelner Items bezüglich den Lehrinhalten und schlugen Verbesserungen für die Antwortalternativen vor. Die Testversion für die erste Hauptstudie (Wintersemester 2013/2014) bestand schließlich aus 18 Items. Für die zweite Hauptstudie (WS 14/15) wurde der Test überarbeitet, indem zu einfache Items gestrichen und weitere Items ergänzt wurden. Bei manchen Items wurden sprachliche Verbesserungen vorgenommen oder Antwortalternativen ersetzt.<sup>2</sup> Die insgesamt 25 verschiedenen KiRC-Items, die in den Studien eingesetzt wurden, befinden sich ausformuliert im Anhang (ab S. 205) und werden in Tabelle 4 (obere Hälfte) charakterisiert. Zwei Beispiele sind in Abb. 13 gezeigt. Während die Tabelle eine genaue Differenzierung zwischen den Items hinsichtlich Repräsentationsform und Inhalt trifft, werden die dahinterliegenden Grundideen in den folgenden Paragraphen dargelegt.

#### Algebraisch-formale Items

Bekanntlich haben Studierende Schwierigkeiten mit den formalen Zusammenhängen kinematischer Größen und den damit verknüpften mathematischen Operationen (Vektoroperationen, Differentiation, Integration; Barniol & Zavala, 2014). Das Item K1 adressiert diese Schwierigkeit durch die Frage nach der Interpretation einer nicht-konstanten Beschleunigung

<sup>2</sup>Die Änderungen sind im digitalen Anhang vollständig dokumentiert und mit Beispielen belegt. Im abgedruckten Anhang befinden sich ausschließlich die aktuellsten Item-Versionen, auf die sich auch der Text bezieht.

$\dot{v}(t) = bt$  ( $b = \text{const.}$ ). Die Studierenden müssen bei diesem Item die zugehörige Ortsfunktion identifizieren, wobei die dargebotenen falschen Antwortalternativen,  $x(t) = 1/2 b x^2$  und  $x(t) = bt$ , vertraut erscheinen mögen. Andere Beispiele sind die Items K18 und K20, bei denen arbiträre Positionsvektoren komponentenweise miteinander verglichen oder Aussagen über formale Ausdrücke getroffen werden.

### Items zum Graphenverständnis

Die Items K2, K4, K17, K19 und K21 hingegen wurden in das KiRC-Inventar aufgenommen, um das Graphenverständnis zu testen. Studierende treffen qualitative Aussagen über (Änderungen in der) Geschwindigkeit und Beschleunigung, wählen zueinander korrespondierende Graphen aus und bewerten formale Aussagen über graphisch dargestellte Bewegungen. Die Items K2, K4 und K19 sind an Aufgaben des TUG-K angelehnt (Beichner, 1994), benut-

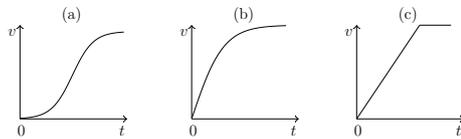
**Tab. 4:** Konstruktionsangaben zu den Items der Leistungstests

Item	Typ <sup>a</sup>	Kurzbeschreibung	Inhalt <sup>b</sup>	RF	Quelle <sup>c</sup>	HS1	HS2
<b>Repräsentationstest (KiRC)</b>							
K1	MC3	Zeitabhängige Beschleunigung	1	F	-	•	•
K2	MC3	$v(t)$ -Diagramm	1	D	TUG-K	•	•
K3	MC3	Freier Fall mit Reibung	2	B	-	•	•
K4	MC3	$x(t)$ -Diagramm	1	D	TUG-K	•	•
K5	MC3	Bewegungen vergleichen	1	B	FCI	•	•
K6	MC3	Bahnkurve ermitteln	3	B↔D	-	•	•
K7	MC3	Rennstrecke	4	B↔D	PISA	•	•
K8	MC3	Freier Fall mit Reibung	2	D	-	•	•
K9	MC3	Bahnkurve ermitteln	3	B↔D	-	•	•
K10	MC3	Bewegungen vergleichen	1	B	FCI	•	•
K11	MC3	Kurvenfahrt	5, 6	B	-	•	•
K12	MC3	Kugel auf Bahn	4	B↔D	DRZ	○	•
K13	MC3	Kugel auf Bahn	4	B↔D	DRZ	○	•
K14	MC3	Uhrzeiger	6	B↔F	RKCI	○	•
K15	MC3	Kugel auf Bahn	4	B↔D	DRZ	○	•
K16	MC3	Zeitabhängige Beschleunigung	1	B↔F	-	○	•
K17	MTF	$x(t)$ -Diagramm	1	D↔F	-	•	•
K18	MTF	Positionsvektor	5, 7	F	-	•	•
K19	MTF	Konstante Beschleunigung	1	D	TUG-K	•	○
K20	MTF	Positionsvektor	5, 7	F	-	•	•
K21	MTF	Zeitabhängige Beschleunigung	1	D↔F	-	•	•
K22	MTF	Rotiertes Bezugssystem	3, 5, 7	B↔F	-	•	•
K23	MTF	siehe K16	1	B↔F	-	•	○
K24	MTF	Freier Fall mit Reibung	2	F	-	○	•
K25	MTF	Rotationskinematik	4, 6	B↔F	RKCI	○	•
<b>Konzepttest</b>							
Kon1	MC5	Kugel fliegt aus Auto	4, 8	B	GRE	•	•
Kon2	MC5	Puck bekommt Stoß	4, 8	B	FCI	•	•
Kon3	MC5	Raumschiff beschleunigt	4, 8	B	FCI	•	•
Kon4	MC5	Flugzeug wirft Kugel ab	4, 8	B	FCI	•	•
Kon5	MC5	Kugel auf Kreisbahn	4, 8	B	FCI	•	•

*Bemerkungen:* Ausformulierte Items befinden sich im Anhang (ab S. 205). RF = Repräsentationsform (B = Bildhaft, D = Diagramm, F = Formal), HS = Hauptstudie.

<sup>a</sup>MC3/5 = Multiple-Choice-Item mit 3/5 Alternativen, MTF = Multiple-True-False-Item (siehe Kap. 5.1.3). <sup>b</sup>Inhalte: 1 = Zusammenhang  $x - v - a$ , 2 = Freier Fall mit Reibung, 3 = Bezugssysteme, 4 = Bahnkurven, 5 = Vektorzerlegung, 6 = Rotationskinematik, 7 = Wurfbewegungen, 8 = Superpositionprinzip und zweites Newtonsches Gesetz. <sup>c</sup>Quellen: TUG-K = Test of Understanding Graphs in Kinematics (Beichner, 1994), FCI = Force Concept Inventory (Hestenes et al., 1992), PISA = Aufgabe (PISA, 2007), DRZ = in Anlehnung an McDermott et al. (1987), RKCI = Rotational Kinematics Concept Inventory (Mashood & Singh, 2012), GRE = Graduate Record Examinations Physics (GRE, 2011).

**Item K8.** Ein Gegenstand fällt zum Zeitpunkt  $t = 0$  von dem Dach eines Hochhauses. Die Luftreibung führt nach einiger Zeit zu einer konstanten Fallgeschwindigkeit. Welches Geschwindigkeit-Zeit-Diagramm beschreibt die Fallbewegung korrekt?

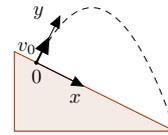


Wie sicher sind Sie sich?



sehr sicher    etwas Zweifel    eher unsicher    geraten

**Item K22.** Eine Kugel wird zum Zeitpunkt  $t = 0$  s von einem Hang aus mit der Anfangsgeschwindigkeit  $v_0$  senkrecht zur Hangebene nach oben geworfen. Welche Aussagen sind richtig bzw. falsch?



- |  |   |
|--|---|
|  | r    f                                      |
| (a) Die Anfangsgeschwindigkeit $v_0$ besitzt nur eine Komponente in $y$ -Richtung. | <input type="radio"/> <input type="radio"/> |
| (b) Die maximale Höhe $y_{max}$ wird nach der halben Wurfdauer erreicht.           | <input type="radio"/> <input type="radio"/> |
| (c) Die Bahnkurve $y(x)$ ist eine Parabel.   | <input type="radio"/> <input type="radio"/> |
| (d) Die Bewegung der Kugel ist in $x$ -Richtung beschleunigt.                      | <input type="radio"/> <input type="radio"/> |
| (e) Die Starthöhe und die Endhöhe sind identisch.                                  | <input type="radio"/> <input type="radio"/> |

**Abb. 13:** Zwei Beispielitems des Leistungstests KiRC. Item K8 (links) thematisiert den Geschwindigkeit-Zeit-Verlauf eines Objekts beim Fallen unter dem Einfluss der Luftreibung (Inhalt „2“, Repräsentationsform „D“; vgl. Tab. 4) und liegt im Multiple-Choice-Format vor (MC3). Item K22 (rechts) ist in Anlehnung an das Wagen-Kugel-Experiment konstruiert und liegt als Multiple-True-False-Item vor (MTF). Das Item akzentuiert inhaltlich die Rolle des Bezugssystems und die Dekomposition einer Wurfbewegung in  $x$ - und  $y$ -Richtung (Inhalte „3, 5, 7“), wobei ein Bezug zwischen der bildhaften Repräsentationsform (Schemazeichnung) und der formalen Beschreibung der Bewegung erforderlich ist („B $\leftrightarrow$ F“).

zen aber ein anderes Antwortformat (vgl. Kap 5.1.3). Die Verbindung solcher Graphen zu realen Bewegungen fällt den Lernenden besonders schwer (McDermott et al., 1987) und wird z.B. durch die Items K7, K12, K13 und K15 aufgegriffen. In diesen Aufgaben wird der Wechsel zwischen der bildhaften Repräsentation einer Kugel entlang einer beliebig geformten Rampe und dem zugehörigen Kinematik-Graph thematisiert. Die Antwortalternativen adressieren die Verständnisprobleme, die während der Entwicklung des TUG-K identifiziert wurden, zum Beispiel die Verwechslung von Position und Steigung, die Verwechslung von Geschwindigkeit und Beschleunigung oder die Darstellung einer Richtungsänderung im  $v(t)$ -Diagramm.

**Bildhafte Items**

Items, die die Interpretation von Punktspuren erfordern (Stroboskopbilder), sind K3, K5, K10, K16 und K23. Items dieser Art stellen die Bewegung eines Objekts durch Punkte mit äquidistantem zeitlichen Abstand dar und werden auch im FCI genutzt (Hestenes et al., 1992). Die Interpretation der Bewegung erfolgt semi-quantitativ und setzt ein konzeptionelles Verständnis der kinematischen Größen Ort, Geschwindigkeit und Beschleunigung sowie des Änderungsbegriffs voraus.

Es sei explizit darauf hingewiesen, dass bildhafte Abbildungen nicht auf Punktspuren beschränkt sind. Die o.g. Items K7, K12, K13 und K15 beispielsweise stellen experimentelle Situationen bildhaft dar.

**Inhaltliche Differenzierung der Items**

Die Items K6, K9 und K22 betonen die besondere Rolle des Koordinatensystems bei der Beschreibung einer Bewegung, womit ebenfalls Schwierigkeiten verbunden werden (Bowden et al., 1992; Sayre & Wittmann, 2008). Besonders die Beschreibung eines schiefen Wurfs in einem rotierten Koordinatensystem, in dem die vertikale Achse nicht in Richtung des Erdbeschleunigungsvektors orientiert ist (K22, vgl. Abb. 13 rechts), erwies sich im Vorfeld der Untersuchung als besonders anspruchsvoll (Klein et al., 2015d).

Die Idee von Nieminen et al. (2012) aufgreifend wurde ein Aufgabentripel entwickelt, welches den gleichen physikalischen Inhalt in verschiedenen Darstellungsformen präsentiert und somit der Erfassung repräsentationaler Kohärenz dient. Die Items K3, K8 und K24 besitzen allesamt einen identischen Itemstamm, während die Antwortalternativen als Stroboskopbilder, als Graphen oder als formale Aussagen über die Bewegung gegeben sind. Inhaltlich handelt es sich um eine eindimensionale Bewegung mit nicht-konstanter Beschleunigung — dem freien Fall eines Gegenstands unter dem Einfluss der Luftreibung, vgl. Abb. 13 (links).

Schließlich thematisieren die Items K11, K14 und K25 die Rotationskinematik mit Fokus auf den Größen Winkelgeschwindigkeit und Winkelbeschleunigung. Die Items K14 und K25 sind eng an die Items von Mashood und Singh (2012) angelehnt.

### Kontexte, Sprache und Bezug zur Instruktion

Bei der Itemkonstruktion und Adaption wurde darauf geachtet, Kontexte zu vermeiden, da sich diese bekannterweise auf die Problemlösung auswirken können (und damit mit der zu messenden Kompetenz konfundieren; vgl. Roth & Bowen, 2001). Die Items beziehen sich auf abstrakte Inhalte und verwenden eine präzise formale Sprache, was KiRC von anderen Tests aus der physikdidaktischen Forschung (wie z.B. dem FCI) unterscheidet. Der Test ist nicht auf die Interventionssequenz videobasierter Aufgaben zugeschnitten, aber hinsichtlich Testlänge mit dem Einsatz in einer Unterrichtsstunde (Schule) oder Übungsstunde (Hochschule) kompatibel. Einige Items zeigen eine Ähnlichkeit zu den Inhalten des Instruktionmaterials. So erinnert das in Aufgabe K21 dargestellte  $a(t)$ -Diagramm an den Beschleunigungsverlauf eines Bungee-Springers, der Gegenstand von Aufgabe 3.2 der Instruktionssequenz ist (Aufgabe „Bungee-Springer“, vgl. Materialien im Anhang, ab S. 220). Die Situation in Item K22 ist der von Aufgabe 3.1 sehr ähnlich (Aufgabe „Schuss von schiefer Ebene“). Es wurde bei der Testkonstruktion darauf geachtet, dass bei der Bearbeitung der Testitems keine Untersuchungsgruppe durch instruktionsspezifische Inhalte bevorzugt wird.

#### 5.1.3 Itemformate und Testsumme

Der Leistungstest besitzt ein geschlossenes Antwortformat. Die Items 1–16 liegen im Multiple-Choice-Format mit einer richtigen und zwei falschen Alternativen vor (MC3), wohingegen die Items 17–25 ein Multiple-True-False-Format (MTF) aufweisen. MTF-Items bestehen aus einem einleitenden Text (Itemstamm) und fünf darauffolgenden Aussagen, die als entweder richtig (true) oder falsch (false) beurteilt werden müssen (vgl. Abb. 13). Ein MTF-Item besteht also aus fünf True-False-Items (TF). MTF-Items sind von den Lernenden in kurzer Zeitdauer zu beantworten und bieten sich vor allem dann an, wenn mehrere voneinander unabhängige Kurzaussagen zu einem Themengebiet gestellt werden sollen. Im Vergleich zu komplexen Mehrfachwahlaufgaben, wie sie zum Beispiel im TUG-K eingesetzt werden, erreichen MTF-Items höhere Reliabilitäten (Frisbie, 1992) und werden von Studierenden im Allgemeinen bevorzugt (Mobalegh & Barati, 2012). MC3-Items hingegen ermöglichen eine differenziertere Untersuchung von Fehlvorstellungen und Lernschwierigkeiten, wenn dementsprechend plausible Distraktoren dargeboten werden. In einer ersten Version des Tests wurden in Anlehnung an vergleichbare Instrumente zunächst fünf statt drei Antwortalternativen angeboten, bevor die beiden Distraktoren mit der geringsten Attraktivität in weiteren Entwicklungsschritten verworfen wurden. Obwohl die meisten Konzepttests mit vier oder fünf Antwortalternativen arbeiten, hat die Einschränkung auf drei Alternativen im Allgemeinen keinen signifikanten Einfluss auf die psychometrische Qualität der Testwerte (Rodriguez, 2005).

Als Maß für die Leistungsfähigkeit einer Person wird die Testsumme (= Summe aller Items) verwendet.<sup>3</sup> Beinhaltet ein Test zwei verschiedene Itemformate (MC3 und MTF),

<sup>3</sup>Die Frage, inwiefern es gerechtfertigt ist, bloße Testsummen zu betrachten und nicht zu berücksichtigen, welche Aufgaben gelöst wurden, wird in Kap. 7 aufgegriffen.

werden zwar deren komplementären Vorteile genutzt, aber eine Gleichbehandlung der Testkomponenten bei Bildung einer Testsumme erscheint fragwürdig: Erstens besitzen die beiden Formate unterschiedliche Ratewahrscheinlichkeiten, zweitens können den Items formatspezifische Lösungsstrategien zugrunde liegen (z.B. Ausschluss von Distraktoren statt Kenntnis der richtigen Antwort) und drittens bleibt unklar, ob die Wertigkeit richtiger Antworten formatübergreifend vergleichbar ist. Eine getrennte Analyse beider Testhälften würde diese Probleme umgehen, aber weniger zuverlässige Messungen zur Folge haben und nicht alle Inhalte des Tests in einer Testsumme abbilden. Die Auswertung und der Ergebnisbericht wären aufwendiger. Als Konsequenz wird für die Zusammenführung beider Testhälften eine Verrechnungsvorschrift eingeführt, die die Testsummen beider Hälften unter Berücksichtigung eines Gewichtungsfaktors addiert. Bezeichnet  $T_i$  ( $i = 1, 2$ ) die Summe der richtigen Antworten aus der  $i$ -ten Testhälfte ( $i = 1$ : MC3-Items,  $i = 2$ : TF-Items), dann sei die gesamte Testsumme  $T$  definiert als

$$T = T_1 + \kappa T_2. \quad (1)$$

Der dimensionslose Parameter  $\kappa \in [0, \infty)$  ermöglicht eine beliebige relative Gewichtung beider Testhälften zu einer gemeinsamen Testsumme. Insbesondere ermöglicht er es, dass nur eine Testhälfte effektiv zur Testsumme beiträgt ( $\kappa = 0$  bzw.  $\kappa \rightarrow \infty$ ) und dass beide Testhälften gleich gewichtet werden ( $\kappa = 1$ ). Für die Itemanalysen in Kap. 6 und 7 wird die heuristische Gleichgewichtung der Items verwendet. Für die Vergleichsanalysen wird ein Verfahren zur Behandlung gemischter Testsummen entwickelt (Testmischungsanalyse, vgl. Kap. 10.5.5).

## 5.2 Konzepttest zu Bewegungen

Zusätzlich zum Repräsentationstest KiRC wurde ein zweites Inventar zur Leistungsmessung eingesetzt, der im Folgenden als Konzepttest bezeichnet wird. Der Konzepttest, dargestellt im Testbogen im Anhang und zusammengefasst in Tab. 4 (unten), besteht aus fünf Items, wovon vier aus dem in der physikdidaktischen Forschung wohl etablierten Force Concept Inventory stammen (FCI; Hestenes et al., 1992). Ein inhalts- und strukturähnliches Item aus dem amerikanischen Standardtest GRE wurde ergänzt. Die Aufgaben sind kontextbezogen, zielen auf Fehl- und Alltagsvorstellungen ab, sind in Alltagssprache formuliert und liegen allesamt in der bildhaften Repräsentationsform vor. Die komplementären Eigenschaften des Konzepttests und KiRC gewährleisten damit die systematische, möglichst konfundierungsfreie Untersuchung des Konzeptverständnisses und der Repräsentationskompetenz vor dem gemeinsamen fachlichen Hintergrund der Kinematik.

Im Speziellen erfordern die Items des Konzepttests eine Vorhersage der Trajektorien von Körpern, nachdem auf sie in verschiedener Art Einfluss genommen wird: Item Kon1 diskutiert die Bahnkurve eines Balles, der aus einem fahrenden Fahrzeug geworfen wird; Item Kon2 stellt die Bewegung eines zunächst gleichförmig bewegten Pucks dar, der einen Stoß bekommt; in Item Kon3 beschleunigt ein Raumschiff senkrecht zu seiner ursprünglichen Bewegungsrichtung; Item Kon4 zeigt ein Flugzeug, aus dem eine Kugel fällt, und in Item Kon5 reißt die Schnur, an der eine Kugel im Kreis geschwungen wird. Alle Items erfordern die Kenntnis des Superpositionsprinzips sowie ein konzeptionelles Verständnis des zweiten Newtonschen Gesetzes. Neben der richtigen Trajektorie werden vier falsche Alternativen präsentiert (MC5-Format). Diese sind an typische Fehlvorstellungen angelehnt, die im Rahmen traditionsreicher Forschung über Bewegungen identifiziert wurden (Kaiser et al., 1985; McCloskey, 1983). Auf eine Reduktion auf drei Antwortalternativen wurde verzichtet.

## 5.3 Erhebung von Antwortsicherheiten

Die Multiple-Choice-Items der beiden Leistungsinstrumente werden von einer Beurteilungsskala begleitet, auf der die Studierenden ihre Zuversicht in ihre Lösung in vier Stufen

(„geraten“ bis „sehr sicher“) ausdrücken können, vgl. Abb. 13 (unten links). Eine solche Beurteilungsskala wird auch Sicherheitsskala genannt (*confidence scale*, vgl. Pieschl, 2008). Die Erhebung von Antwortsicherheiten zu leistungsbezogenen Aufgaben hat eine lange Tradition in der psychologischen Diagnostik und nimmt erst seit jüngster Zeit Einzug in die physikdidaktische Forschung. Im Folgenden wird ein kurzer Abriss über diese Fragetechnik gegeben und in Aussicht gestellt, welches Ziel mit der Erhebung von Antwortsicherheiten im Allgemeinen und in dieser Arbeit im Speziellen verfolgt wird.

### 5.3.1 Mehrstufige Multiple-Choice-Items und Ziele von confidence rating

Die Technik, zu einem Multiple-Choice-Item neben der Antwort auf die eigentliche Frage zusätzliche Informationen über Denk- oder Problemlösungsprozesse zu erheben, ist ursprünglich auf Tobin und Capie (1981) zurückzuführen (vgl. Treagust, 1986, S. 3). Sie boten der Testperson zu jeder Alternative mehrere mögliche *Begründungen* an, die beispielsweise typische Denkfehler adressieren.<sup>4</sup> Diese Fragetechnik erhebt zwar zusätzliche Informationen über den Denkprozess, gibt aber nur wenig Auskunft darüber, ob die Testperson die Antwort wirklich wusste, schlicht erraten hat oder, feiner differenziert, wie sicher sie sich ihrer Antwort war. Deshalb wurde als nächste Stufe das Erheben fragebezogener Antwortsicherheiten (sog. *confidence rating* oder *confidence scoring*) entwickelt, das seinen Ursprung in der Psychologie findet (Echternacht, 1971).<sup>5</sup> Die Erhebung von Antwortsicherheiten ist aus praktischer Sicht der Testkonstruktion und Testdurchführung sehr ökonomisch, denn für jedes Item wird stets die gleiche Frage „Wie sicher sind Sie sich Ihrer Antwort?“ gestellt, und verlängert die Testzeit nur unerheblich. Im Sinne einer klassischen Item- und Skalenanalyse werden die Sicherheitsabfragen als eigenständige Beurteilungs-Items aufgefasst und separat von den eigentlichen Antworten auf die physikalische Fragestellung analysiert. Die Untersuchung des Zusammenhangs zwischen der Richtigkeit der Antwort und der angegebenen Sicherheit wird in dieser Arbeit in einem eigenständigen Kapitel (siehe Kap. 8) aufgegriffen.

Das Ziel bei der Erhebung von Antwortsicherheiten ist es im Allgemeinen, auf metakognitive Fähigkeiten der Testpersonen zu schließen. Annahme ist, dass die Angabe einer Antwortsicherheit die Reflexion der eigenen Leistung hinsichtlich der unmittelbar gelösten Aufgabe erfordert (Lindsey & Nagel, 2015). Diese Selbsteinschätzung ist nicht mit einer allgemeinen Zuversicht in disziplinbezogene Fähigkeiten zu verwechseln (Selbstkonzept, vgl. Kap. 5.4.2), aber es ist anzunehmen, dass zwischen diesen beiden Konstrukten eine Verwandtschaft besteht. Das Erkennen von Defiziten durch die Selbstreflexion des eigenen Verständnisses wird als Voraussetzung gesehen, den Lernprozess effektiv zu strukturieren und zu steuern (*selbstreguliertes Lernen*, Garavalia & Gredler, 2002; Rickey & Stacy, 2000; Schunk & Pajares, 2004). Studenten, die ihre eigene Fähigkeit zutreffend beurteilen können, verfügen über die Möglichkeit, zielgerecht zu lernen. Das Planen und Überwachen des Lernfortschritts führt zu besseren Leistungen (Boekaerts & Rozendaal, 2006).

### 5.3.2 Confidence rating in dieser Arbeit

In dieser Arbeit werden Antwortsicherheiten erhoben, um die metakognitiven Fähigkeiten von Studierenden zu diagnostizieren. Die Leitfrage dabei ist, wie zutreffend Studierende ihre eigene Fähigkeit bzgl. repräsentationsbezogenen Aufgaben einschätzen und ob die Genauigkeit der Selbsteinschätzung von dem Leistungsniveau der Studenten oder vom Geschlecht abhängt. Die Untersuchung knüpft an bestehende physikdidaktische Forschung über metakognitive Fähigkeiten an, und reichert diese vor dem Hintergrund multipler Repräsentationen mit

<sup>4</sup>Solche Items werden *two-tier-items* bzw. *assertion-reasoning-tasks* genannt, zu dt. etwa „zweistufiges Item“ bzw. „Auswahl-Begründungs-Items“.

<sup>5</sup>Tests, die alle drei Stufen nutzen werden als *three-tier-tests* bezeichnet und sind in der Leistungsdiagnostik wenig verbreitet (Caleon & Subramaniam, 2010).

Evidenzen an. Die bisherigen Erkenntnisse in diesem noch jungen Forschungsgebiet der Physikdidaktik werden in Klein et al. (2016) berichtet und sind in Kap. 8 zusammengefasst.

Wie in Kapitel 5.1 dargelegt wurde, sind in der Literatur an vielfacher Stelle Verständnisschwierigkeiten beim Umgang mit verschiedenen Repräsentationen dokumentiert. Ein ökonomischer, *quantitativer* Nachweis dieser Schwierigkeiten durch Antwortsicherheiten blieb bisher aus und wird in dieser Arbeit untersucht. Leitfrage hierbei ist, ob Items existieren, bei denen sich Studierende überdurchschnittlich sicher sind, obwohl sie die Items falsch lösten.

Es wird nicht untersucht, ob sich die Selbsteinschätzungen der Studenten zeitlich ändern oder ob videobasierte Aufgaben einen Einfluss auf die gegebenen Sicherheiten haben. Die im Rahmen der Studie entwickelten Instruktionsmaterialien (Videoanalyse-Aufgaben) zielen nicht explizit darauf ab, einen Einfluss auf metakognitive Prozesse zu nehmen. Perspektivisch ist es aber denkbar, Strategien zur Förderung metakognitiver Prozesse im Rahmen der universitären Physikausbildung instruktional zu fördern — auch mit videobasierten Aufgaben. Die Erkenntnisse dieser Untersuchung sollen einen Grundstein dafür legen (vgl. Kap. 8).

## 5.4 Fragebogen zur Erfassung verschiedener Motivationskomponenten

Der konzipierte Fragebogen zur Erfassung der Motivation basiert im Wesentlichen auf dem MAI-Fragebogen<sup>6</sup>, der u.a. von Kuhn (2010) unter Berücksichtigung bestehender Testinventare entwickelt und eingesetzt wurde (vgl. Bolte & Stork, 1996; Häußler & Hoffmann, 1995). Ziel des MAI-Fragebogens ist die Erfassung bestimmter Motivationsfaktoren vor und nach experimentell gegebenen Lernsituationen. In der ursprünglichen Version liegen hochreliable und gut validierte Items zur Erfassung des *Realitätsbezugs*, der *intrinsischen Motivation / des Engagements* sowie des *akademischen Selbstkonzepts* vor, die sich inhaltlich auf Physikunterricht in der Schule beziehen (Kuhn, 2010, S. 332). Unter Beibehaltung dieser Motivationskomponenten wurden einige Items hinsichtlich des Untersuchungsgegenstandes dieser Studie modifiziert, indem sie einen eindeutigen Gegenstandsbezug zu den Aufgaben der Experimentalphysik sowie einen klaren Situationsbezug zu den Übungen erhielten. Außerdem wurden Items zu den bestehenden Komponenten ergänzt sowie zwei neue Komponenten hinzugefügt, die wichtige motivationale Faktoren speziell des Physikstudiums abbilden. Alle Items liegen als Aussagen mit numerischer, sechsfach abgestufter Beurteilungsskala vor, die die Werte 1 („Die Aussage trifft voll und ganz zu“) bis 6 („Die Aussage trifft überhaupt nicht zu“) umfasst. Im Folgenden wird die Konzeptualisierung der verwendeten Fragebögen mit beispielhaften Items dargestellt. Für einen besseren Lesefluss werden ausformulierte Items als Fußnote dargestellt, wenn deren Inhalte nicht unmittelbar aus dem Text hervorgehen. Die vollständigen Fragebögen befinden sich mit Quellenangaben im Anhang (Tab. A-1).

### 5.4.1 Interesse / Engagement (IE)

Das Interesse ist ein wichtiger motivationaler Faktor im Studium, eine Bedingung für erfolgreiche Lernprozesse und das Ergebnis universitärer Lehre (Krapp, 1989). Konsequenterweise fordern Krapp et al. (1993) eine höhere Berücksichtigung des Interessenkonstrukts in der empirischen Lernforschung gerade im Hochschulbereich und verweisen auf die nachgewiesenen Zusammenhänge zwischen Interesse und akademischen Leistungen. Außerdem seien Phänomene wie Studienabbruch und -Fachwechsel zum Teil auf das Interesse zurückzuführen.

Interesse wird als affektive Variable aufgefasst, die mehrere, voneinander nicht unabhängige Komponenten besitzt. Die IE-Skala von Kuhn (2010), die dort „intrinsische Motivation / Engagement“ bezeichnet wird, zielt überwiegend auf die Gefühle ab, die mit

<sup>6</sup>MAI = Modified-Anchored-Instruction; Der Fragebogen bezieht sich auf einen Realisierungsansatz dieser Theorie.

dem Physikunterricht verbunden werden<sup>7</sup> sowie auf einen freiwilligen, zusätzlichen Lernaufwand<sup>8</sup> (vgl. Kuhn, 2010, S. 332–333, IE-Items). Die hier eingenommene theoretische Position erfolgt in Anlehnung an das Interessenkonstrukt von Krapp et al. (1993), die dem Interesse neben dieser *intrinsischen* und *emotionalen* Komponente auch eine *wertbezogene* Komponente beimisst. Die genannten Autoren entwickelten den *Fragebogen für Studieninteresse* (FSI), aus dem einige Items übernommen wurden (vgl. Tab. A-1). Die emotionale Komponente wird zum Beispiel durch das Item M35 erfasst, welches die Übungsaufgaben mit positiven Gefühlen verbindet<sup>9</sup>, während ein Itembeispiel für die wertbezogene Komponente M10 darstellt.<sup>10</sup> Auch Item M28, welches das völlige Aufgehen in einer Tätigkeit beschreibt, gilt als Indiz für gefühlsbetonte Motivation. Ein wichtiges Merkmal des Interesses ist der intrinsische Charakter, d.h. die Auseinandersetzung mit den Übungsaufgaben sollte einer Eigeninitiative folgen und nicht von außen getrieben sein.<sup>11</sup> Bei allen Items wird deshalb ein klarer Gegenstands- und Situationsbezug (Aufgaben und Übung) deutlich, um eine Verbindung positiver Valenzen zu anderen Sachverhalten während der Messung auszuschließen. Zwar wurden einige Items der originären IE-Komponente des MAI-Fragebogens beibehalten (z.B. M02, M19), doch ist a priori unklar, ob sich diese Items im Gefüge der ergänzten IE-Items psychometrisch zu einem gemeinsamen Faktor gruppieren.

### 5.4.2 Akademisches Selbstkonzept (SK)

Unter dem akademischen Selbstkonzept wird allgemein die Vorstellung über die eigenen disziplinbezogenen kognitiven Fähigkeiten verstanden. Diese wahrgenommene Kompetenz wird ebenso wie das Interesse als positiver Prädiktor intrinsisch motivierter Verhaltensweisen erachtet (Bandura et al., 2001; Deci & Ryan, 1993; Hattie, 2009). Im originären MAI-Fragebogen sowie in der hier adaptierten Version bezieht sich das wahrgenommene Selbstkonzept konkret auf die spezielle Lernsituation des Untersuchungsgegenstandes (Physikunterricht bzw. Übungen zur Mechanik), die als Komponenten eines hierarchisch übergeordneten mathematisch-naturwissenschaftlichen Selbstkonzepts aufgefasst werden können. Das Selbstkonzept ist in der Lehr-/Lernforschung vom besonderen Interesse, denn es nimmt eine wichtige Rolle im Entscheidungsprozess des beruflichen Werdegangs ein (Filipp, 2006), wirkt sich auf die Leistungsbereitschaft aus und bestimmt das Selbstwertgefühl (Stiensmeier-Pelster & Schöne, 2008) — laut Bandura et al. (2001) sogar in größerem Maße als die eigentliche Leistungsfähigkeit selbst. Geschlechtsspezifische Unterschiede in der Ausprägung des Selbstkonzepts werden darüber hinaus als Erklärung für Studien- und Berufswahl herangezogen, was gerade in der Physikdidaktik großes Forschungsinteresse anstößt (vgl. Klein et al., 2015b). In diesem Zusammenhang führten Häußler und Hoffmann (1995) im Rahmen der Kieler Interessenstudie (über 8000 Teilnehmer) Interessensunterschiede zwischen Schülerinnen und Schülern fast vollständig auf das Selbstkonzept zurück.

Die hier relevante Komponente des akademischen Selbstkonzepts betrifft die Einschätzung, die Übungsaufgaben lösen zu können bzw. verstanden zu haben (vgl. Items M01, M26, M36, M43 im Testbogen auf S. 202), das Zutrauen in die eigenen Fähigkeiten (M11, M30) sowie die Beurteilung, wie die eigene Leistung von anderen wahrgenommen wird (M50).

### 5.4.3 Realitätsbezug / Authentizität der Aufgaben (RA)

Die RA-Items bringen die Alltagsbezogenheit der Übungsaufgaben und die daraus resultierende Relevanz des Lerngegenstandes zum Ausdruck (vgl. Tab. A-1). Authentische Problemstel-

<sup>7</sup>z.B. „Mir gefällt unser Physikunterricht“

<sup>8</sup>z.B. „In meiner Freizeit beschäftige ich mich auch über die Hausaufgaben hinaus mit Themen, die mit Physik zu tun haben“

<sup>9</sup>M35: „Auch wenn es anstrengend ist, Übungsaufgaben zu lösen, so bereitet mir die Beschäftigung damit doch Vergnügen“

<sup>10</sup>M10: „Die Inhalte der Übungsaufgaben sind für mich von persönlicher Bedeutung“

<sup>11</sup>z.B. M52: „Ohne äußeren Druck würde ich mich wohl nicht so regelmäßig mit den Inhalten und Problemen meines Studiums beschäftigen“

lungen, die einen Bezug zur realen Welt herstellen, vermeiden es nach Kuhn (2010), dass „das [ansonsten] in einem isolierten Kontext erworbene Wissen punktuell und unvernetzt bleibt“ [S. 17]. Darüber hinaus wird durch die Alltagsrelevanz des Inhalts ein Anwendungsbezug hergestellt, was im Sinne eines kontextorientierten Physikunterrichts als motivationsfördernd erachtet wird (Duit & Mikelskis-Seifert, 2007; vgl. „faktische Authentizität“, Barab et al., 2000). Das große Motivationspotential des Realitätsbezugs für Schülerinnen und Schüler kann auch auf Physikstudierende übertragen werden (Henning et al., 2012), wobei weitere Aspekte der Authentizität berücksichtigt werden sollten, siehe nächster Abschnitt.

#### 5.4.4 Disziplinäre Authentizität (dA)

Für Physikstudierende spielt neben der Authentizität des Inhalts auch die Identifikationsintensität mit dem Fach Physik eine große Rolle, die durch die wahrgenommene persönliche Bedeutsamkeit der Inhalte bestimmt wird (Watkins et al., 2012). Damit ist gemeint, dass die Möglichkeit zu spezifischen Arbeitsweisen, die den Vorstellungen der Studierenden über die Disziplin Physik entsprechen, zu höherer Motivation führt. Werden durch die Aufgaben vermehrt *fachliche Methoden* vermittelt, wird der Professionsbezug und damit die persönliche Relevanz gestärkt. Zur Erhebung der disziplinären Authentizität schätzen die Studierenden beispielsweise ein, ob ihnen die Aufgaben die Möglichkeit zum physikalischen Arbeiten geben (M05), ob sie Strategien kennenlernen, die ihnen physikalisches Problemlösen ermöglicht (M25) oder ob sie ihre fachmethodischen Fähigkeiten durch die Übungen erweitern können (M34). Als Vorlage für die Itemkonstruktion dienten überwiegend Beschreibungen und Aussagen über den Charakter der disziplinbezogenen Arbeitsweisen aus den Arbeiten von Enghag und Niedderer (2007), Rainer und Matthews (2002) sowie Watkins et al. (2012); vgl. Bemerkung zu Tab. A-1.

Eine Subkomponente der disziplinären Authentizität kann darin verstanden werden, eigene Experimente durchzuführen (M39, M49), eigene Fragestellung zu formulieren (M17, M29) und eigene Daten zu erheben (M44). Durch die Betonung der Eigenständigkeit im Lernprozess („*Ownership of learning*“) wird mit dieser Komponente ein besonders hoher Grad der Authentizität angesprochen, die sich nach Osborne (2003) bedeutsam auf die intrinsische Motivation auswirkt und deshalb im Fragebogen verortet wird. Die Lernenden übernehmen eine aktive Rolle im Lernprozess, wodurch die persönliche Relevanz (Wertbezug) und die Autonomie erhöht werden (Enghag & Niedderer, 2007). Der im Motivationsfragebogen abgebildete Authentizitätsbegriff reicht von der Authentizität der Lernmaterialien selbst (RA) über die Authentizität der verwendeten fachspezifischen Methoden und Arbeitsweisen (dA) bis hin zur Entwicklung eigener Fragestellungen und eigener Produkte aus der Lernsituation (dA-Ownership). In Kapitel 6.4.3 wird untersucht, ob sich diese unterschiedlichen Facetten der Authentizität auch psychometrisch voneinander abtrennen.

#### 5.4.5 Autonomie (Atnm)

Ein wesentliches Element der Motivation besteht nach Ryan und Deci (2000) in der Selbstständigkeit und Selbstbestimmung beim Lernen, was eng an eine Entscheidungsfreiheit bzgl. des Lernprozesses geknüpft ist. Die strikten Rahmenbedingungen der verpflichtenden Übungsaufgaben erlauben es kaum, den Studenten Freiheiten in Bezug auf die Themeninhalte oder die Art und Weise der Aufgabenbearbeitung zu gewähren. Trotz dieser untergeordneten Rolle der instruktionsbedingten Autonomie in diesem Untersuchungsszenario wurden vier Items formuliert, die nach der Strenge der Aufgabenführung fragen (M08, M16, M32, M47) und eng an etablierte Items des SDS (*Self-Determination Scale*) angelehnt sind (Sheldon & Deci, o. J.). Insbesondere die mobilen Videoanalyse-Aufgaben besitzen durch die Interdependenz von theoretischen und experimentbezogenen Teilaufgaben einen höheren Offenheitsgrad als traditionelle Aufgaben, wodurch der Problemlöseprozess stärker von Studierenden beeinflusst werden kann.

## 5.5 Neugierdefragebogen

Neugierde wird als Reiz aufgefasst, Neuartiges erfahren zu wollen, und spielt eine entscheidende Rolle beim Lernen. Auf Grundlage ihrer Meta-Analyse über die prädiktive Kraft verschiedener Variablen zur Erklärung akademischer Leistungsfähigkeit fanden von Stumm et al. (2011) heraus, dass die Neugierde — ein wissbegieriger Geist („*hungry mind*“) — neben der Intelligenz und der Gewissenhaftigkeit ein ganz wesentliches Differenzierungsmerkmal bzgl. der Leistungsfähigkeit darstellt. Auch wenn das Instruktionsmaterial nicht explizit darauf abgestimmt ist, Neugierde zu wecken, wird die Wirkung des Instruktionsmaterials auf diese Variable untersucht und der Einfluss der Neugierde auf den Lernzuwachs analysiert. In der Entwicklungsphase der Instrumente konnten auf existierende Neugierde-Fragebögen zurückgegriffen werden, aus denen bedeutsame Items selektiert und auf den Untersuchungsgegenstand adaptiert wurden. Dies erfolgte in enger Zusammenarbeit mit dem Kollegen Michael Hirth, in dessen Studie die Neugierde eine übergeordnete Untersuchungsvariable bildet (Hirth, 2016). Deshalb beschränkt sich die Darstellung der Fragebogengengese auf die Angabe der Items mit den zugehörigen Quellen (siehe Anhang, Tab. A-2) sowie auf eine Einordnung des Begriffsverständnisses der Neugierde in bestehende Theorien.

Berlyne (1978) trifft eine Unterscheidung zwischen einer *perzeptuellen* Neugierde-Komponente und einer *epistemischen*. Die perzeptuelle Komponente steht in Verbindung mit Reizen (kognitiv wie sensorisch), die ein Gefühl der Ungewissheit oder eine Lust an Sensationen erwecken (*sensation seeking*). Die epistemische Komponente der Neugierde beschreibt den Drang nach Wissen. Ausgehend von dieser Abgrenzung wurde das Konzept der Neugierde in den letzten Jahrzehnten stets weiter verfeinert. So unterscheidet Naylor (1981) eine der Person innewohnende, zeitlich weitestgehend stabile Ausprägung der Neugierde (*curiosity trait*) von der Wissbegierde, die aus einem bestimmten Zustand resultiert und über Situationen hinweg variiert (*curiosity state*). Der aus dem Neugierde-Reiz resultierende Drang nach Erkundung wird schließlich noch bezüglich der Dimension des Fokus, nämlich *diversiv* (ungerichtet) und *spezifisch* (auf einen bestimmten Reiz gerichtet), unterschieden (Litman & Spielberger, 2003).

In der vorliegenden Studie wurde ein Fragebogen mit zwei Subdimensionen konzipiert (siehe Tab. A-2). Da hauptsächlich die epistemische Komponente der Neugierde für die Lehr-/Lernforschung von Bedeutung ist, wurden die Items mit engem Bezug auf diese Komponente ausgewählt. Die erste Subdimension dient der Erfassung der *Neugierde als Personeneigenschaft* (NP) und beinhaltet neun Items mit allgemeinen Aussagen über die epistemische Neugierde.<sup>12</sup> Diese Subdimension erfasst die oben angesprochene stabile Neugierde-Komponente in diversiver Ausrichtung und dient der Untersuchung hauptsächlich als Kontrollvariable. Um einen Einfluss des Instruktionsmaterials auf die *Neugierde als aktuellen Zustand* (NZ) messbar zu machen, fragen die acht Items der zweiten Subdimension nach der spezifisch durch die Übungsaufgaben geweckten epistemischen Neugierde.<sup>13</sup>

## 5.6 Belastungsfragebogen

Um Aussagen über die Belastung der Studierenden während der Intervention zu erhalten, wurde ein Fragebogen mit zwei Komponenten entwickelt. Die erste Komponente beinhaltet zehn Items, die primär Aussagen über den *Cognitive Load* der Aufgaben (CL) enthalten (Tab. A-3). Die Items entstammen größtenteils dem Fragebogen zu Ankereigenschaften von Aufgaben aus der Arbeit von Kuhn (2010), die auf die Bearbeitungssituation in den Übungen angepasst und durch eigene Items ergänzt wurden. Im hier aufgefassten Sinne umfasst der Cognitive Load die Schwierigkeit der Aufgabeninhalte<sup>14</sup> sowie die durch die Aufgabenstellung und das Aufgabenformat generierte notwendige Anstrengung zur Aufgabenbearbeitung.<sup>15</sup>

<sup>12</sup>z.B. NP5: „Ich freue mich, etwas über Themen zu lernen, die mir unbekannt sind“

<sup>13</sup>z.B. NZ8: „Die Übungsaufgaben haben mich neugierig gemacht, mehr über Mechanik zu erfahren“

<sup>14</sup>z.B. CL01: „Die Übungsaufgaben waren schwierig“

<sup>15</sup>z.B. CL05: „Ich wusste während der Bearbeitung der Übungsaufgaben genau, was ich zu tun hatte“ oder CL07 „Es war schwierig, wichtige von unwichtigen Informationen in der Aufgabenstellung zu unterscheiden“

Die zweite Komponente fragt nach der *psychischen Beanspruchung* durch das Studium allgemein<sup>16</sup> und der Physik-Übung im Speziellen.<sup>17</sup> Die Items entstammen ausnahmslos dem *Perceived Stress Questionnaire* (PSQ), der von Levenstein et al. (1993) zur Diagnose von Belastung entwickelt und von Fliege et al. (2001) an einer deutschen Stichprobe faktoranalytisch untersucht und validiert wurde. Beide Subdimensionen dienen dazu, quantitative Aussagen über die Belastung und deren Veränderung zu erhalten und insbesondere um Unterschiede in der aufgabeninduzierten Belastung zwischen den Untersuchungsgruppen festzustellen.

## 5.7 Betreuerfragebogen

Mit Blick auf eine sorgfältige methodische Kontrolle der Vergleichsstudie wurden fünf Items zur Erhebung des Betreuerverhaltens konstruiert. Wie im Studiendesign beschrieben wird (vgl. Kap. 10.2), betreute ein Übungsgruppenleiter jeweils eine Kontroll- und eine Interventionsgruppe. Die Items zum Betreuerverhalten fragen danach, wie motivierend der Übungsgruppenleiter auf die Studierenden wirkt (vgl. Tab. A-4). Die Erhebung dieser Variable dient dem Zweck, unterschiedliches Verhalten der Betreuer zwischen den Gruppen zu überprüfen, um eine Beeinflussung (beispielsweise durch ideologische Überzeugungen) im Idealfall auszuschließen.

---

<sup>16</sup>z.B. PSQ1: „Ich habe für mein Studium zu viele Dinge zu erledigen“

<sup>17</sup>z.B. PSQ3: „Ich fühle mich durch die Abgabetermine für die Übungsaufgaben unter Druck gesetzt“



## Kapitel 6

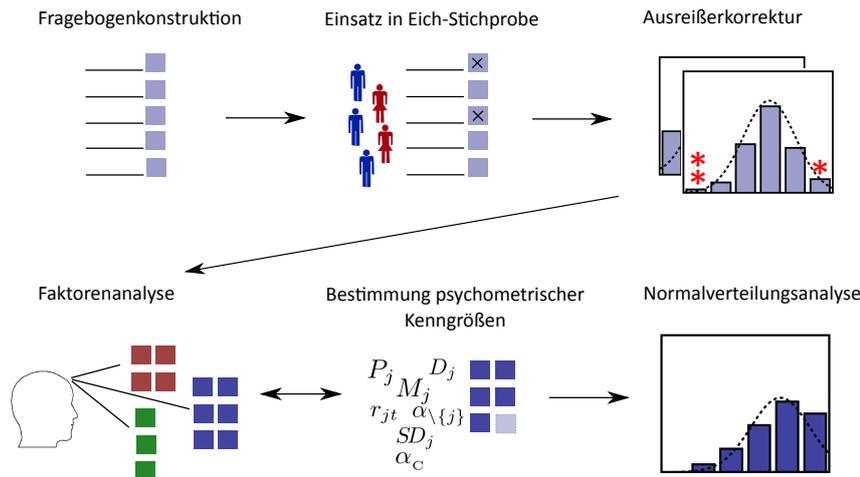
# Testanalysen nach Klassischer Testtheorie (KTT)

Das übergeordnete Ziel der Testanalysen besteht in der Bildung inhaltlich valider und psychometrisch reliabler Skalen. Unter einer Skala wird eine Batterie mehrerer Items zur Messung eines bestimmten latenten Konstrukts verstanden. Die in Kap. 5 konstruierten Fragebögen stellen die Grundlage für die Skalen dar. Unter Verwendung psychometrischer Verfahren wird ermittelt, welche Items der Fragebögen in eine Skala aufgenommen werden. In diesem Kapitel wird also eine konzeptionelle Unterscheidung zwischen dem Fragebogen (als materiellem Gegenstand) mit allen konstruierten Items und der Skala, die eine ausgewählte Teilgruppe von Items beinhaltet, getroffen. Fragebögen erfassen mehrere (latente) Merkmale, während Skalen näherungsweise eindimensional sind. Die Bildung psychometrisch guter Skalen ist für diese Arbeit deshalb bedeutungsvoll, weil aus ihnen die Ausprägungen der Untersuchungsvariablen berechnet werden (Testsummen / Mittelwerte). Da diese Testsummen und Skalenmittelwerte die Grundlage der Vergleichsanalysen bilden, müssen die Skalen hinreichend messscharf sein, sodass die aus den Analysen gezogenen Schlussfolgerungen belastbar sind. In diesem Kapitel werden zunächst die zur Skalenbildung nötigen methodischen Grundlagen der klassischen Testtheorie zusammengefasst. Es folgen die Planung und Strukturierung der Testanalysen auf Basis einer Eich-Stichprobe sowie der Ergebnisbericht. Das Resultat dieses Kapitels ist die Zusammenstellung bestmöglicher Skalen nach inhaltlichen und psychometrischen Kriterien sowie die Quantifizierung derer Güte.

### 6.1 Methodische Grundlagen

#### 6.1.1 Prinzip der Skalenbildung

Nachdem die Fragebögen in einer Erhebungsstichprobe eingesetzt wurden, werden die erhobenen Daten aller Items einer Voranalyse unterzogen. Diese beinhaltet die Bestimmung skalenunabhängiger Item-Eigenschaften (Mittelwerte, Standardabweichungen) und die Prüfung auf Extremwerte und Ausreißer. Die ausreißerkorrigierten Messwerte werden anschließend einer Faktorenanalyse unterzogen, um die Dimensionalität der eingesetzten Tests zu beurteilen. Über die korrelativen Beziehungen der Items untereinander wird so eine Gruppierung zu sogenannten Faktoren hergestellt. Damit wird einerseits überprüft, ob die nach theoretischen Annahmen zusammengestellten Itemgruppen innerhalb der untersuchten Stichprobe auch tatsächlich als statistisch bedeutsame Faktoren erhalten bleiben und sich voneinander abgrenzen — zum Beispiel die beiden Neugierde-Komponenten. Andererseits identifiziert die Faktorenanalyse auffällige Items, die das zu messende Konstrukt nicht hinreichend abbilden. Die Faktorenanalyse dient damit als erster Validierungsschritt der verwendeten Instrumente. Da das Verfahren in Statistik-Lehrbüchern ausreichend beschrieben wird, beschränkt sich



**Abb. 14:** Vorgehensweise bei Skalenbildung (schematisch).

dessen Darstellung in Kapitel 6.1.2 auf die grundlegende Prinzipien und Anwendungsspezifikationen für die hier erhobenen Variablen.

In einem nächsten Schritt werden item- und skalenbezogene Kenngrößen ermittelt, um die Güte einzelner Items und des Gesamttests zu bestimmen und mit den Ergebnissen anderer Studien vergleichbar zu machen. Dazu zählen die Konsistenz der Instrumente (Reliabilität) sowie die Eigenschaft, fähige oder motivierte Studenten von weniger fähigen oder unmotivierten zu unterscheiden (Diskrimination). Diese Kenngrößen stellen nach der Faktorenanalyse eine zweite Entscheidungshilfe dar, Items aus den Skalen zwecks Optimierung der Messgenauigkeit auszuschließen. In Kapitel 6.1.3 werden die erwähnten psychometrischen Maße definiert, sowie deren Bedeutung für die Testentwicklung näher erläutert.

In einem letzten Schritt werden die gefundenen Skalen einer Normalverteilungsanalyse unterzogen, um bei späteren Vergleichsanalysen die Wahl zwischen parametrischen und nicht-parametrischen Testverfahren auf eine statistisch begründete Entscheidung zu stützen (Kap. 6.1.4). Das Prinzip der Skalenbildung ist in Abb. 14 schematisch dargestellt und wird für festgelegte Itemgruppen (siehe Kap. 6.2) wiederholt angewendet.

### 6.1.2 Faktorenanalyse

Zur Untersuchung der faktoriellen Struktur der eingesetzten Instrumente wird die explorative Faktorenanalyse genutzt (Backhaus et al., 2011; Bortz & Schuster, 2010; Tabachnick & Fidell, 2001). Das Verfahren analysiert Items auf ihre korrelativen Beziehungen mit dem Ziel, diese in wenige voneinander unabhängige Gruppen (Faktoren) zu ordnen. Als Extraktionsverfahren wird die Hauptkomponentenanalyse verwendet, die ausgehend von der Korrelationsmatrix in einem ersten Schritt paarweise orthogonale Faktoren ermittelt, die sukzessiv maximale Varianz in den Variablen (Items) erklären, d.h. absteigende Eigenwerte besitzen. Das Ergebnis sind Indexzahlen, die — vereinfacht ausgedrückt — angeben, wie gut ein Item durch den jeweiligen Faktor beschrieben wird („Faktorladungen“). Für die Anzahl zu extrahierender Faktoren existieren unterschiedliche Kriterien. Nach Kaiser werden alle Faktoren mit Eigenwert kleiner 1 verworfen, da diese weniger Varianz als die einer einzigen Variable erklären und somit unbedeutend sind. Da dieses Kriterium bei vielen Variablen allerdings zu einer großen Anzahl schwierig zu interpretierenden Faktoren führt, empfehlen Bortz und Döring (2006) zusätzlich die Betrachtung des *ScreepLOTS*. Im Screeplot sind alle Faktoren geordnet nach ihrem Eigenwert dargestellt. Weist die Eigenwertkurve einen „Knick“ auf, fällt also der Absolutwert der Eigenwerte nach einem bestimmten Faktor rapide ab, werden nur die Faktoren vor dem Knick betrachtet, da alle weiteren Faktoren weniger bedeutsam sind. Neben Kaiser- und

Scree-Kriterium spielen inhaltliche Überlegungen zur Bestimmung der Faktorenzahl eine zentrale Rolle, um die Lösung mit größtem wissenschaftlichem Nutzen zu erhalten (Bortz & Döring, 2006).

In einem zweiten Schritt wird die gewonnene Faktorlösung zwecks besserer Interpretierbarkeit rotiert. Das hier verwendete Varimax-Rotationsverfahren mit anschließender Kaiser-Normalisierung maximiert die Ladung eines Items auf genau einem Faktor und minimiert die Ladungen dieses Items auf den anderen Faktoren. In der Gesamtheit ergibt sich bestenfalls eine sogenannte Einfachstruktur, die sich durch hohe Ladungen ( $> 0.5$ ) der Items auf jeweils einem Faktor auszeichnet. Wichtige Kenngrößen der Faktorlösung wie die Item-Kommunalitäten (Ausmaß der Varianzaufklärung eines Items durch die Faktoren) und die erklärte Gesamtvarianz sind unter dieser Rotation invariant. Als Qualitätskriterium der gefundenen Faktorlösung wird die Stabilität  $FS$ , gegeben durch

$$FS = 1 - \left( 1.10 \cdot \frac{1}{\sqrt{N}} - 0.12 \cdot l_{\min} + 0.066 \right), \quad (2)$$

berechnet (Bortz & Schuster, 2010, S. 396). Dabei bezeichnet  $l_{\min}$  den minimalen Ladungswert, der bei der Interpretation der Faktoren berücksichtigt wird. Ab einem Richtwert von  $FS = 0.90$  kann davon ausgegangen werden, dass die gefundene Lösung replizierbar ist und stichprobenunabhängig interpretiert werden kann.

Das Anwenden der Faktorenanalyse setzt voraus, dass die Items ausreichend hoch miteinander korrelieren. Dies wird durch komplementäre Untersuchungsverfahren der zugrunde liegenden Korrelationsmatrix geprüft (KMO-Maß, Bartlett-Test auf Sphärizität), die an dieser Stelle als bekannt vorausgesetzt werden (Bortz & Schuster, 2010). Neben Erfüllung dieser formalen Kriterien ist es außerdem ratsam, vor der Analyse Ausreißerwerte auf Itemebene zu überprüfen (Tabachnick & Fidell, 2001). Der Grund liegt darin, dass die Verwendung von Produkt-Moment-Korrelationen (als Ausgangsbasis der Faktorenanalyse) streng genommen normalverteilte Zufallsvariablen voraussetzt.<sup>1</sup> Dies wurde im strukturellen Vorgehen berücksichtigt, vgl. Abb. 14.

### 6.1.3 Psychometrische Kenngrößen der KTT

Zur Beurteilung der Güte von Multiple-Choice-Tests existieren zahlreiche Kenngrößen, die routinemäßig berechnet und mit vorgegebenen Zielintervallen verglichen werden können (Moosbrugger & Augustin, 2007). Itembezogene Kenngrößen geben beispielsweise Auskunft über die Streuung der Daten innerhalb der Stichprobe, die Unterscheidungsfähigkeit zwischen Personen oder über den Zusammenhang eines Items mit der gesamten Skala. Dazu zählen das arithmetische Mittel eines Items (bei kognitiven Items auch Schwierigkeitsindex genannt), die Standardabweichung, der Diskriminationsindex und der Trennschärfeparameter. Skalenbezogene Kennwerte wie die Reliabilität ( $\alpha_c$ ) geben eine Auskunft über die interne Konsistenz der Itemgruppen. Diese Größen werden im Folgenden erläutert und sind in Tabelle 5 gemeinsam mit ihrem möglichen Werte- und Zielbereich zusammengefasst.

1. Der Schwierigkeitsindex  $P_j$  eines Items  $j$  wird über die Formel

$$P_j = \frac{R_j}{N_j} \quad (3)$$

berechnet und liegt zwischen 0 und 1. Dabei sind  $R_j$  die Anzahl richtiger Antworten und  $N_j$  die Gesamtanzahl gegebener Antworten (d.h. richtige + falsche + fehlende Antworten). Entgegen der Bezeichnung dieser Kenngröße ist das Item  $j$  umso schwerer, je kleiner der Wert  $P_j$  ausfällt. Im üblichen Sprachgebrauch wird  $P_j$  auch als Lösungswahrscheinlichkeit des Items  $j$  interpretiert.

<sup>1</sup>Im Falle dichotomer Daten (z.B. bei Leistungstests) ist es üblich, statt Produkt-Moment-Korrelationen sog. tetrachorische Korrelationen zu nutzen (Scott, Schumayer & Gray, 2012), da infolge der starken Diskretisierung des Wertebereichs die Normalverteilungsannahme häufig verletzt wird.

2. Bei Beurteilungs-Items ist der Mittelwert  $M_j$  die zu  $P_j$  korrespondierende Größe. Der Mittelwert  $M_j$  ist das arithmetische Mittel aller Beurteilungen des Items  $j$  und wird als durchschnittliches Zustimmungsmaß interpretiert. Da die Primärdaten von Beurteilungs-Items in Werten zwischen 1 und 6 vorliegen, erfolgt für eine bessere Interpretierbarkeit der Kenngrößen eine lineare Transformation (Umskalierung) der Primärdaten auf das Intervall  $[0, 1]$ .
3. Die Standardabweichung  $SD_j$  als Maß der Differenzierungsfähigkeit ergibt sich bei kognitiven (dichotomen) Items direkt über  $SD_j^2 = P_j(1-P_j)$ , weshalb sie nur für Beurteilungs-Items berichtet wird. In diesem Zusammenhang sei angemerkt, dass die Standardabweichung der Skalenwerte  $SD$  ungleich dem Mittelwert der Item-Standardabweichungen ist. Die Skalenwerte  $SD$  werden immer berichtet, unabhängig vom Itemtyp.
4. Der Diskriminationsindex  $D_j$  misst, wie stark ein Testitem zwischen Personen mit hohen Testsummen und niedrigen Testsummen unterscheidet. Dazu wird die Differenz zwischen der Anzahl richtiger Antworten im oberen Quantil der Testsummen ( $N_{O,j}$ ) und der Anzahl richtiger Antworten im unteren Quantil der Testsummen ( $N_{U,j}$ ) gebildet und durch die Quantilbreite  $N_{Q,j}$  (typischerweise  $N_{Q,j} = N_j/4$ ) geteilt, also

$$D_j = \frac{N_{O,j} - N_{U,j}}{N_{Q,j}} = P_{O,j} - P_{U,j}. \quad (4)$$

Für Rating-Skalen gilt Gl. (4) entsprechend mit den Bezeichnungen  $M_{O,j}$  bzw.  $M_{U,j}$ . Die beiden Größen  $P_j$  und  $D_j$  sollten stets zusammen betrachtet werden, da die möglichen Beobachtungswerte von  $D_j$  durch die  $P_j$ -Werte eingeschränkt werden — Items mit zu hohen oder zu niedrigen Schwierigkeitsindizes können per Definition nicht diskriminieren. Der von Ding und Beichner (2009) postulierte absolute Schwellenwert

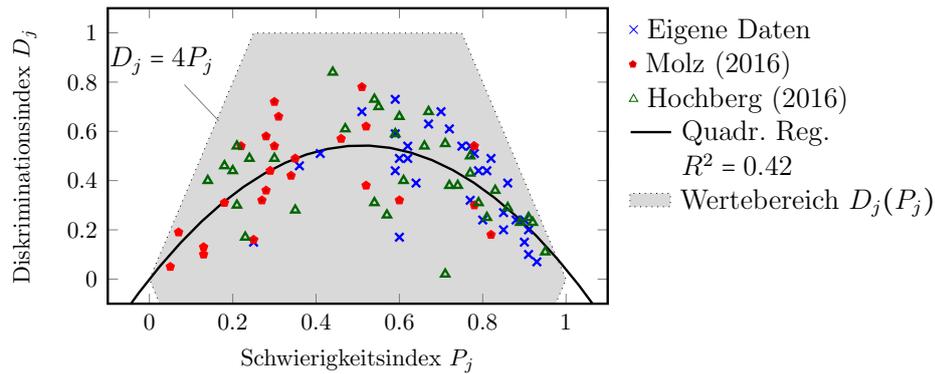
**Tab. 5:** Psychometrische Kenngrößen der KTT

Kenngröße	Wertebereich	Zielbereich <sup>a</sup>
<u>Kenngrößen auf Itemebene</u>		
Schwierigkeitsindex $P_j$	$[0, 1]$	$[0.35, 0.85]$
Mittelwert $M_j$	$[0, 1]$	-
Standardabweichung* $SD_j$	$[0, 0.5]$	-
Diskriminationsindex $D_j$	$[-1, 1]$	$\geq 0.3$
Trennschärfeparameter $r_{jt}$	$[-1, 1]$	$\geq 0.2$
Reliabilität bei Itemausschluss $\alpha_{\setminus\{j\}}$	$(-\infty, 1]$	$(\leq \alpha_c)$
<u>Kenngrößen auf Skalenebene</u>		
Mittlerer Schwierigkeitsindex $P$	$[0, 1]$	$[0.35, 0.85]$
Skalenmittelwert $M$	$[0, 1]$	-
Standardabweichung $SD$	$[0, 0.5]$	-
Mittlerer Diskriminationsindex $D$	$[-1, 1]$	$\geq 0.3$
Reliabilität $\alpha_c$	$(-\infty, 1]$	$\geq 0.7$

*Bemerkungen:* Kenngrößen auf Itemebene werden durch den Index  $j$  gekennzeichnet. Schwierigkeitsindizes beziehen sich auf kognitive Items / Variablen und werden auch Lösungswahrscheinlichkeit genannt; Mittelwerte beziehen sich auf Beurteilungsskalen.

<sup>a</sup>Empfehlung gemäß Ding und Beichner (2009) mit geringfügiger Anpassung beim Schwierigkeitsindex wegen formatbedingten Rate-Effekten (ursprünglich  $[0.3, 0.9]$ ).

\*wird auf Itemebene nur bei Beurteilungs-Items angegeben (Erläuterungen im Text).



**Abb. 15:** Empirische  $(P_j, D_j)$ -Wertepaare aggregiert aus eigenen Test-Daten und aus Studien der gemeinsamen Arbeitsgruppe. Der theoretisch mögliche Wertebereich von  $D_j$  ergibt sich unmittelbar aus Gl. (3) und (4).

von  $D_j > 0.3$  kann also für sehr schwere oder sehr leichte Items per Definition nicht eingehalten werden. Dieser Sachverhalt wird in Abb. 15 veranschaulicht und durch empirische  $P_j$ - $D_j$ -Wertepaare unterstützt, die in verschiedenen Studien parallel zu dieser Arbeit entstanden. Es kann studienübergreifend festgestellt werden, dass Items mit extremalen Schwierigkeitsindizes ( $P_j < 0.2$  bzw.  $P_j > 0.8$ ) systematisch niedrigere Diskriminationsindizes besitzen als mittelschwere Items. Dies bedeutet, dass vor allem mittelschwere Items mit niedrigem Diskriminationsindex kritisch geprüft werden müssen. Mögliche Ursachen für zu niedrige Diskrimination liegen in mehrdeutigen Formulierungen, Fehlvorstellungen der Studierenden, fehlerbehafteter Auswertemethodik oder Mehrdimensionalität der Skala.

- Der Trennschärfekoeffizient  $r_{jt}$  (auch Item-Test-Korrelation) gibt den korrelativen Zusammenhang zwischen dem Item  $j$  und der Gesamtttestsumme an. Bei dichotomen Items heißt  $r_{jt}$  auch punktbiserialer Korrelationskoeffizient und es gilt die Formel

$$r_{jt} = \frac{T_j \text{ richtig} - T_j \text{ falsch}}{SD} \sqrt{P_j(1 - P_j)}, \quad (5)$$

wobei  $T_j$  richtig/falsch die durchschnittliche Testsumme der Personen ist, die das Item  $j$  richtig bzw. falsch gelöst haben. Zur Trennschärfenbestimmung bei Skalen mit wenigen Items empfiehlt es sich, das jeweilige Item aus der Summenbildung auszuschließen, da  $r_{jt}$  sonst überschätzt wird (korrigierte Item-Test-Korrelation, vgl. Moosbrugger & Augustin, 2007, S. 83).

- Die Reliabilität  $\alpha_c$  (nach Cronbach) gilt als Maß für die interne Konsistenz einer Skala. Es gilt die Formel

$$\alpha_c = \frac{n}{n-1} \left( 1 - \frac{\sum_{j=1}^n SD_j^2}{SD^2} \right), \quad (6)$$

wobei  $n$  die Anzahl der Items bezeichnet. Für den Spezialfall dichotomer Items heißt Gl. (6) auch Kuder-Richardson-Formel („KR20“) mit  $SD_j^2 = P_j(1 - P_j)$ . Es bietet sich an, die Veränderung der Reliabilität zu beobachten, wenn einzelne Items zur Skala hinzugefügt oder entfernt werden ( $\alpha_{\setminus\{j\}}$ ). Für den Itemausschluss sollten inhaltliche Gründe Vorrang vor statistischen Empfehlungen haben. Durch eine Eliminierung zu vieler Items können unter Umständen zwar exzellent homogene Skalen geschaffen werden, jedoch sinkt auch der wissenschaftliche Nutzen der Skala. In diesem Zusammenhang verweist Streiner (2003) darauf, dass ab einem Wert von  $\alpha_c = 0.90$  kein Mehrwert

in einer weiteren Steigerung der Reliabilität besteht. Außerdem ist zu beachten, dass die Reliabilität  $\alpha_C$  von der Testlänge abhängt und kein Maß für die Dimensionalität eines Tests ist (Green et al., 1977). Tavakol und Dennick (2011) argumentieren, dass der Reliabilitätsanalyse eine Dimensionsanalyse vorangehen sollte und raten davon ab, Reliabilitäten für gemischte Skalen anzugeben (z.B. eines mehrdimensionalen Gesamttests). Dieser Ratschlag wurde in dem hier gewählten Vorgehen berücksichtigt (vgl. Abb. 14).

#### 6.1.4 Prüfung auf Normalverteilung und Ausreißerkorrektur

Die Prüfung auf Normalverteilung der Skalenausprägungen in der untersuchten Stichprobe entscheidet darüber, ob für die Vergleichsanalysen der Hauptstudie parametrische Verfahren genutzt werden können. Neben einem visuellen Eindruck der Verteilungen mithilfe von Histogrammen umfasst die Normalverteilungsanalyse *drei* Schritte. Zunächst prüft der Normalverteilungstest von Shapiro-Wilk (SW-Test) die Nullhypothese, dass die beobachteten Daten einer Normalverteilung folgen. Unterschreitet der  $p$ -Wert des SW-Tests das kritische Signifikanzniveau von  $p = 0.05$ , liegt eine signifikante Abweichung der Häufigkeitsverteilung von der Normalverteilung vor. Bortz und Döring (2006) sowie Ahad et al. (2011) weisen in diesem Zusammenhang darauf hin, dass der SW-Test — wie andere Normalverteilungstests auch — mit wachsender Stichprobe selbst bei geringen Differenzen der empirischen Daten zur Normalverteilung signifikant wird. Aus diesem Sensitivitätsargument wird bei signifikantem Ergebnis der Empfehlung von Mayers (2013, S. 52–53) gefolgt, Schiefe und Kurtosis der Skalenmittelwerte zu betrachten. Die Schiefe steht im Zusammenhang mit der Asymmetrie der Verteilungsfunktion und die Kurtosis bewertet die Steilheit (Gipfligkeit, Wölbung). Miles und Shevlin (2001, S. 74) geben für die Absolutwerte von Schiefe und Kurtosis die Faustregel an, dass Werte unter 1 tolerierbar sind. Eine statistische Absicherung wird über das Verhältnis des Absolutwertes zu seinem Standardfehler ( $z$ -Wert) erreicht, der stichprobenabhängige Grenzen betragsmäßig nicht überschreiten darf. Sind die bisher aufgeführten Kriterien verletzt, wird in einem letzten Schritt versucht, eine Normalverteilung durch das Korrigieren von Ausreißern näherungsweise herzustellen. Werte gelten als Ausreißer, wenn der  $|z|$ -Wert der standardisierten Daten größer als 3.29 ist (Mayers, 2013, S. 53). Bei Verteilungsanalysen einzelner Items bietet sich eine Einschränkung auf den letzten Schritt an (Ausreißeranalyse), denn bei nur wenigen Ausprägungskategorien stellen parametrische Normalverteilungstests in diesem Fall keine sinnvolle Analyseemethode dar.

## 6.2 Definition der Eich-Stichprobe und Strukturierung der Testanalysen

Für die Testanalyse werden die Daten aus den beiden Hauptstudien aggregiert, um eine möglichst große Datengrundlage zu erhalten. Items, die sich zwischen beiden Teilstudien der Hauptuntersuchung stark verändert haben, neu hinzugefügt oder entfernt wurden, werden nicht in der Testanalyse berücksichtigt, sondern separat untersucht, um eine maximale Stichprobengröße zu gewährleisten. Zum Prä-Zeitpunkt liegen  $N = 165$  ausgefüllte Fragebögen vor, die zur Bildung der Skalen genutzt werden (*Eich-Stichprobe*; eine Ausnahme bilden die Neugierde und das Betreuerverhalten, siehe unten). Unter Berücksichtigung der engen inhaltlichen Verwandtschaft der erhobenen Variablen werden die Items

- (1) zur Aufgabenschwierigkeit und zur Beanspruchung,
- (2) zur Neugierde als Persönlichkeitsmerkmal und zur Neugierde als aktueller Zustand,
- (3) zum Realitätsbezug, Selbstkonzept, Interesse, zur disziplinierten Authentizität und Autonomie sowie

## (4) zur Antwortsicherheit bei KiRC- und Konzept-Items

jeweils gemeinsam analysiert, vgl. Tab. 6. Die Tabelle stellt auch die Erhebungszeitpunkte dar, aus denen sich unter Berücksichtigung weiterer Variablenmerkmale folgende Spezifikationen ergeben:

- Der NP-Fragebogen stand erst im zweiten Teil der Hauptuntersuchung zur Verfügung (WS 2014/15) und der NZ-Fragebogen wurde jeweils erst zum ersten Post-Zeitpunkt eingesetzt. Deshalb wird die Faktorenanalyse für die Neugierde-Items anhand der Post1-Daten der zweiten Hauptstudie durchgeführt ( $N = 43$ ). Die gefundene Dimensionsstruktur bildet die Basis für die anschließende Itemanalyse, die auf Grundlage der jeweils größeren Stichprobe durchgeführt wird.
- Zur Analyse des Repräsentationstests (KiRC) wird vor der Skalenanalyse eine Kern-Itemgruppe definiert, die die Schnittmenge von unveränderten Items über beide Hauptstudien enthält.
- Wegen der geringen Itemanzahl des Konzepttests wird auf eine Faktorenanalyse verzichtet. Wie in Kap. 5.2 begründet, wird der Konzepttest eigenständig und unabhängig von KiRC behandelt.
- Die beiden Leistungstests werden auf *Änderungssensitivität* geprüft. Da zwischen den beiden Testzeitpunkten  $t_0$  und  $t_1$  vier Wochen liegen, in denen Aufgabenstellungen zum Lerninhalt bearbeitet werden, ist davon auszugehen, dass sich die Leistung in der Gesamtkohorte über die Zeit verbessert. Leistungstests sollten solche Änderungen in den Merkmalausprägungen auch detektieren können. Zur weiteren Validierung werden die berechneten Itemstatistiken studienübergreifend verglichen.
- Die Analyse der Items zum Betreuerverhalten beruht auf den Beurteilungen zum ersten Post-Zeitpunkt ( $N = 96$ ).

In Anlehnung an das strukturelle Vorgehen der Skalenbildung aus Abb. 14 werden zunächst Ausreißerkorrekturen auf Itemebene berichtet (Kap. 6.3), bevor sich die Ergebnisse der Faktoren- und Itemanalysen anschließen (Kap. 6.4) und die Verteilungen der Skalenwerte überprüft werden (Kap. 6.5). Die Ausreißer-, Normalverteilungs- und Faktorenanalysen wurden mit SPSS durchgeführt, tetrachorische Korrelationen wurden mit R berechnet, und die psychometrischen Kenngrößen hauptsächlich mit Microsoft EXCEL.

**Tab. 6:** Zeitpunkte der eingesetzten Fragebögen in beiden Hauptstudien

Fragebogen	HS1			HS2		
	$t_0$	$t_1$	$t_2$	$t_0$	$t_1$	$t_2$
Beanspruchung (1)	●	○	○	●	○	○
Aufgabenschwierigkeit (1)	●	○	○	●	○	○
Neugierde als Personeneigenschaft (2)	-	-	-	○	●	○
Neugierde als akt. Zustand (2)	-	○	○	-	●	○
Motivation (3)	●	○	○	●	○	○
Repräsentationskompetenz	●	○	-	●	○	-
Konzeptverständnis Trajektorien	●	○	○	●	○	○
Antwortsicherheiten Repräsentationskompetenz (4)	●	○	-	●	○	-
Antwortsicherheiten Konzeptverständnis (4)	●	○	○	●	○	○
Beurteilung des Übungsleiters	-	●	○	-	●	○

*Bemerkungen:* HS = Hauptstudie, Testzeitpunkte  $t_0$  = Prä,  $t_1$  = Erster Post-Test,  $t_2$  = Zweiter Post-Test (siehe auch Kap. 10.2). Striche (-) markieren Zeitpunkte ohne Fragebogeneinsatz; volle Markierungen (●) kennzeichnen Zeitpunkte zur Dimensionsanalyse. Fragebögen mit gleicher Kennziffer werden gemeinsam analysiert.

### 6.3 Ergebnisse und Diskussion I: Ausreißeranalyse aller Items

Zur Voranalyse wurden alle Beurteilungs-Items auf Ausreißer geprüft, ehe die Faktorenanalysen durchgeführt wurden. Tabelle 7 gibt einen Überblick über die Häufigkeit der Beobachtungen, deren  $z$ -Werte den kritischen Wert von 3.29 betragsmäßig überschreiten. Im Falle eines Ausreißers wurden die zugehörigen Rohwerte durch den Itemmittelwert  $M_j \pm 3SD_j$  ersetzt (plus / minus je nach Richtung des Ausreißers). Im Falle einer Korrektur beträgt die Differenz zwischen dem korrigierten Wert und der Beobachtung typischerweise 5 bis 15 Prozentpunkte.

Der Anteil an Ausreißern liegt bei jedem Fragebogen unter einem Prozent und ist damit tolerierbar. Es ist deshalb nicht zu erwarten, dass die o.g. Ausreißerkorrektur einen wesentlichen Einfluss auf die Ergebnisse der Skalenbildung nimmt. Im digitalen Anhang ist die Suche nach Ausreißern und deren Korrektur dokumentiert.

**Tab. 7:** Ergebnisse der Ausreißeranalyse auf Itemebene

Fragebogen	Beobachtungen <sup>a</sup>	Ausreißer ( $ z  > 3.29$ )	Ausreißer in %
Beanspruchung	988	2	.20
Aufgabenschwierigkeit	1639	0	-
Neugierde P	746	2	.27
Neugierde Z	775	0	-
Motivation	7833	14	.18
Antwortsicherheiten	2714	2	.07
Beurteilung Betreuer	480	1	.21
Gesamt	15175	21	.14

<sup>a</sup>Summe der Beobachtungen aller Items innerhalb eines Fragebogens.

### 6.4 Ergebnisse und Diskussion II: Faktorenanalysen und psychometrische Kenngrößen

Vor Durchführung der Faktorenanalysen erfolgte eine Prüfung auf Eignung des Datensatzes (vgl. Kap. 6.1.2). Die Ergebnisse sind in Tab. 8 zusammengefasst und zeigen anhand der KMO-Maße und den Ergebnissen der Bartlett-Tests, dass die Voraussetzungen zur Durchführung einer Hauptkomponentenanalyse erfüllt sind. Die berichteten Stabilitätswerte ( $FS > 0.89$ ) sprechen dafür, dass die faktorenanalytischen Ergebnisse, die in den folgenden Kapiteln berichtet werden, stichprobenunabhängig interpretierbar sind.

**Tab. 8:** Prüfung der Datensätze auf Eignung zur Hauptkomponentenanalyse

Fragebogen	Gesamtvarianz <sup>a</sup>	KMO	Bartlett-Test			FS
			$\chi^2$	df	p	
Belastung	60 <sub>ex</sub>	.91	1481	120	<.001	.92
Neugierde	45 <sub>ex</sub>	.67	312	136	<.001	.92
Motivation	39 <sub>in</sub>	.77	3220	1128	<.001	.90
KiRC-Test	57 <sub>ex</sub>	.80	730	231	<.001	(-) <sup>b</sup>
Antwortsicherheiten	57 <sub>ex</sub>	.88	920	120	<.001	.89
Betreuer	59 <sub>ex</sub>	.75	172	10	<.001	.90

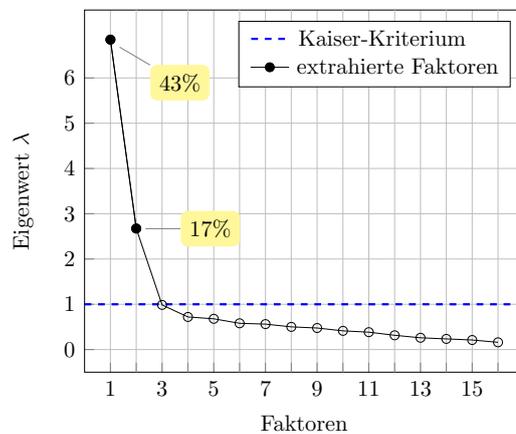
<sup>a</sup>Erklärte Gesamtvarianz (in %) der extrahierten (ex) bzw. der interpretierten Faktoren (int). <sup>b</sup>Faktorlösung wurde nicht interpretiert.

### 6.4.1 Belastungsfragebogen

#### Faktorenanalyse

Die Hauptkomponentenanalyse führte unter Verwendung des Kaiser-Kriteriums (Eigenwert  $\lambda > 1$ ) zu zwei Faktoren, die gemeinsam 60% der Varianz erklären, siehe Abb. 16. Die in Tabelle 9 dargestellte rotierte Faktorlösung ist eindeutig interpretierbar: Die CL-Items laden einheitlich auf dem ersten Faktor (Aufgabenschwierigkeit) und zeigen nur geringe Nebenladungen auf dem zweiten Faktor, auf dem alle PSQ-Items eindeutig laden (Beanspruchung). Es liegt eine Einfachstruktur vor. Trotz der inhaltlichen Verwandtschaft beider Belastungskomponenten liegen also psychometrisch klar trennbare Faktoren vor, die im Folgenden getrennt analysiert werden.

**Abb. 16:** Screeplot: Aus der Korrelationsmatrix berechnete Eigenwerte absteigend sortiert nach Größe. Zwei Eigenwerte erfüllen das Kaiser-Kriterium und tragen gemeinsam 60% zum Erklärungsgehalt des Modells bei.



**Tab. 9:** Rotierte Faktorlösung (Belastung)

Item	Faktoren		$\kappa$
	1	2	
CL4	.89	.13	.81
CL9	.86	.11	.75
CL1	.83	.17	.71
CL3	.81	.01	.65
CL6	.80	.08	.64
CL5	.78	.04	.61
CL10	.77	.24	.66
CL2	.77	.18	.62
CL8	.71	.10	.52
CL7	.58	.16	.36
PSQ4	.03	.80	.63
PSQ1	.14	.78	.63
PSQ6	.11	.72	.54
PSQ2	-.02	.69	.47
PSQ5	.27	.66	.51
PSQ3	.18	.62	.42

*Bemerkung:* Faktorladungen mit Betrag  $< 0.30$  sind kontrastarm dargestellt.

#### Item- und Skalenanalyse

Tabelle 10 zeigt die Statistiken aller CL- und PSQ-Items, bezogen auf ihre jeweilige Skala. Die Mittelwerte der CL-Items liegen im unteren bis mittleren Bereich ( $M_j = 0.32 - 0.54$ ), d.h. der Schwierigkeitsgrad der Aufgaben wird zum Prä-Testzeitpunkt (nach Bearbeitung des ersten Übungsblattes) als einfach bis moderat wahrgenommen. Die Mittelwerte der PSQ-Items liegen ebenfalls im mittleren Bereich ( $M_j = 0.45 - 0.67$ ), wobei Item PSQ1 („Ich habe in meinem Studium viele Dinge zu erledigen“) den höchsten Wert aufweist, was in Einklang mit der Erfahrung aus den ersten Semesterwochen steht. Jedes Item zeigt sowohl ausreichend hohe Diskriminationsindizes ( $D_j > 0.30$ ) als auch substantielle Trennschärfen ( $r_{jt} > 0.20$ ). Einzig der Ausschluss des Items CL7 („Es war schwierig, wichtige von unwichtigen Informationen in der Aufgabenstellung zu unterscheiden“) würde die Reliabilität der Schwierigkeitsskala weiter erhöhen, was aber angesichts der ohnehin sehr hohen Reliabilität ( $\alpha_c = 0.93$ ) nicht notwendig ist. Dieses Item zeigt im Vergleich die niedrigste Faktorladung, kleinste Kommunalität und geringste Trennschärfe. Auf dieses Item kann aus inhaltlichen Gründen nicht verzichtet werden, da damit eine wichtige Information des extraneous CL gemessen wird: Da sich die notwendigen Angaben zur Bearbeitung von V-Aufgaben im Gegensatz zu T-Aufgaben nicht nur im Aufgabentext, sondern teilweise auch im Screenshot zum Versuchsaufbau befinden, kann das Item wichtige Hinweise auf unterschiedliche Anforderungen geben.

Tabelle 10 fasst weiterhin die Kennwerte beider Skalen zusammen. Die Reliabilitäten sind gut bis exzellent ( $\alpha_c = 0.81$  bzw.  $0.93$ ), und der mittlere Diskriminationsindex  $D$  ist akzeptabel. Die vergleichsweise etwas geringere Reliabilität der Beanspruchungsskala ist auf die Skalenkürze (nur 6 Items) zurückzuführen.

**Tab. 10:** Ergebnisse der Itemanalyse (Belastung)

Item	Kurzformulierung	$M_j$	$SD_j$	$D_j$	$r_{jt}$	$\alpha_{\setminus\{j\}}$
<u>Schwierigkeits-Items</u>						
CL1	Aufgaben sind schwierig	.52	.22	.45	.79	.92
CL2	Bearbeitung verläuft holprig	.44	.22	.41	.73	.93
CL3	Inhalt schwer verständlich	.47	.25	.51	.74	.93
CL4	Lösungsschwierigkeiten	.45	.23	.53	.86	.92
CL5	Ahnungslosigkeit beim Bearbeiten	.41	.19	.38	.71	.93
CL6	Herausfordernde Aufgaben	.50	.24	.46	.74	.93
CL7	Wesentliches schwer identifizierbar	.32	.23	.34	.54	.94
CL8	Informationen schwer auffindbar	.37	.23	.41	.66	.93
CL9	Anstrengung notwendig	.54	.25	.55	.82	.92
CL10	Verständnisschwierigkeiten	.38	.23	.47	.76	.92
<u>Beanspruchungs-Items</u>						
PSQ1	Viele Dinge erledigen müssen	.67	.20	.39	.67	.77
PSQ2	Zu wenig Zeit für sich selbst	.57	.23	.42	.51	.80
PSQ3	Gefühl unter Druck zu sein	.59	.23	.38	.50	.80
PSQ4	Rastlosigkeit spüren	.48	.25	.49	.65	.77
PSQ5	Es wird zu viel verlangt	.45	.22	.40	.56	.79
PSQ6	Sich in Eile fühlen	.55	.23	.42	.58	.78
<u>Skalenkennwerte</u>		$M$	$SD$	$D$	$\alpha_c$	
	Aufgabenschwierigkeit	.44	.18	.45	.93	
	Beanspruchung	.55	.15	.42	.81	

*Bemerkung:* Große Mittelwerte bedeuten hohe Belastungen — Aussagen, die im Originalwortlaut (Tab. A-3) entgegen der Schlüsselrichtung formuliert waren, sind hier entsprechend umformuliert.

## 6.4.2 Neugierdefragebogen

### Faktorenanalyse

Nach dem Scree-Kriterium wurde die Anzahl extrahierter Faktoren auf zwei festgesetzt (vgl. Abb. A-1 im Anhang). Die rotierte Faktorlösung bestätigt die Erwartung, dass die NP-Items (Neugierde als Personeneigenschaft) auf einem Faktor und die NZ-Items (Neugierde als aktueller Zustand) überwiegend auf dem anderen gruppieren (siehe Tab. 11). Die Items NP6, NZ1 und NZ5 fallen durch niedrige Primär- bzw. hohe Nebenladungen auf. Item NP6 („In meiner Freizeit gehe ich spannenden Aktivitäten nach“) zielt eher auf eine Erlebniskomponente der Neugierde ab als auf eine epistemische. Item NZ1 („Durch die Übungsaufgaben wurde meine Phantasie angeregt“) ist vermutlich zu allgemein formuliert, sodass neben dem aktuellen Zustand auch die zeitlich robuste Neugierde-Komponente angesprochen wird und die negative Ladung von Item NZ5 („Ich möchte die behandelten Inhalte der Übungsaufgaben gerne weiter vertiefen“) deutet darauf hin, dass das Vertiefen von Lerninhalten primär nicht mit allgemeiner Neugierde korreliert.

### Item- und Skalenanalyse

Die Ergebnisse der Item- und Skalenanalyse befinden sich ebenfalls in Tabelle 11. Es zeigt sich, dass die Mittelwerte der NP-Items mit Ausnahme von NP6 im oberen Bereich liegen ( $M_j = 0.74 - 0.85$ ), d.h. die Stichprobe ist allgemein sehr neugierig. Dementsprechend fallen die Diskriminationsindizes der Items niedrig aus, weshalb diese nicht zur Itemselektion

verwendet werden (vgl. Kap. 6.1.3). Ferner wird deutlich, dass die Reliabilität der NP-Skala durch Ausschluss von NP6 von 0.77 auf 0.80 verbessert werden kann. Aus diesem Grund und wegen oben diskutierter inhaltlicher Abweichung dieses Items wurde es aus der Skala entfernt. Eine Neuberechnung der testsummenbezogenen Itemkennwerte zeigt geringfügige Verbesserungen der Diskriminationsindizes und Trennschärfeparameter. Die Items NZ1 und NZ5 werden trotz auffälliger Nebenladungen im Instrument behalten, da die Itemkennwerte tolerierbar und der Inhalt der Items (Vertiefung von Lerninhalten) bedeutungsvoll ist.

**Tab. 11:** Ergebnisse der Faktoren- und Itemanalyse (Neugierde)

Item	Faktorenanalyse			Itemanalyse ( $\alpha = 0.77$ )					Itemanalyse <sup>a</sup>		
	1	2	$\kappa$	$M_j$	$SD_j$	$D_j$	$r_{jt}$	$\alpha_{\setminus\{j\}}$	$D_j$	$r_{jt}$	$\alpha_{\setminus\{j\}}$
NP4	.81	-.15	.68	.82	.21	.36	.51	.75	.35	.51	.77
NP1	.76	-.04	.58	.85	.14	<b>.22</b>	.52	.75	<b>.23</b>	.54	.77
NP9	.71	.13	.52	.76	.15	<b>.17</b>	.38	.76	<b>.20</b>	.44	.78
NP2	.70	.29	.59	.77	.15	<b>.23</b>	.51	.75	<b>.25</b>	.53	.77
NP7	.68	.25	.52	.74	.22	.41	.52	.74	.42	.49	.78
NP3	.63	-.03	.39	.78	.17	<b>.26</b>	.46	.75	<b>.26</b>	.53	.77
NP5	.62	.02	.39	.81	.14	<b>.25</b>	.56	.74	<b>.25</b>	.54	.77
NP8	.54	.29	.38	.76	.16	.32	.56	.74	.32	.53	.77
NP6	.32	-.15	.12	.62	.25	.33	.28	.80	-	-	-
NZ3	-.05	.76	.58	.61	.19	.40	.73	.81			
NZ4	.13	.72	.53	.64	.21	.39	.70	.81			
NZ2	.30	.71	.59	.60	.18	.38	.71	.82			
NZ8	-.18	.66	.47	.58	.21	<b>.29</b>	.45	.85			
NZ1	.39	.61	.52	.56	.21	.44	.62	.82			
NZ6	.17	.59	.38	.61	.20	.39	.66	.82			
NZ5	-.37	.45	.34	.64	.19	<b>.27</b>	.43	.85			
NZ7	-.06	.41	.17	.55	.22	.32	.40	.85			
Skalenkennwerte (nach Re-Analyse)				$M$	$SD$	$D$	$\alpha_c$				
Neugierde als Personeneigenschaft				.79	.11	<b>.28</b>	.80				
Neugierde als aktueller Zustand				.60	.14	.36	.85				

*Bemerkungen:* Itemformulierungen befinden sich im Anhang (Tab. A-2); Faktorladungen mit Betrag  $< 0.30$  sind aus Übersichtsgründen kontrastarm dargestellt und Itemkennwerte außerhalb konventioneller Grenzen (siehe Kap. 6.1.3) sind fettgedruckt.

<sup>a</sup>Re-Analyse der NP-Items nach Ausschluss von NP6 ( $M_j$  bzw.  $SD_j$  bleiben unverändert).

### 6.4.3 Motivationsfragebogen

#### Faktorenanalyse

Die Hauptkomponentenanalyse ergibt nach dem Kaiser-Kriterium 14 Faktoren mit Eigenwerten  $\lambda > 1$ , die zusammen 68% der Varianz erklären. Die hohe Anzahl extrahierter Faktoren ist auf die hohe Itemzahl (bei vergleichsweise geringer Stichprobengröße) zurückzuführen und ist nach anschließender Varimax-Rotation nicht vollständig interpretierbar. Der Screeplot legt eine Drei-Faktoren-Lösung nahe (Abknicken des Kurvenverlaufs zwischen dem dritten und vierten Eigenwert, vgl. Abb. A-1), die erfolglos zu interpretieren versucht wurde. Da weder das Kaiser- noch das Scree-Kriterium zu einer interpretierbaren Lösung führen, wird der Vorschlag von Bortz und Schuster (2010) aufgegriffen, die Faktorenlösung mit größtem wissenschaftlichem Nutzen zu verwenden. In diesem Fall bedeutet dies konkret, die rotierte Lösung mit 14 Faktoren zu verwenden, daraus interpretierbare Faktoren als Skalen anzunehmen (*Faktorlösung*) und die Skalen unter Beachtung der psychometrischen Qualität inhaltlich zu ergänzen (*inhaltlich ergänzte Lösung*).

Die Faktorladungsmatrix ist aufgrund ihrer Größe nicht vollständig abgebildet. Tabelle 12 zeigt den interpretierten Teil der Matrix, der 29 Items mit ihren Faktorladungen auf den ersten 6 Faktoren beinhaltet. Die restlichen 19 Items verteilen sich nicht eindeutig auf weitere

acht Faktoren, die im Folgenden nicht weiter berücksichtigt werden. Der erste Faktor mit dem größten Eigenwert beinhaltet sieben RA-Items und weist sich somit als RA-Subskala aus. Faktor 2 enthält in erster Linie Items, die die gefühlsbezogene Komponente des Interessenkonstrukts betreffen. Das Bearbeiten der Übungsaufgaben wird mit positiven Gefühlen (Freude, Vergnügen, Spaß) verbunden. Item M20 zielt auf die persönliche Bedeutsamkeit der Aufgabenbearbeitung und spricht somit die wertbezogene Komponente des Interesses an. Faktor 3 beinhaltet fünf SK-Items. Die niedrige Faktorladung von Item M30 ist dadurch erklärbar, dass neben der Kompetenz zur fachlichen Diskussion auch äußere Faktoren wie besondere Lernarrangements oder soziographische Gegebenheiten eine Rolle spielen. Nichtsdestotrotz ist dieses Item im weiteren Sinne kompetenzthematisch und trägt zur Identifikation dieses Faktors als Selbstkonzeptskala bei. Auf Faktor 4 laden fünf Items zur Erfassung der disziplinären Authentizität, während sich die experimentbezogenen Items auf Faktor 5 gruppieren. Faktor 6 wird von drei Autonomie-Items gebildet (kurz Atnm-Skala). Die sechs interpretierten Faktoren erklären eine Varianz von 39%.

### Inhaltliche Ergänzung der Skalen

Das Ergebnis der Faktorenanalyse zeigt, dass die theoretisch konstruierten Skalen auch psychometrisch eigenständige Faktoren bilden, die allerdings nicht alle ursprünglich konstruierten Items enthalten (Tab. 12). Der statistisch bedeutsamste Faktor beinhaltet den

**Tab. 12:** Ausschnitt der rotierten Faktorlösung (Motivation)

Item	Kurzformulierung	TZ <sup>a</sup>	Faktoren						$\kappa$
			1	2	3	4	5	6	
M18	Alltagsbezogene Aufgaben	RA	.75	.12	.00	.23	-.04	-.05	.75
M48	Physik ist alltagsrelevant	RA	.74	.04	.00	.11	.09	.09	.69
M07	Themen bilden Realität ab	RA	.72	.15	.05	.01	.03	.08	.69
M38	Bezug Theo. und Realität	RA	.63	.14	.15	.06	.20	.11	.69
M46	Alltägliche Dinge	RA	.60	-.04	.03	-.02	.21	.18	.69
M33	Bezug zum phys. Alltag	RA	.59	.04	.02	-.16	.21	.24	.77
M31	Nützlich im Alltag	RA	.57	-.04	.13	.09	.08	.12	.70
M35	Problemlösen Vergnügen	IE	.09	.78	.15	.06	.09	.07	.71
M41	Aufgaben machen Spaß	IE	.11	.74	.11	-.01	.00	.07	.72
M12	Bearbeite gerne Aufgaben	IE	.06	.59	.09	.11	-.04	.16	.63
M20	Hilfreich für Verständnis	IE	.03	.56	.07	.35	.14	.05	.56
M28	Merke nicht wie Zeit vergeht	IE	.19	.43	.03	.00	.13	-.01	.53
M36	Eigene Leistungen gut	SK	.09	.16	.84	-.03	.01	-.03	.79
M50	Andere denken ich sei gut	SK	-.04	.01	.77	.09	.10	.07	.66
M26	Löse erfolgreich	SK	.04	.16	.67	.06	.07	.04	.74
M11	In Zukunft gute Leistung	SK	.14	-.01	.47	-.01	-.10	-.05	.64
M30	Eigene Lösung diskutieren	SK	.11	.20	.44	.16	.16	.17	.52
M05	Arbeit wie Physiker	dA	.19	-.18	.06	.74	.08	-.06	.69
M04	Lerne wichtige Methoden	dA	.01	.28	.10	.73	.03	-.06	.73
M25	Lerne Lösungsstrategien	dA	.01	.33	.14	.56	.26	.12	.75
M03	Phys. Denkweisen lernen	dA	.15	.31	-.16	.51	-.12	.11	.67
M09	Ermöglichen wiss. Arbeiten	dA	.39	.09	.16	.40	.14	.38	.69
M39	Nutze Experimente	dA	.27	.03	-.05	.03	.81	.05	.78
M49	Erstelle eigene Exp.	dA	.10	.06	.11	.07	.78	.05	.70
M44	Erhebe eigene Daten	dA	.01	.05	.18	.10	.55	.20	.57
M29	Formuliere Hypothesen	dA	-.03	.12	.23	.08	.47	.06	.67
M08	Lege Arbeitsweise fest	Atnm	.29	-.03	.11	.00	.16	.73	.69
M47	Kann Einfluss nehmen	Atnm	.01	.07	.04	-.03	.16	.72	.75
M32	Aufg. erlauben Freiheiten	Atnm	.13	.32	-.03	.05	-.11	.65	.68
Summe der Items									19.9
Summe nicht dargestellter Items									12.9
Eigenwerte			4.0	3.2	3.1	3.1	2.6	2.5	

*Bemerkungen:* Ausformulierte Items befinden sich im Anhang (Tab. A-1). Faktorladungen mit Betrag  $< 0.30$  sind aus Übersichtsgründen kontrastarm dargestellt.

<sup>a</sup>Theoriebasierte Zuordnung zu den Motivationsskalen gemäß Kap. 5.4 (siehe auch Tab. A-1).

Realitätsbezug der Aufgaben, was auf die hohen Interkorrelationen der RA-Items infolge der parallelen Formulierungen zurückzuführen ist. Das Item M15 (Physik im Alltag verstehen) lädt in der empirisch gefundenen Lösung nicht auf diese Skala, wird aber aus inhaltlichen Gründen ergänzt, da die Vermittlung von Verständnis alltagsbezogener Inhalte vor dem Hintergrund der Authentizität wesentlich ist (vgl. Kap. 3.1.2 und 5.4). Wie in Tabelle 13 dargestellt ist, verändert sich die Skalenreliabilität durch das Hinzufügen des Items nicht. Es zeigt sich weiter, dass sich die gefühlsbezogene Komponente des Interessenkonstrukts fast eigenständig gruppiert (Faktor 2) während sich die wertbezogenen Items mit kleinen Ladungen auf der kompletten Faktorlösung verteilen (nicht gezeigt) und keinen eigenen Faktor bilden. Eine mögliche Erklärung ist darin zu vermuten, dass zwischen einer persönlichen Relevanz der Übungsaufgaben und den anderen Faktoren zum Zeitpunkt der Erhebung noch nicht unterschieden wird. Das Wertgefühl der Übungsaufgaben für die Erweiterung der eigenen Kompetenz konfundiert vermutlich mit dem Ziel, die Leistungsanforderungen der Veranstaltung zu erfüllen. Um die Anzahl von Faktoren überschaubar zu halten, werden zwei der wertbezogenen Interesse-Items zur empirisch gefundenen Skala hinzugenommen (M10 - persönliche Relevanz, M14 - Lernen für sich selbst), wodurch sich die Skalenreliabilität sogar verbessern lässt (vgl. Tabelle 13). Ferner weist die Faktorenanalyse auf eine Unterscheidung zwischen einer disziplinären Authentizität hinsichtlich des reinen Professionsbezugs (Arbeitsweise der Physik) und einer spezifischen Komponente, eigene Experimente durchzuführen, Daten zu erheben etc. hin (*Experimentbezug*, kurz: Exp-Skala). Obwohl beide Faktoren disziplinbezogene Methoden betreffen, werden sie theoriekonform nach dem Ausmaß an empfundener Authentizität unterschieden (Kap. 5.4.4). Der vierte Faktor *disziplinäre Authentizität* (kurz dA-Skala) wird im Folgenden von der Exp-Skala abgegrenzt. Die Items M22 (fachspezifische Relevanz erkennen) und M34 (phys. Fähigkeiten erweitern) werden zur dA-Skala und Item M17 (eigene Fragen entwickeln) wird zur Exp-Skala hinzugefügt, wodurch sich die Reliabilitäten in beiden Fällen steigern lassen.

Items, die weder in der Faktorlösung vorkommen noch inhaltlich ergänzt wurden, zeigen entweder eine hohe Redundanz zu Neugierde-Items, weshalb auf sie verzichtet werden kann (z.B. M19 und M37), fragen Inhalte ab, die zum Prä-Zeitpunkt noch nicht valide eingeschätzt werden können (z.B. M1, M6, M27 und M43) oder wurden nicht sorgfältig genug konstruiert (z.B. M2, M21, M23). Beispielsweise ist Item M2 („Ich schaue zu Hause in Büchern, im Internet oder ähnlichem nach, um mehr zu den Übungsaufgaben zu erfahren“) schlichtweg ungeschickt formuliert, da die Übungsaufgaben per se im Selbststudium — d.h. zu Hause — bearbeitet werden, und somit nicht zu erwarten ist, dass das Item eine interessensspezifische Komponente anspricht.

### Item- und Skalenanalyse

Auf Darstellung der Ergebnisse auf Itemebene wird verzichtet. Analog zu dem Vorgehen vorangegangener Analysen wurden die Itemkennwerte bestimmt und unter Berücksichtigung konventioneller Grenzbereiche bewertet. Es stellte sich heraus, dass keines der 35 Items Auffälligkeiten bzgl. der Kennwerte zeigte, weshalb die oben diskutierte Skalenbildung nicht weiter modifiziert wurde. Die Ergebnisse auf Skalenebene sind in Tabelle 13 für die Faktorlösung sowie für die inhaltlich ergänzten Skalen dargestellt. Wie zu sehen ist, zeigen die inhaltlich optimierten Skalen gleichwertige oder bessere psychometrische Kennwerte als die originäre Faktorlösung.

**Tab. 13:** Skalenkennwerte der Motivationskomponenten

	Art der Skalenbildung <sup>a</sup>	# Items	<i>M</i>	<i>SD</i>	<i>D</i>	$\alpha_c$
<i>Realitätsbezug</i>	Faktorlösung	7	.64	.13	.32	.83
	inkl. M15	8	.65	.12	.32	.83
<i>Interesse</i>	Faktorlösung	5	.72	.14	.35	.74
	inkl. M10 u. M14	7	.72	.12	.31	.76
<i>Selbstkonzept</i>	Faktorlösung	5	.59	.14	.33	.74
<i>Disz. Authentizität</i>	Faktorlösung	5	.68	.12	.30	.72
	inkl. M22 u. M34	7	.71	.11	.27	.77
<i>Experimentbezug</i>	Faktorlösung	4	.42	.16	.41	.71
	inkl. M17	5	.45	.15	.38	.72
<i>Autonomie</i>	Faktorlösung	3	.69	.15	.37	.71

<sup>a</sup>Unterscheidung zwischen rein psychometrischer Lösung (Faktorlösung) und inhaltlich ergänzter Lösung (durch Hinzufügen der angegebenen Items).

#### 6.4.4 KiRC-Instrument

##### Bildung einer Kern-Itemgruppe

Aus allen KiRC-Items wurde zunächst eine *Kern-Itemgruppe* identifiziert, die ausschließlich Items beinhaltet, die in beiden Hauptstudien eingesetzt und dazwischen nicht wesentlich verändert wurden. Zu diesem Zweck sind in der linken Hälfte von Tabelle 14 die Lösungswahrscheinlichkeiten  $P_j(t_0)$  aller Items aufgelistet, die in beiden Teilstudien eingesetzt wurden (vgl. Tab. 4). Hier und im Folgenden werden einzelne TF-Items einer MTF-Gruppe als eigenständig betrachtet (vgl. Kap. 5.1.3). Die *p*-Werte geben die Wahrscheinlichkeit dafür an, dass der Unterschied zwischen den gemessenen Lösungswahrscheinlichkeiten eines Items in beiden Teilstudien rein zufällig zustande kommt. Die Ergebnisse basieren auf zweiseitigen *t*-Tests, die nach einer Korrektur des  $\alpha$ -Fehlers nach Benjamini-Hochberg vier auffällige Items identifizieren.<sup>2</sup> Die verschiedenen Lösungswahrscheinlichkeiten dieser Items zwischen beiden Teilstudien können auf inhaltliche Änderungen im Zuge der Instrumentenüberarbeitung zurückgeführt werden und sind im digitalen Anhang dokumentiert. Zum Beispiel wurden bei Item K1 andere Antwortalternativen bei gleichem Itemstamm angeboten, was zu einer bedeutsamen Erniedrigung des Schwierigkeitsindex führte (von  $P_j = 0.70$  auf  $P_j = 0.19$ ). Dies bedeutet, dass die Aufgabe schwieriger wurde.<sup>3</sup>

Um aus den verbleibenden 32 Kern-Items eine Skala zu bilden, wurden zunächst die tetrachorischen Korrelationen zwischen den Variablen bestimmt und die resultierende Matrix als Grundlage für eine Faktorenanalyse verwendet (vgl. Faktorenanalyse bei dichotomen Daten, Fußnote 1 auf S. 81). Es stellte sich heraus, dass die berechnete Matrix singulär ist und sich somit nicht für das Verfahren eignet. Der Grund hierfür liegt in zu hohen Inter-Item-Korrelationen zwischen Items mit Deckeneffekten. Aus diesem Grund wird die Testanalyse vorgezogen, um die Itemanzahl zu reduzieren und die Faktorenanalyse mit einem optimierten Datensatz durchzuführen.

<sup>2</sup>Bemerkung zu dem Problem des multiplen Testens: Je mehr Tests mit demselben Datensatz durchgeführt werden, desto höher wird die Wahrscheinlichkeit, dass ein Ergebnis fehlerhaft signifikant wird ( $\alpha$ -Fehler-Inflation). Diesem Fehler wird durch eine Anpassung des  $\alpha$ -Niveaus begegnet, wofür mehrere Verfahren existieren (Victor et al., 2010): Bei der konservativsten Methode (Bonferroni-Korrektur) wird das globale  $\alpha$ -Niveau (typischerweise 0.05) durch die Anzahl der Tests geteilt. Die Benjamini-Hochberg-Korrektur ist weniger konservativ und vergleicht die *p*-Werte mit einer wachsenden  $\alpha$ -Schranke. Bei *k* Tests gilt für den kleinsten *p*-Wert die Bonferroni-Schranke  $\alpha/k$ , für den zweitkleinsten  $\alpha/(k-1)$  usw. (Benjamini & Hochberg, 1995). Es werden alle Nullhypothesen abgelehnt, die zu *p*-Werten gehören, die kleiner als der größte *p*-Wert sind, der eine zugehörige Schranke unterschreitet. Damit stellt die Methode einen Kompromiss zwischen der strengen Bonferroni-Korrektur und keiner Korrektur des  $\alpha$ -Niveaus dar.

<sup>3</sup>Es sei an dieser Stelle erneut auf die kontra-intuitive Bezeichnung dieses Index hingewiesen.

**Item- und Skalenanalyse**

Die Analyseergebnisse der Kern-Itemgruppe ist in der rechten Hälfte von Tabelle 14 dargestellt. Durch den Ausschluss von 10 Items mit auffälligen Werten wurde die sogenannte *optimierte Kern-Itemgruppe* mit 22 Items gebildet.

Mit Ausnahme von Item K17b und K22b zeigen die ausgeschlossenen Items ausnahmslos zu hohe Lösungswahrscheinlichkeiten, sind also zu einfach. Sie besitzen entsprechend niedrige Diskriminationswerte und Trennschärfe-Parameter. Item 17b weist trotz moderater Lösungswahrscheinlichkeit ( $P_j = 0.60$ ) einen sehr kleinen Diskriminationsindex ( $D_j = 0.17$ ) auf und ist nicht trennscharf ( $r_{jt} = 0.01$ ), d.h. das Item misst eine andere Kompetenz als die restlichen. Inhaltlich fragt das Item nach der Stetigkeit der Geschwindigkeit bei einem gegebenen  $x(t)$ -Diagramm. Das Item erforderte als einziges Testitem ein Begriffsverständnis für Stetigkeit, wodurch die geringe Testkonformität erklärt wird. Erwähnenswert ist die sehr

**Tab. 14:** Festlegung der Kern-Itemgruppen und Itemanalysen (KiRC)

	Prüfung auf Kern-Item				Itemanalyse						
	$P_j(t_0)$	$P_j(t_0)$	$p$ -	Kern- Item	$P_j$	Kern-Items			opt. Kern-Items		
	HS1	HS2	Wert			$D_j$	$r_{jt}$	$\alpha_{\setminus\{j\}}$	$D_j$	$r_{jt}$	$\alpha_{\setminus\{j\}}$
K1	.70	.19	.000*	○	-	-	-	-	-	-	-
K2	.73	.70	.642	●	.72	.61	.52	.80	.63	.50	.80
K3	.74	.80	.437	●	.77	.32	.24	.81	.32	.24	.81
K4	.78	.81	.673	●	.79	.44	.44	.80	.49	.44	.81
K5	.56	.70	.067	●	.62	.49	.35	.80	.46	.34	.81
K6	.70	.65	.545	●	.67	.63	.45	.80	.68	.47	.80
K7	.87	.93	.194	●	<b>.90</b>	<b>.24</b>	<b>.33</b>	<b>.81</b>	-	-	-
K8	.74	.52	.002*	○	-	-	-	-	-	-	-
K9	.77	.84	.225	●	.81	.44	.41	.80	.44	.41	.81
K10	.77	.87	.100	●	.82	.49	.48	.80	.49	.47	.80
K11	.32	.41	.219	●	.36	.46	.22	.81	.44	.21	.82
K17a	.72	.68	.611	●	.70	.68	.52	.80	.73	.53	.80
K17b	.65	.57	.340	●	.60	<b>.17</b>	<b>.01</b>	<b>.82</b>	-	-	-
K17c	.89	.82	.187	●	<b>.85</b>	<b>.20</b>	<b>.20</b>	.81	-	-	-
K17d	.88	.98	.016	●	<b>.93</b>	<b>.07</b>	<b>-.03</b>	.81	-	-	-
K17e	.63	.63	1	●	.64	.39	.25	.81	.39	.23	.82
K18a	.88	.91	.429	●	<b>.90</b>	<b>.15</b>	<b>.14</b>	.81	-	-	-
K18b	.95	.88	.099	●	<b>.91</b>	<b>.20</b>	<b>.17</b>	.81	-	-	-
K18c	.91	.78	.017	●	<b>.85</b>	<b>.27</b>	.25	.81	-	-	-
K18d	.64	.61	.673	●	.62	.54	.33	.81	.51	.32	.81
K18e	.93	.84	.100	●	<b>.88</b>	<b>.24</b>	.27	.81	-	-	-
K20a	.83	.77	.332	●	.80	<b>.24</b>	.27	.81	<b>.27</b>	.27	.81
K20b	.90	.69	.001*	○	-	-	-	-	-	-	-
K20c	.86	.85	.823	●	<b>.86</b>	.39	.39	.80	.39	.41	.81
K20d	.77	.78	.853	●	.77	.54	.45	.80	.49	.43	.81
K20e	.62	.58	.633	●	.60	.49	.26	.81	.56	.29	.81
K21a	.90	.93	.579	●	<b>.91</b>	<b>.10</b>	<b>.20</b>	.81	-	-	-
K21b	.98	.57	.000*	○	-	-	-	-	-	-	-
K21c	.58	.60	.751	●	.59	.73	.53	.80	.76	.55	.80
K21d	.54	.64	.203	●	.59	.59	.40	.80	.61	.41	.81
K21e	.78	.78	1	●	.78	.51	.42	.80	.49	.40	.81
K22a	.70	.79	.208	●	.75	.54	.38	.80	.51	.38	.81
K22b	.26	.25	.858	●	<b>.25</b>	<b>.15</b>	<b>.04</b>	<b>.82</b>	-	-	-
K22c	.43	.40	.635	●	.41	.51	.24	.81	.51	.24	.82
K22d	.56	.62	.428	●	.59	.44	.24	.81	.44	.26	.81
K22e	.48	.54	.435	●	.51	.68	.44	.80	.73	.42	.81

*Bemerkung:* Für Kurzformulierungen der Items siehe Tab. 4; ausformulierte Items befinden sich im Testbogen (ab S. 205).

\* $p$ -Wert unterschreitet kritisches  $\alpha$ -Niveau nach Benjamini-Hochb.-Verfahren (s. Fußnote 2, S. 92).

niedrige Lösungswahrscheinlichkeit von Item 22b, welches nach der Wurfdauer einer Kugel bis zum Erreichen der maximalen Höhe  $y_{\max}$  fragt. Die Bewegung startet vom obersten Punkt einer schiefen Ebene, endet auf dieser und wird in einem gedrehten Koordinatensystem beschrieben. Drei Viertel der Studierenden waren der fälschlichen Ansicht, dass  $y_{\max}$  nicht nach der halben Wurfdauer erreicht ist, was vermutlich auf eine falsche Identifikation von  $y_{\max}$  zurückzuführen ist. Auch die schlechte Trennschärfe ( $D_j = 0.15$ ;  $r_{jt} = 0.04$ ) spricht für den Ausschluss des Items. Die Item-Kennwerte der optimierten Kerngruppe wurden nach dem Ausschluss auffälliger Items neu berechnet und sind ebenfalls in Tab. 14 aufgeführt, während Teststatistiken für beide Itemgruppen in Tab. 15 zusammengefasst sind. Trotz geringerer Itemzahl belegen die Kennwerte eine Verbesserung der Test-Reliabilität durch den Ausschluss psychometrisch unpassender Items. Die höhere Testschwierigkeit der optimierten Itemgruppe ist auf den Ausschluss überwiegend leichter Items aus der Kern-Itemgruppe zurückzuführen.

**Tab. 15:** Skalenkennwerte (KiRC)

	Anzahl Items	$P$	$SD$	$D$	$\alpha_C$
Kern-Itemgruppe	32	.71	.16	.40	.81
opt. Kern-Itemgruppe	22	.67	.21	.52	.82

### Subdimensionen des Instruments

Da die Faktorenanalyse der Kern-Itemgruppe zu keiner Lösung führte, wurde das Verfahren mit der optimierten Kern-Itemgruppe wiederholt. Die resultierende Faktorstruktur ist nicht eindeutig, siehe Tab. 16. Der erste Faktor beinhaltet primär Aufgaben zum Umgang mit Bezugssystemen, ein Item zum graphischen Ableiten sowie ein Stroboskopbild. Ferner

**Tab. 16:** Rotierte Faktorenlösung der optimierten KiRC-Itemgruppe

Item	Inhalt <sup>a</sup>	RF <sup>b</sup>	Faktoren							$\kappa$
			1	2	3	4	5	6	7	
K6	3	B↔D	.80	.19	.00	.07	.05	.10	-.02	.69
K9	3	B↔D	.70	-.02	.11	.13	.13	.13	-.01	.54
K17a	1	D↔F	.61	.29	.21	.06	-.02	-.06	.26	.58
K10	1	B	.49	-.10	.46	.17	-.01	.08	.30	.59
K21d	1	D↔F	.01	.76	.02	.17	-.03	.30	.11	.71
K21c	1	D↔F	.11	.72	.17	.15	.23	.14	.02	.68
K21e	1	D↔F	.18	.67	.20	-.03	.05	-.18	.10	.57
K5	1	B	-.02	-.07	.65	.12	.24	.06	.25	.57
K4	1	D	.19	.24	.63	.08	-.12	.08	.02	.52
K20a	5, 7	F	.00	.30	.51	-.13	.15	.07	-.13	.41
K20d	5, 7	F	.25	.39	.45	.10	.00	-.11	-.10	.44
K2	1	D	.41	.12	.44	-.03	.06	.24	.28	.51
K22d	3, 5, 7	B↔F	.00	.00	.08	.70	.37	-.21	-.03	.68
K22a	3, 5, 7	B↔F	.20	.24	-.06	.70	-.18	.15	.14	.66
K22c	3, 5, 7	B↔F	.03	.04	.12	.62	-.02	.04	-.05	.67
K22e	3, 5, 7	B↔F	.44	.05	-.08	.56	.11	.24	.02	.59
K3	2	B	.22	.25	-.16	.01	.67	-.24	.14	.66
K20e	5, 7	F	.02	.02	.26	.07	.64	.17	.03	.51
K17e	1	D↔F	.04	.08	.10	.06	-.01	.80	.08	.67
K20c	5, 7	F	.26	.20	.30	.11	.31	.35	-.30	.52
K11	5, 6	B	.12	.06	-.05	.16	.29	.19	.69	.64
K18d	5, 7	F	.08	.16	.28	.22	-.12	-.10	.56	.50

*Bemerkung:* Faktorladungen mit Betrag  $< 0.30$  sind kontrastarm dargestellt.

<sup>a</sup>Inhalte: 1 = Zusammenhang  $x - v - a$ , 2 = Freier Fall mit Reibung, 3 = Bezugssysteme, 4 = Bahnkurven, 5 = Vektorzerlegung, 6 = Rotationskinematik, 7 = Wurfbewegungen.

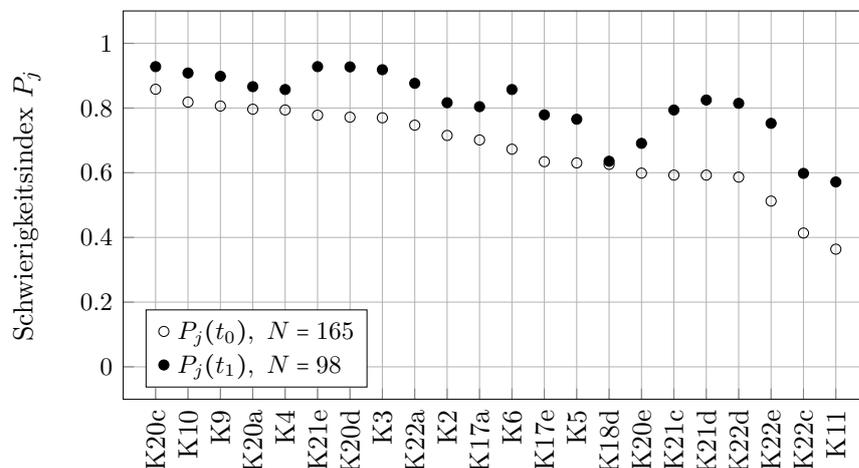
<sup>b</sup>Repräsentationsform (B = Bildhaft, D = Diagramm, F = Formal).

gruppieren Items innerhalb einer MTF-Gruppe zusammen. Die Faktoren 6 und 7 werden aus drei Items mit sehr niedrigen Faktorladungen gebildet, die sowohl inhaltlich auch als formatspezifisch unterschiedlich sind. Aus dieser Lösung können keine Subdimensionen des Instruments erkannt werden, die inhaltshomogen sind. Darüber hinaus zeigt das Ergebnis, dass sich Items mit gleichem Repräsentationstyp (graphisch, G, bildhaft, B, formal F und Transfer dazwischen) psychometrisch nicht gruppieren.

Um die verschiedenen Repräsentationen im Instrument in der späteren Vergleichsanalyse dennoch berücksichtigen zu können, wird auf die normative Zuordnung der Aufgaben zu den Repräsentationsformen zurückgegriffen (vgl. Tab. 4 auf S. 68, Spalte Repräsentationsform). Die Übereinstimmung dieser Zuordnung wurde durch ein Interrater-Verfahren abgesichert. Dazu wurden vier Rater aufgefordert, jedem Item genau eine der sechs Kategorien (G, F, B,  $G \leftrightarrow F$ ,  $G \leftrightarrow B$  und  $F \leftrightarrow B$ ) zuzuschreiben, die zur Lösung des jeweiligen Items primär benötigt wird. Die anschließend berechnete Interrater-Reliabilität nach Fleiss beträgt  $\kappa_F = 0.82$ , was auf eine „ausgezeichnete Übereinstimmung“ der Rater hinsichtlich der normativ zugeteilten Repräsentationsformen der Items deutet (Fleiss, 1971; Sim & Wright, 2005).

### Änderungssensitivität

Um festzustellen, ob die optimierte KiRC-Itemgruppe Lernerfolge auch statistisch nachweisen kann, werden die Prä- und Post-Daten miteinander verglichen. Die Lösungswahrscheinlichkeiten zu beiden Zeitpunkten sind in Abb. 17 dargestellt, wobei auf die jeweiligen Maximalstichproben zurückgegriffen wurde ( $N_{\text{Prä}} = 165$ ,  $N_{\text{Post}} = 98$ ). Wie zu erkennen ist, steigt die Lösungswahrscheinlichkeit jedes Items von der Prä- zur Postmessung an. Einzig bei Item K18d bleibt die Lösungswahrscheinlichkeit annähernd stabil. Die mittlere Lösungswahrscheinlichkeit steigt von  $P(t_0) = 0.67$  auf  $P(t_1) = 0.81$  an. Der Unterschied ist signifikant ( $t(261) = 5.62$ ,  $p < 0.001$ ) und entspricht einem mittelgroßen bis großen Effekt (Cohen's  $d = 0.71$ ). Innerhalb der verbundenen Stichprobe ( $N = 98$ ) steigt die Lösungswahrscheinlichkeit von  $P(t_0) = 0.69$  auf  $P(t_1) = 0.81$ . Der Leistungszuwachs ist ebenfalls signifikant ( $t(98) = 8.47$ ,  $p < 0.001$ ,  $d = 0.60$ ).



**Abb. 17:** Schwierigkeitsindizes (Lösungswahrscheinlichkeiten)  $P_j$  der optimierten Kern-Items des KiRC-Inventars zum Prä-Zeitpunkt ( $t_0$ ) und Post-Zeitpunkt ( $t_1$ ). Die Items sind absteigend nach dem Schwierigkeitsindex  $P_j(t_0)$  sortiert.

### 6.4.5 Konzepttest

#### Item- und Skalenanalyse

Wie die Analyseergebnisse in Tabelle 17 zeigen, liegen die Itemkennwerte allesamt im gewünschten Bereich. Die sehr hohen Diskriminationsindizes sind aufgrund der geringen Itemzahl unzuverlässig. Bei maximal fünf möglichen Punkten fallen einzelne richtig beantwortete Items stark ins Gewicht und beeinflussen dementsprechend stark die Testsumme, welche die Quantilaufspaltung in Gl. (4) definiert. Deshalb ist es kein Widerspruch, dass die korrigierten Trennschärfeparameter, die von dem Einfluss des jeweiligen Items selbst bereinigt sind, nur niedrig bis mittelmäßig ausfallen. Dies erklärt auch die schlechte Gesamtreliabilität dieses Kurztests ( $\alpha_c = 0.51$ ), die nicht zuletzt auch Konsequenz der geringen Itemanzahl ist.

**Tab. 17:** Ergebnisse der Itemanalyse (Konzepttest)

Item	Kurzbeschreibung	$P_j$	$D_j$	$r_{jt}$	$\alpha_{\setminus\{j\}}$
Kon1	Kugel wird seitlich aus fahrenden Wagen geworfen	.43	.88	.42	.36
Kon2	Scheibe ( $v = const.$ ) bekommt Kick $\perp v$	.61	.62	.25	.48
Kon3	Raumschiff ( $v = const.$ ) beschleunigt $\perp v$	.57	.68	.29	.45
Kon4	Flugzeug verliert Kugel	.51	.65	.22	.49
Kon5	Kugel auf Kreisbahn, Schnur reißt	.66	.74	.24	.48
<u>Skalenkennwerte</u>		$P$	$SD$	$D$	$\alpha_c$
Konzepttest		.56	.29	.71	.51

#### Studienübergreifender Vergleich der Itemstatistiken

*Lösungswahrscheinlichkeiten.* Zum Vergleich der gemessenen Item-Schwierigkeiten wurden die Daten einer leistungsschwächeren Population (Schulklasse) sowie einer vergleichbaren Population (Studierende einer anderen Hochschule) recherchiert. Aus Publikationen konnten Daten (a) von bayrischen Schülerinnen und Schülern der 11. Jahrgangsstufe vor und nach einem konventionellen Mechanikunterricht (Wilhelm & Heuer, 2005), (b) von kroatischen Gymnasiasten vor dem Unterricht und (c) von kroatischen Studierenden nach einem Mechanikkurs (Planinic et al., 2010) extrahiert werden, siehe Tab. 18. Zudem wurden die Post1 sowie die Post2-Daten aus der vorliegenden Studie ergänzt. Bei jedem Item schneiden die beiden Studierendengruppen im Prä-Test besser ab als die Schülerinnen und Schüler. Die bayrischen Schülerinnen und Schüler schließen nach dem Unterricht zwar zu den Studierenden auf, liegen aber mit Ausnahme von Item Kon5 noch unterhalb der Studierendenleistungen. Die Leistungen der beiden Studierendengruppen sind untereinander sehr ähnlich (Mittelwerte  $P = 0.59$  bzw.  $P = 0.62$  für die hiesigen / kroatischen Studierenden). Ferner zeigt sich, dass sich alle Schwierigkeitsindizes in der durchgeführten Hauptstudie im Laufe des fortschrei-

**Tab. 18:** Stichprobenübergreifender Vergleich der Schwierigkeitsindizes einzelner FCI-Items

Item	TU KL (HS)			Bayern		Kroatien	
	Prä (134)	Post1 (96)	Post2 (81)	Prä (258)	Post (258)	Gym (1676)	Uni (141)
Kon2-Scheibe	.61	.73	.77	.37	.37	.39	.55
Kon3-Raumschiff	.57	.70	.74	.21	.31	.25	.51
Kon4-Flugzeug	.51	.63	.68	.28	.49	.20	.77
Kon5-Kreis	.66	.83	.88	.48	.72	.48	.65
Mittelwert	.59	.73	.77	.34	.47	.33	.62

*Bemerkungen:* TU KL = Physik-Studierende der Technischen Universität Kaiserslautern, HS = Hauptstudie. Die jeweilige Stichprobengröße ist in Klammern angegeben.

tenden Semesters verbesserten. Diese Ergebnisse indizieren eine stichprobenübergreifende Diskriminationseigenschaft und eine Änderungssensitivität der Items.

*Reliabilitäten.* Schecker und Gerdes (1999) sowie Wilhelm und Heuer (2005) berichten von niedrigen Reliabilitäten des FCI-Instruments insgesamt und einzelnen Subskalen im Besonderen. Wilhelm und Kollegen verweisen diesbezüglich auf die Bemerkung von Hestenes und Halloun (1995), dass „Lernende, die nicht über das Newtonsche Konzept verfügen, keine konsistente Alltagstheorie [haben]“. Hier zeigt sich insbesondere, dass die Reliabilität  $\alpha_C$  als Maß der internen Skalenkonsistenz nicht nur von der Güte der Items selbst abhängt, sondern gerade im Fall von Leistungstests auch von der Wissensstruktur der Lernenden bestimmt wird. Die gemessene niedrige Reliabilität legt den Schluss nahe, dass die fünf Items trotz ihres sehr ähnlichen Charakters (sowohl von der Darstellungsform als auch vom Inhalt) keine konsistente Einheit für die Lernenden bilden. Eine Erklärung liegt vermutlich in den stark unterschiedlichen Kontexten, die in den Items abgebildet werden (Auto, Puck, Flugzeug, Raumschiff, Hammerwerfer; vgl. Huffman & Heller, 1995). Aus der physikdidaktischen Forschung ist bekannt, dass der Kontext einer Aufgabe mit der Aufgabenschwierigkeit zusammenhängt (Härtig, 2014; Stewart et al., 2007). Zwar beziehen sich diese Befunde auf Lernende der Sekundarstufe 2, aber die Ergebnisse sprechen dafür, dass diese Kontexteinflüsse auch bei Physikstudierenden eine Rolle spielen.

#### 6.4.6 Antwortsicherheiten

Bezüglich den Sicherheitsangaben liegen vollständige Datensätze zu den fünf Konzept-Items und zu 11 KiRC-Items vor. Es wurden zwei Faktoren extrahiert, die zusammen 57% der Varianz aufklären (vgl. Screeplot, Abb. A-1). Eine anschließende Rotation der Lösung verdeutlicht, dass sich die Sicherheits-Items des KiRC-Tests und des Konzept-Tests jeweils zu einem Faktor gruppieren (siehe Tab. 19). Dies bedeutet, dass Sicherheits-Items zu Repräsentationsaufgaben stärker untereinander korrelieren als mit Sicherheits-Items zu Konzeptaufgaben (und umgekehrt), was den unterschiedlichen Charakter der zugrundeliegenden Aufgaben betont: Die Einschätzung der eigenen Leistung hängt vom Aufgabentyp (repräsentations-/konzeptbezogen) ab. Eine getrennte Analyse beider Skalen zeigt, dass sowohl Item- als auch Skalenkennwerte im erwünschten Bereich liegen (Tab. 19).

**Tab. 19:** Faktoren- und Itemanalyse, sowie Skalenkennwerte (Antwortsicherheiten)

Items	Faktorenanalyse			Itemanalyse				
	1	2	$\kappa$	$M_j$	$SD_j$	$D_j$	$r_{jt}$	$\alpha_{\setminus\{j\}}$
KiRC4-S	.79	.24	.68	.83	.25	.42	.66	.87
KiRC10-S	.77	.23	.65	.81	.27	.48	.67	.87
KiRC2-S	.76	.17	.62	.87	.02	.36	.68	.87
KiRC1-S	.69	.08	.49	.72	.27	.46	.59	.87
KiRC3-S	.67	.09	.46	.79	.26	.42	.60	.87
KiRC9-S	.67	.29	.53	.71	.30	.55	.62	.87
KiRC7-S	.66	.18	.47	.77	.27	.43	.57	.88
KiRC8-S	.65	.29	.51	.81	.23	.40	.62	.87
KiRC6-S	.62	.30	.47	.73	.30	.50	.59	.87
KiRC5-S	.54	.25	.36	.65	.29	.51	.56	.88
KiRC11-S	.52	.39	.42	.53	.33	.60	.53	.88
Konzept4-S	.23	.86	.80	.58	.32	.70	.80	.83
Konzept3-S	.31	.78	.70	.58	.31	.65	.74	.84
Konzept2-S	.25	.78	.66	.65	.31	.66	.70	.85
Konzept1-S	.22	.77	.64	.64	.29	.59	.66	.86
Konzept5-S	.12	.76	.59	.74	.32	.61	.64	.87
<b>Testsummenwerte (Skalenanalyse)</b>				$M$	$SD$	$D$	$\alpha_C$	
Antwortsicherheiten (KiRC)				.75	.18	.47	.88	
Antwortsicherheiten (Konzepttest)				.64	.25	.64	.88	

*Bemerkung:* Faktorladungen mit Betrag  $< 0.30$  sind kontrastarm dargestellt.

### 6.4.7 Betreuerfragebogen

Eine Faktorenanalyse ergab, dass alle fünf Betreuer-Items auf einem Faktor laden. Die Skala kann demnach als eindimensional betrachtet werden. Die Item- und Skalenstatistiken sind in der Ergebnistabelle aufgeführt und zeigen keine psychometrischen Auffälligkeiten, sodass alle Items zur Skalenbildung beibehalten werden (vgl. Tab. 20).

**Tab. 20:** Faktoren- und Itemanalyse, sowie Skalenkennwerte (Betreuer-Items)

Item	Kurzformulierung	FA		Itemanalyse				
		1	$\kappa$	$M_j$	$SD_j$	$D_j$	$r_{jt}$	$\alpha_{\setminus j}$
B1	Übungsleiter (ÜL) spornt an	.83	.69	.61	.23	.49	.71	.75
B2	ÜL ist begeistert von Phy	.67	.45	.73	.17	.31	.52	.81
B3	ÜL ist engagiert	.80	.64	.79	.15	.30	.65	.78
B4	ÜL geht auf Fragen ein	.69	.48	.83	.17	.31	.52	.81
B5	ÜL motiviert zur Mitarbeit	.82	.67	.63	.23	.46	.69	.76
<u>Skalenkennwerte</u>				$M$	$SD$	$D$	$\alpha_c$	
Betreuerskala				.75	.18	.47	.88	

## 6.5 Ergebnisse und Diskussion III: Normalverteilungsanalyse der Skalenmittelwerte

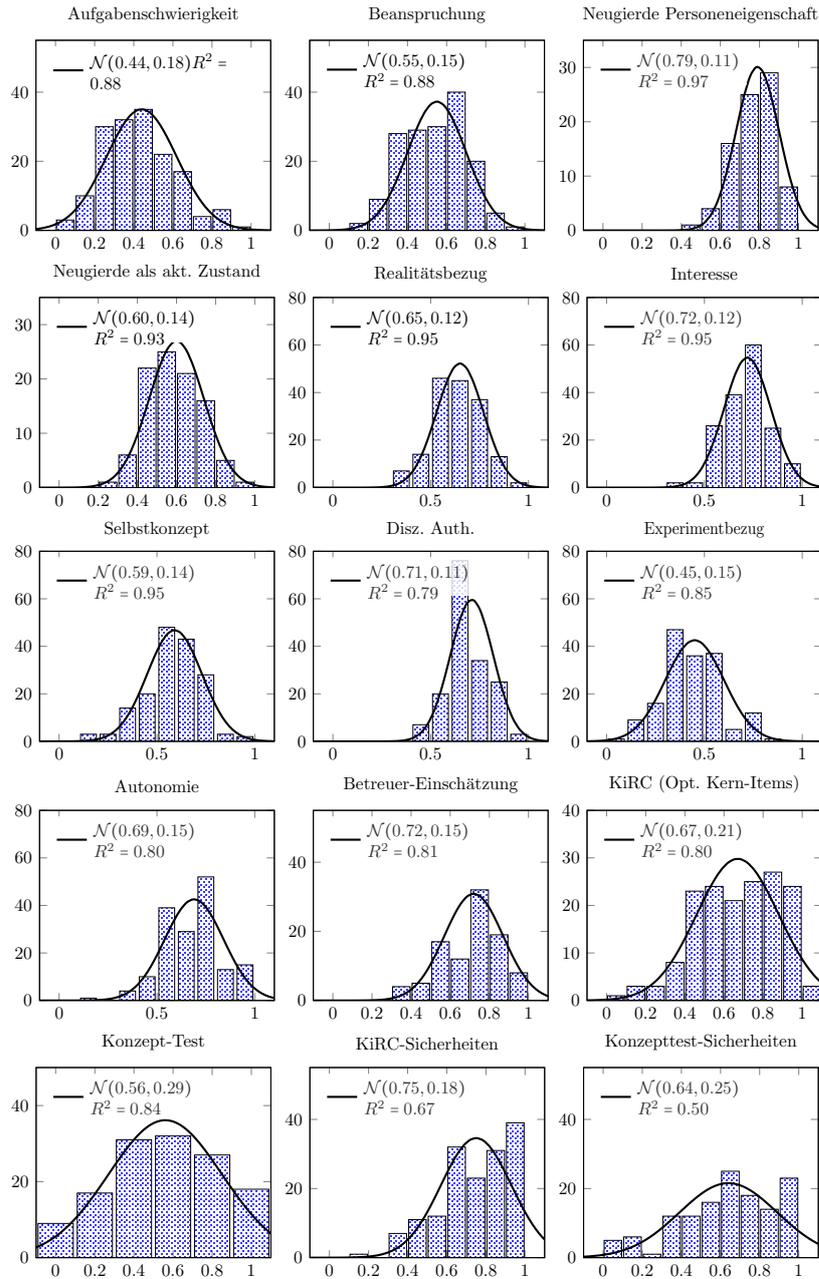
Die nach Itemausschluss gebildeten Skalen wurden einer Normalverteilungsanalyse unterzogen, deren Ergebnisse in Tab. 21 dargestellt sind. Die zugehörigen Häufigkeitsverteilungen befinden sich in Abb. 18.

Der Shapiro-Wilk-Test (SW-Test) wird für vier Motivations-Skalen signifikant (Selbstkonzept, disziplinäre Authentizität, Experimentbezug, Autonomie), weshalb Schiefe und Kurtosis betrachtet werden. Die Absolutwerte dieser Parameter sind für die genannten Variablen kleiner als 1, d.h. die empirischen Verteilungen weichen nur unbedeutend von einer Normalverteilung ab, was durch den visuellen Eindruck der Histogramme bestätigt werden kann.

**Tab. 21:** Ergebnisse der Normalverteilungsanalyse

Skala	SW-Test <sup>a</sup>			Schiefe		Kurtosis		$R^2$
	$W$	$df$	$p$ -Wert	$s$	$z$ -Wert	$k$	$z$ -Wert	
Beanspruchung	.99	164	.13	-.93	-.49	-.55	-1.46	.88
Aufgabenschwierigkeit	.99	164	.16	.31	1.65	-.55	-1.46	.88
Neugierde Personeneig.	.98	83	.31	-.37	-1.39	-.14	-.27	.97
Neugierde akt. Zustand	.99	97	.89	.11	.46	-.20	-.40	.93
Realitätsbezug	.99	163	.50	-.12	-.63	-.22	-.57	.95
Interesse / Engagement	.99	164	.09	-.29	-.15	.43	1.13	.95
Selbstkonzept	.98	163	.01	-.39	-2.03	.37	.98	.95
Disz. Authentizität	.97	162	< .01	-.48	-2.51	.30	.79	.79
Experimentbezug	.98	164	.01	.22	1.16	.19	.49	.85
Autonomie	.98	162	.01	-.19	-.97	.05	.12	.80
Repr. Komp. (KiRC)	.96	165	< .01	-.49	-2.57	-.44	-1.16	.80
Konzeptverständnis	.93	140	< .01	-.11	-.56	-.84	-2.07	.84
Sicherheit (KiRC)	.94	159	< .01	-.95	-4.95	1.61	4.20	.67
Sicherheit (Konzept)	.95	124	< .01	-.58	-2.65	-.26	-.61	.50
Betreuerverhalten	.97	97	.02	-.46	-1.89	-.33	-.77	.81

<sup>a</sup>Shapiro-Wilk-Test mit Teststatistik  $W$ , Freiheitsgraden  $df$  und  $p$ -Wert. Variationen in  $df$  sind auf fehlende Skalenwerte zurückzuführen.



**Abb. 18:** Histogramme der Skalenmittelwerte mit Normalverteilungskurven  $\mathcal{N}(M, SD)$  und Bestimmtheitsmaß  $R^2$ . Die Ordinate weist die absolute Häufigkeit aus.

Auch die  $z$ -Werte von Schiefe und Kurtosis liegen allesamt innerhalb der Toleranzgrenze ( $|z| < 3.29$ ), was vorherige Aussage statistisch absichert. Auf das Korrigieren von Ausreißern wird deshalb verzichtet.

Auch die Verteilung der KiRC-Mittelwerte weicht nach dem SW-Test signifikant von einer Normalverteilung ab. Die empirische Verteilung ist leicht linksschief ( $s = -0.49$ ,  $z = -2.57$ ) und flachspitziger als eine Normalverteilung ( $k = -0.44$ ,  $z = -1.16$ ). Sowohl die Absolutwerte als auch die  $z$ -Werte dieser Parameter attestieren allerdings eine hinreichende Übereinstimmung mit einer Normalverteilung für parametrische Auswerteverfahren. Die

Verteilung der erreichten Konzepttest-Mittelwerte ist annähernd symmetrisch (Schiefe  $s = -0.11$ ,  $z = -0.56$ ) und flachgipflig (Kurtosis  $k = -0.84$ ,  $z = -2.07$ ) und im Rahmen der Parameterwerte für Schiefe und Kurtosis als normal zu betrachten.

Die Normalverteilungstests der Sicherheitsskalen fallen signifikant aus. Die Parameterwerte von Schiefe und Kurtosis zu den Antwortsicherheiten des Konzepttests liegen im unkritischen Bereich (Beträge  $< 1$ ). Wie die Verteilung der KiRC-Sicherheiten zeigt (vgl. Histogramm in Abb. 18), liegt ein starker Deckeneffekt vor (d.h. viele Studierende waren sich beim Beantworten der Fragen sehr sicher). Die  $z$ -Werte der Parameter überschreiten entsprechend die Toleranzgrenzen. Folglich liegt keine Normalverteilung vor und die Antwortsicherheiten des KiRC-Tests eignen sich nicht für parametrische Analyseverfahren.

Der SW-Test zu den Mittelwerten der Betreuerskala fällt zwar signifikant aus, die Parameterwerte von Schiefe und Kurtosis sprechen aber für eine hinreichende Übereinstimmung zur Normalverteilung dieser Kontrollvariablen.

## 6.6 Zusammenfassung

In diesem Kapitel wurden die zuvor theoretisch konstruierten bzw. aus der Literatur übernommenen und adaptierten Items zur Erfassung relevanter latenter Konstrukte analysiert. Ausgehend von dem dimensionsreduzierenden und -überprüfenden Verfahren der Faktorenanalyse sowie der Bestimmung von psychometrischen Kenngrößen zur Quantifizierung der Güte einzelner Items bzw. ganzer Skalen wurden die Itemanalysen nach einem einheitlichen Leitfaden durchgeführt (vgl. Abb. 14). Dadurch war es möglich, aus allen eingesetzten Fragebögen zuverlässige Skalen zu extrahieren, die die zu messende Merkmalausprägungen valide quantifizieren (vgl. Tab. 22):

Obwohl die im Studium wahrgenommene Beanspruchung (PSQ) sowie die empfundenen Leistungsanforderungen (CL) inhaltlich eng verwandte Konstrukte sind, bestätigte sich die a priori Unterteilung der Items eindeutig. Alle Items wurden in die Skalenbildung aufgenommen. Die Kenngrößen beider Skalen liegen ausschließlich im gewünschten Bereich und die Skalenmittelwerte sind innerhalb der Eich-Stichprobe normalverteilt. Dies gilt ebenso für die beiden Neugierde-Skalen (Neugierde als Personeneigenschaft, NP, und als aktueller Zustand, NZ) mit Ausnahme des Items NP6, welches sich als unpassend herausstellte.

Bei der Analyse des Motivationsfragebogens zeigte sich durch das Zusammenwirken der ergänzten Items zur Authentizität und Autonomie mit den originären MAI-Items erwartungsgemäß eine Vermischung der neuen und originären Skalen. Die Lösung der Hauptkomponentenanalyse wurde unter Berücksichtigung der statistisch bedeutsamsten Faktoren interpretiert. Es wurde deutlich, dass sich die experimentbezogenen Items (Datenerhebung, Hypothesenprüfung, usw.) eigenständig gruppieren und sich von den Items zum Professionsbezug (Arbeitsweisen der Physik) abgrenzen. Für spätere Vergleichsanalysen wird der Experimentbezug demnach nicht ausschließlich dem Motivationskonstrukt zugeschrieben, sondern wird auch als Variable im Sinne eines „Manipulation Checks“ genutzt. Damit ist gemeint, dass die Variable ein wichtiges Prüfkriterium darstellt, ob der im Studiendesign vorgesehene Unterschied zwischen den Aufgaben auch tatsächlich — d.h. statistisch signifikant und mit relevanter Effektstärke — durch die Lernenden wahrgenommen wird. Die SK-Skala enthält überwiegend kognitiv-fachbezogene Items, die IE-Skala sowohl Items zur persönlichen Bedeutsamkeit des Lerninhalts als auch zu gefühlsbezogenen Valenzen und die originäre RA-Skala wird gut reproduziert. Alle Skalen sind in Anbetracht konventioneller Entscheidungsregeln als annähernd normalverteilt anzunehmen.

Um den Repräsentationstest KiRC zu analysieren, wurde zunächst eine Kern-Itemgruppe identifiziert, die eine Schnittmenge von studienübergreifend vergleichbaren Items enthielt. Von diesen 32 Items stellten sich einige als zu einfach dar (hohe Lösungswahrscheinlichkeiten bzw. Schwierigkeitsindizes  $P_j \geq 0.85$ ) und wurden in Folge des geringen Informationsgehalts entfernt. Die resultierende optimierte Kern-Itemgruppe (22 Items) weist zufriedenstellende psychometrische Kennziffern auf und ist sensitiv gegenüber Leistungsänderungen. Der Test ist

für einen Leistungstest trotz der Kürze bemerkenswert reliabel ( $\alpha_C = 0.82$ ), was vermutlich auf den gemeinsamen Grundinhalt Kinematik zurückgeführt werden kann, der bei der Beantwortung jedes Items eine Rolle spielt. Die Faktorenanalyse lieferte keine eindeutigen Hinweise auf Subdimensionen. Insbesondere ließen sich keine Repräsentationsgruppen finden (z.B. Graphen oder Bilder) und auch inhaltlich eng verwandte Items (z.B. Wurfbewegungen) gruppierten nicht zusammen. Um die Zuteilung einzelner Items zu den Repräsentationsformen dennoch statistisch abzusichern, wurde ein Expertenrating durchgeführt. Mit hoher Übereinstimmung kann jedem Item eine Repräsentationskategorie normativ zugeordnet werden, die bei Vergleichsanalysen dazu genutzt werden, Lernunterschiede auf Repräsentationsebene zu messen.

Da der Kinematik-Konzepttest über Trajektorien vier FCI-Items enthält, konnten die Lösungswahrscheinlichkeiten der Eich-Stichprobe mit denen anderer Studien verglichen werden. Erwartungskonform zeigten sich mit dem Leistungsstand wachsende Lösungswahrscheinlichkeiten. Die in der Eich-Stichprobe gemessene Reliabilität des Kurztests ist inakzeptabel, was auf geringe Inter-Item-Korrelationen sowie auf die Testlänge (insgesamt fünf Items) zurückgeführt werden kann. Dieses Ergebnis steht in Einklang mit bisherigen Befunden bezüglich des FCI und daraus entnommenen Subskalen und wird mit einem inkohärenten Verständnis der Lernenden begründet. Nichtsdestotrotz trägt die Skala zur Einschätzung des Interventionserfolgs bei und wird gerade wegen der hohen Etablierung in der Forschungsgemeinschaft für die Vergleichsuntersuchung verwendet.

Die Items zur Messung der Antwortsicherheit, die zu einigen Aufgaben der beiden Leistungstests erhoben wurden, gruppieren sich testspezifisch und zeigen keine psychometrischen Auffälligkeiten. Die Skalenwerte zu den KiRC-Sicherheiten sind nicht normalverteilt, weshalb sich diese nicht für parametrische Analyseverfahren eignen. Die Auswertung der Antwortsicherheiten erfolgt in Kap. 8 aus dort dargestellten inhaltlichen Gründen hauptsächlich auf Itemebene, weshalb das Ergebnis bezüglich den dort gewählten Methoden keine Einschränkung darstellt. Für die Vergleichsuntersuchung werden die Skalenwerte der Antwortsicherheiten nicht genutzt.

**Tab. 22:** Zusammenfassung der Ergebnisse zur Skalenbildung

	# Items		$P/M$	$SD$	$D$	$\alpha_C$	$\mathcal{N}^a$
	a priori	final					
Zielbereich	-	-	[0.35, 0.85]	-	$\geq 0.30$	$\geq 0.70$	
Aufgabenschwierigkeit	10	10	.44	.18	.45	.93	●
Beanspruchung	6	6	.55	.15	.42	.81	●
Neugierde Personeneig.	9	8	.79	.11	.28	.80	●
Neugierde akt. Zustand	8	8	.60	.14	.36	.85	●
Realitätsbezug	9	8	.65	.12	.32	.83	●
Interesse / Eng.	16	7	.72	.12	.31	.76	●
Selbstkonzept	7	5	.59	.14	.33	.74	●
Disziplinäre Auth.	12	7	.71	.11	.27	.77	●
Experimentbezug	12	5	.45	.15	.38	.72	●
Autonomie	4	3	.69	.15	.37	.71	●
Repräsentationskomp.	32	22	.67	.21	.52	.82	●
Konzeptverständnis	5	5	.56	.29	.71	.51	●
Sicherheit (Repr.)	11	11	.75	.18	.47	.88	○
Sicherheit (Konzept)	5	5	.64	.25	.64	.88	●
Betreuerskala	5	5	.72	.15	.37	.82	●

<sup>a</sup>Skalenmittelwerte können als normalverteilt betrachtet werden (●); vgl. Kap. 6.5 und Histogramme in Abb. 18 im Anhang, S. 99.



# Kapitel 7

## Testanalysen nach Item-Response-Theorie (IRT)

In Kapitel 6.4.4 und 6.4.5 wurden die Items der beiden Leistungstests gemäß der klassischen Testtheorie (KTT) analysiert, um möglichst reliable und ökonomische Skalen zur Erfassung der Repräsentationskompetenz und des Konzeptverständnisses im Themengebiet Kinematik zu erhalten. In diesem Kapitel werden zunächst einige Unzulänglichkeiten der KTT herausgestellt, die von der Item-Response-Theorie (IRT) methodisch gelöst werden. Das einfachste IRT-Modell, das Rasch-Modell, wird beschrieben und dessen zentralen Eigenschaften werden genannt. Nach Darstellung dieser methodischen Grundlagen erfolgt die Anpassung der empirischen Leistungstestdaten an die Gleichungen des Rasch-Modells. Es wird eine Beziehung zwischen Testsumme und Personenfähigkeit hergestellt, die stichprobenübergreifend gültig ist. Im Gegensatz zur klassischen Testtheorie kann mittels IRT *überprüft* werden, ob es überhaupt gerechtfertigt ist, mit den Testsummen als valide Leistungsindikatoren zu arbeiten. Wie aus der Diskussion der Modelleigenschaften hervorgeht, ist die Gültigkeit des Rasch-Modells nämlich die Voraussetzung dafür, dass die Testsumme eine sinnvolle Aussage über die Fähigkeit einer Person macht (Rost, 1996, S. 128).

Trotz dieser herausragenden Bedeutung, Skalen mit höchster statistischer Qualität zu identifizieren, ist die Rasch-Analyse in der Forschung noch wenig etabliert (Amelang & Schmidet-Atzert, 2006, S. 61).<sup>1</sup> Die Rasch-Analyse des KiRC-Instruments liefert stichproben-unabhängige Itemparameter zu repräsentationsbezogenen Aufgaben, die studienübergreifend verglichen werden können. Dies ist mit Blick auf die in der Einleitung dargestellten Übergangproblematik zwischen Schule und Hochschule besonders nützlich, denn damit sind Leistungsvergleiche zwischen verschiedenen Lerner-Gruppen — zum Beispiel zwischen Studierenden verschiedener Universitäten, zwischen Studierenden aus verschiedenen Erstsemester-Jahrgängen sowie zwischen Lernenden zum Ende ihrer Schullaufbahn und Studierenden in der Studieneingangsphase — möglich.

### 7.1 Methodische Grundlagen der IRT und des Rasch-Modells

In der klassischen Testtheorie wird üblicherweise die Anzahl richtig gelöster Aufgaben (Testsumme) als Maß für die Fähigkeit einer Person genutzt (Moosbrugger & Augustin, 2007). In dieser Zuordnung liegen einige Unzulänglichkeiten (vgl. Neumann, 2014, S. 356):

---

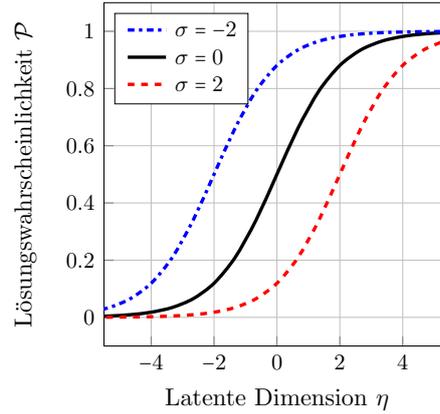
<sup>1</sup>Hauptgründe dafür sind fehlende Routinen bei den Programmpaketen zur Durchführung der Analyse (vgl. Kap. 7.2.1) und der große Aufwand zur Konstruktion rasch-skalierbarer Tests (Amelang & Schmidet-Atzert, 2006, S. 61).

- Zwei Personen mit der gleichen Anzahl richtig gelöster Aufgaben werden als gleich fähig betrachtet, unabhängig davon, *welche* Aufgaben sie gelöst haben. In der klassischen Testtheorie wird dies durch folgende Argumentation begründet: Löst eine Person eine Aufgabe mit Schwierigkeit  $P'_j$ , so sollte diejenige Person auch alle Aufgaben mit  $P_j < P'_j$  lösen. In der KTT wird stillschweigend angenommen, dass eine personenunabhängige Hierarchie der Aufgabenschwierigkeiten existiert und dass allein die Personenfähigkeit darüber entscheidet, ob eine Aufgabe gelöst wird oder nicht. Beide Annahmen sind in der KTT methodisch nicht überprüfbar.
- In der KTT ist die latente Variable (Personenfähigkeit) beidseitig beschränkt und der Zusammenhang zwischen der Anzahl gelöster Aufgaben und Fähigkeit wird als linear angenommen. Dies führt dazu, dass einem Studenten ohne richtig gelöste Aufgabe die niedrigste Fähigkeit 0 (und mit ausschließlich richtigen Aufgaben die Fähigkeit 1) zugeordnet wird, obwohl sich mit Sicherheit ein Student finden lässt, der noch weniger kompetent (bzw. noch kompetenter) ist. Hat ein Student  $k > 0$  Aufgaben richtig gelöst, müsste ein *doppelt so fähiger* Student unter Umständen mehr Aufgaben lösen, als im Test vorhanden sind. Auch auf Itemebene wird das Problem deutlich: Wenn eine lineare Beziehung zwischen Personenfähigkeit und Lösungswahrscheinlichkeit (Item-Schwierigkeit) besteht (lineare Itemfunktion), dann muss die Itemfunktion aufgrund der Beschränktheit der Lösungswahrscheinlichkeit auf das Intervall  $[0, 1]$  zwangsläufig an den Rändern abgeknickt sein. Dies ist unplausibel und psychologisch nicht interpretierbar (Rost, 1996, S. 114). Um diesen Problemen zu entgegnen, muss zum einen das Konzept linearer Itemfunktionen verworfen werden und zum anderen müssen unbeschränkte Fähigkeitsausprägungen zugelassen werden.
- Die Fähigkeiten sind abhängig von der präsentierten Itemstichprobe und umgekehrt werden die Itemparameter (und damit die Testkennwerte) durch die Erhebungsstichprobe bestimmt. Diese Stichprobenabhängigkeit ist aus messtechnischer Sicht unbefriedigend, da damit generalisierte Aussagen zu Testwerten nicht möglich sind.

Die IRT begegnet diesen Problemen mit einer Familie mathematisch-formaler Messmodelle, die dem beobachteten Testverhalten eine Fähigkeit zugrunde legen (Rost, 1996). Das Antwortverhalten wird in der IRT durch eine nicht-lineare Funktion modelliert, die eine Aussage über die *Wahrscheinlichkeit* einer Beobachtung in Abhängigkeit der Fähigkeitsausprägung trifft. Je nach Art des IRT-Modells beinhaltet die Antwortfunktion spezifische Parameter, die den Kurvenverlauf steuern. Im einfachsten IRT-Modell — dem Rasch-Modell — wird das Antwortverhalten einer Person  $i$  mit Fähigkeit  $\eta_i$  auf ein Item  $j$  mit Schwierigkeit  $\sigma_j$  durch eine Funktion  $\mathcal{P}(\eta_i, \sigma_j)$  beschrieben. Andere IRT-Modelle beschreiben das Antwortverhalten durch zusätzliche Parameter wie beispielsweise itemspezifische Diskriminationsparameter  $\delta_j$  (2PL-Modell oder Birnbaum-Modell) oder Rateparameter  $\gamma_j$  (3PL-Modell) und benötigen wegen der größeren Anzahl zu schätzender Parameter größere Stichproben ( $\gg 200$ ), weshalb auf diese Modelle in dieser Arbeit nicht weiter eingegangen wird. Das Rasch-Modell zeichnet sich gegenüber den erweiterten Modellen dadurch aus, dass die Modellparameter unabhängig voneinander geschätzt werden können (Rost, 1996, S. 136), woraus wichtige Messeigenschaften folgen (s.u.). Unabhängig vom gewählten Modell erfolgt die Messung in der IRT im Gegensatz zur KTT auf indirekte Weise und beruht auf *falsifizierbaren* Modellannahmen. Die Möglichkeit, ein Modell ablehnen zu können, stellt eine der größten Stärken der IRT dar und unterscheidet den Ansatz von dem der KTT (Strobl, 2012, S. 2).

### 7.1.1 Das Rasch-Modell

Das Rasch-Modell trifft die Annahme, dass die empirischen Leistungsdaten auf die Personenfähigkeiten und die Aufgabenschwierigkeiten zurückgeführt werden können. Die Wahrscheinlichkeit  $\mathcal{P}$  für die Beobachtung  $x_{ij} \in \{0, 1\}$  (als Realisierung der Indikatorvariable  $X_{ij}$ )



**Abb. 19:** Aufgabencharakteristische Kurven (ICC) im Rasch-Modell: Die Lösungswahrscheinlichkeit eines Items wird als Funktion der Personenfähigkeit  $\eta$  dargestellt. In dieser Darstellung ist die Itemschwierigkeit  $\sigma$  als Parameter aufzufassen, der zu einer Verschiebung der ICC entlang der  $\eta$ -Achse führt.

unter dem gegebenen Personenparameter  $\eta_i$  und Aufgabenparameter  $\sigma_j$  lässt sich gemäß der Modellgleichung ausdrücken als

$$\mathcal{P}(\eta_i - \sigma_j | X_{ij} = x_{ij}) = \frac{\exp\{x_{ij}(\eta_i - \sigma_j)\}}{1 + \exp\{\eta_i - \sigma_j\}}. \quad (7)$$

Gleichung (7) hat die Form einer subtraktiven Parametrisierung, da die Wahrscheinlichkeit für eine richtige Lösung ( $X_{ij} = 1$ ) nur von der Differenz  $\eta_i - \sigma_j$  abhängt. In Anlehnung an das Prinzip einer physikalischen Messung wird das Messen als Vergleich zweier Parameter aufgefasst, wobei explizit deutlich wird, dass beide Größen ( $\eta$  und  $\sigma$ ) auf einer gemeinsamen Skala angeordnet sind. Da die Wahrscheinlichkeit zur Lösung einer Aufgabe nur von der Differenz zweier Parameter abhängt, wird das Rasch-Modell auch 1PL-Modell (einparametriges logistisches Modell) genannt. Abbildung 19 illustriert den Zusammenhang zwischen Personenparameter  $\eta$  und Lösungswahrscheinlichkeit  $\mathcal{P}$  für Items verschiedener Schwierigkeit, wobei Gl. (7) als Funktion von  $\eta$  mit dem Parameter  $\sigma$  aufgefasst wird. Die dargestellten *Aufgabencharakteristischen Kurven* (item characteristic curves; ICCs)

$$P_\sigma(\eta) = \frac{\exp\{\eta - \sigma\}}{1 + \exp\{\eta - \sigma\}} \quad (8)$$

verlaufen parallel zueinander ( $\Delta\sigma = \text{const}$ ),

$$\mathcal{P}_\sigma(\eta) = \mathcal{P}_{\sigma + \Delta\sigma}(\eta + \Delta\sigma), \quad (9)$$

sind differenzierbar und streng monoton steigend,

$$\frac{d\mathcal{P}_\sigma(\eta)}{d\eta} = \frac{\mathcal{P}_\sigma(\eta)}{1 + \exp\{\eta - \sigma\}} > 0, \quad (10)$$

und es gilt

$$\mathcal{P}_\sigma(\eta = \sigma) = \frac{1}{2}, \quad (11)$$

$$\lim_{\eta \rightarrow -\infty} \mathcal{P}_\sigma(\eta) = 0, \quad (12)$$

$$\lim_{\eta \rightarrow \infty} \mathcal{P}_\sigma(\eta) = 1. \quad (13)$$

Die ICCs bilden also das latente Kontinuum  $\eta \in (-\infty, \infty)$  eineindeutig auf das beschränkte Intervall  $[0, 1]$  ab. Aus Gleichung (9) geht hervor, dass sich die ICCs nur durch eine Verschiebung um  $\Delta\sigma$  auf der  $\eta$ -Skala unterscheiden, woraus sofort folgt, dass die Skala nur bis auf Translationen eindeutig ist (Differenzskalenniveau). Diese Freiheit bei Festlegung der Itemparameter wird durch die Normierungsbedingung  $\sum_j \sigma_j = 0$  eingeschränkt. Ferner ist ersichtlich, dass eine Person mit der Fähigkeit  $\eta_1$  stets eine geringere Lösungswahrscheinlichkeit eines Items besitzt als eine Person mit einer größeren Fähigkeit  $\eta_2 > \eta_1$ , unabhängig davon, welche Aufgabe für den Vergleich herangezogen wird. Mit anderen Worten ist die Rangordnung der Lösungswahrscheinlichkeiten eines Items für zwei Personen unterschiedlicher Fähigkeit unabhängig von dem gewählten Item.

Genügen die empirischen Daten dem Rasch-Modell, so kann die Wahrscheinlichkeit zur Lösung eines (jeden) Items durch Gleichung (7) beschrieben werden und es gelten die zentralen Modelleigenschaften:

1. *Eindimensionalität.* Alle Aufgaben sprechen dieselbe latente Dimension  $\eta$  an. Gäbe es eine zweite Dimension, dann ließe sich  $\mathcal{P}$  nicht durch Gl. (7) ausdrücken.<sup>2</sup>
2. *Spezifische Objektivität.* Für den Vergleich zweier Personen ist es irrelevant, welche Items für den Vergleich herangezogen werden. Umgekehrt ist es für den Vergleich zweier Items unerheblich, welche Personengruppen die Fragen beantwortet. Diese Aussagen können nachvollzogen werden, indem der Wettquotient, definiert als

$$\frac{\mathcal{P}(X_{ij} = 1)}{\mathcal{P}(X_{ij} = 0)} = \exp\{\eta_i - \sigma_j\}, \quad (14)$$

logarithmiert und zwischen zwei Personen  $i$  und  $k$  verglichen wird. Der logarithmierte Wettquotient

$$\log \left( \frac{\mathcal{P}(X_{ij} = 1)}{\mathcal{P}(X_{ij} = 0)} \right) = \eta_i - \sigma_j \quad (15)$$

gibt zunächst an, dass die sogenannten „logit-transformierten Lösungswahrscheinlichkeiten“ (kurz Logits) linear von der Personenvariable abhängen, was bei den nicht transformierten Lösungswahrscheinlichkeiten der KTT auf Probleme stieß (Rost, 1996, S.121–125). Für den Vergleich zweier Personen  $i$  und  $k$  unterschiedlicher Fähigkeit  $\eta_i$  und  $\eta_k$  werden deren Logits verglichen, d.h.

$$\log \left( \frac{\mathcal{P}(X_{ij} = 1)}{\mathcal{P}(X_{ij} = 0)} \right) - \log \left( \frac{\mathcal{P}(X_{kj} = 1)}{\mathcal{P}(X_{kj} = 0)} \right) = \eta_i - \eta_k. \quad (16)$$

Da die rechte Seite von Gleichung 16 unabhängig von dem Itemparameter  $\sigma_j$  ist, ist der Vergleich zweier Personen unabhängig davon, mit welchen Items sie getestet werden. Umgekehrt ergibt der Vergleich zweier Items  $j$  und  $l$

$$\log \left( \frac{\mathcal{P}(X_{ij} = 1)}{\mathcal{P}(X_{ij} = 0)} \right) - \log \left( \frac{\mathcal{P}(X_{il} = 1)}{\mathcal{P}(X_{il} = 0)} \right) = \sigma_j - \sigma_l. \quad (17)$$

Die Differenzen zwischen zwei Itemparametern können *stichprobenunabhängig* bestimmt werden.<sup>3</sup>

3. *Suffizienz.* Die gesamte Information über die Personenfähigkeit bzw. Aufgabenschwierigkeit ist in der Anzahl der richtig gelösten Aufgaben enthalten — das Innere der

<sup>2</sup>In der Praxis ist diese Eigenschaft natürlich nur näherungsweise gegeben. Es ist fast unmöglich, vollständig eindimensionale Tests zu konstruieren, aber für Anwendungszwecke kann dies ausreichend genau realisiert werden (siehe z.B. Planinic et al., 2010). Dies gilt auch für die anderen Modelleigenschaften.

<sup>3</sup>Bühner (2006, S. 340) weist darauf hin, dass bei dem Vergleich stark unterschiedlicher Gruppen (z.B. Physik-Studierenden und Gymnasiasten) die Modellgeltung in beiden Gruppen gegeben sein muss.

Datenmatrix ist irrelevant. Diese Eigenschaft ist dadurch gegeben, dass die Likelihood der beobachteten Daten

$$L := \prod_i \prod_j \mathcal{P}(X_{ij} = x_{ij}) \quad (18)$$

nur von den Randsummen abhängt, wie sich durch Einsetzen der Modellgleichung (7) in Gl. (18) leicht prüfen lässt (Rost, 1996, S. 128–129). Die Suffizienz besagt also einerseits, dass von der Gesamtheit der empirischen Daten nur die Randsummen zur Parameterschätzung relevant sind und andererseits, dass das Rasch-Modell alle zur Verfügung stehenden Daten nutzt, um lokale Modellparameter zu schätzen.<sup>4</sup>

4. *Lokale stochastische Unabhängigkeit.* Die Lösung eines Items bedingt nicht die Lösung eines anderen. Auftretende Korrelationen zwischen den Items können durch die Personenfähigkeit (lokal) erklärt werden.

Diese attraktiven Modelleigenschaften folgen, wenn eine Itemgruppe die Voraussetzungen des Rasch-Modells erfüllt<sup>5</sup> und zeigen, wie die anfänglich beschriebenen Unzulänglichkeiten der KTT methodisch gelöst werden. Im Folgenden wird das Prinzip der Parameterschätzung dargestellt und es werden die Modellgeltungstests erläutert, die zur Überprüfung der Modelleigenschaften anhand empirischer Daten angewendet werden.

### 7.1.2 Prinzip der Parameterschätzung

Die rechte Seite der Modellgleichung (7) gibt die Wahrscheinlichkeit für die Beobachtung  $X_{ij}$  bei gegebenen Parametern  $\eta_i$  und  $\sigma_j$  an. Ausgehend von den beobachteten Werten 0 und 1 (Item richtig / falsch gelöst), werden die der Beobachtung zugrunde liegenden Parameter geschätzt (*indirekte Messung*). Hierfür existieren verschiedene Ansätze, die darin unterschieden werden können, ob die Aufgaben- und Personenparameterschätzung gleichzeitig oder nacheinander erfolgt (Strobl, 2012, S. 27–35). Ohne näher auf die Unterschiede der Ansätze einzugehen, liegt das Grundprinzip der Schätzung darin, Parameter zu finden, die mit größtmöglicher Wahrscheinlichkeit zu den beobachteten Daten führen (*Maximum-Likelihood-Schätzung*). Dies bedeutet, die Parameter  $\eta_i$  und  $\sigma_j$  so zu wählen, dass die Likelihoodfunktion  $L$  gemäß Gl. (18) maximiert wird. Aus dem Maximierungsproblem entstehen Bestimmungsgleichungen, die durch spezielle Algorithmen gelöst werden können (Rost, 1996, S. 300–308). Es kann gezeigt werden, dass die Parameterschätzung unter bestimmten Bedingungen eindeutig ist (Rost, 1996, S. 315–316).<sup>6</sup>

Die Interpretation der Parameterschätzwerte ist natürlich nur dann sinnvoll, wenn das Rasch-Modell auch tatsächlich gültig ist. Die zugrunde liegenden Annahmen, dass die Items nur eine Fähigkeit ansprechen (also homogen sind) und die ICCs durch Gl. (8) beschrieben werden können, müssen überprüft werden.

### 7.1.3 Modellgeltungstests nach Invarianzprinzipien

Die im Folgenden dargestellten Kriterien zur Modellprüfung basieren auf der Annahme, dass die Parameterschätzungen innerhalb zweier Teildatensätze nicht signifikant von der gemeinsamen Schätzung abweichen sollen (vgl. Strobl, 2012, S. 39). Die genannten Tests sind in Koller et al. (2012) ausführlich dargestellt.

<sup>4</sup>In der KTT hingegen berechnet sich der Schwierigkeitsindex eines Items völlig unabhängig von den Schwierigkeitsindizes anderer Items.

<sup>5</sup>Ausgehend von der Annahme der lokalen stochastischen Unabhängigkeit lässt sich umgekehrt die Modellgleichung (7) explizit herleiten, wenn die spezifische Objektivität gefordert und geringfügige weiteren Annahmen getroffen werden (Fischer, 2012).

<sup>6</sup>Vereinfacht ausgedrückt darf der Datensatz nicht in zwei separate Datensätze zerlegbar sein, für die nur getrennte Parameterschätzungen möglich sind.

1. *Gleichheit der Itemparameter in verschiedenen Subpopulationen.* Zwischen gezielt gewählten unterschiedlichen *Personengruppen* dürfen sich die Differenzen der geschätzten Itemparameter nicht systematisch unterscheiden. Ansonsten würden unterschiedliche Parameterschätzwerte die Schwierigkeit der Items in den beiden Gruppen besser widerspiegeln als eine gemeinsame Parameterschätzung. Werden zur Stichprobenaufteilung externe Teilungskriterien (z.B. Geschlecht, Alter usw.) herangezogen, wird von der Überprüfung von DIF gesprochen (*Differential Item Functioning*); interne Teilungskriterien, z.B. Teilung am Mittelwert, stehen im Zusammenhang mit der Diskrimination von Items. Die Nullhypothese („Die Parameterschätzung führt in beiden Gruppen zu gleichen Ergebnissen“) wird mithilfe des *Andersen LR-Tests* global geprüft. Im Falle der Modellverletzung wird die Subgruppeninvarianz mithilfe des *Wald-Tests* auf Itemebene untersucht.
2. *Gleichheit der Personenparameter in verschiedenen Itemgruppen (Itemhomogenität).* Zur Überprüfung der Gleichheit von Personenparametern wird der *Martin-Löf-Test* angewandt, nachdem die Teilung der Items in zwei Gruppen erfolgt.
3. *Test auf lokale stochastische Unabhängigkeit.* Ein Item  $j$  wird als Teilungskriterium gewählt, sodass sich zwei Personengruppen ergeben (diejenigen, die das Item  $j$  lösten und die, die es nicht lösten). Anschließend werden die Schwierigkeitsparameter in beiden Personengruppen geschätzt und miteinander verglichen, wobei Item  $j$  nicht in der Schätzung berücksichtigt wird (Wald-Test). Unterscheiden sich die Schwierigkeitsparameter eines Items  $l$  zwischen den Gruppen, so ist die Annahme der lokalen stochastischen Unabhängigkeit verletzt, da die Antwort auf Item  $l$  im Zusammenhang mit der Schwierigkeit des Items  $j$  steht.

Ergänzend zu den statistischen Tests werden graphische Modellkontrollen durchgeführt (Strobl, 2012, S. 40). Dazu werden die Parameterwerte der beiden Subgruppen auf verschiedenen Achsen gegeneinander aufgetragen oder DIF-Plots betrachtet. Grafische Modellkontrollen sind rein deskriptiv und im eigentlichen Sinne keine Modelltests.

#### 7.1.4 Globale Modellgeltungstests

Neben diesen Tests auf Invarianzeigenschaften existieren *globale Modelltests*, die auf einem Vergleich zwischen den empirischen Pattern (= Antwortmuster) und den bei Modellgültigkeit theoretisch zu erwartenden Pattern beruhen. Die *Likelihood-Ratio des saturierten Modells*, definiert durch

$$LR_{\text{sat}}^2 = -2 \ln \frac{L}{L_{\text{sat}}} = -2 \ln \frac{L}{\prod_{\mathbf{x}} \left( \frac{n(\mathbf{x})}{N} \right)^{n(\mathbf{x})}}, \quad (19)$$

stellt eine globale Prüfgröße dar. Gleichung (19) setzt die Likelihood des Rasch-Modells nach Gl. (18) ins Verhältnis zur Likelihood, die das Modell am besten beschreibt (Test gegen das saturierte Modell). Die Likelihood des saturierten Modells  $L_{\text{sat}}$  ergibt sich aus den Produkten der relativen Häufigkeiten der beobachteten Antwortpattern  $\mathbf{x}$ , die mit der Anzahl von Antwortpattern potenziert werden und stellt die größtmögliche Likelihood dar. Die Größe  $LR_{\text{sat}}^2$  ist  $\chi^2$ -verteilt und kann daher auf Signifikanz geprüft werden.<sup>7</sup> Der Test gegen das saturierte Modell stellt das strengste Testkriterium dar (Rost, 1996). Das Problem globaler Modellgeltungstests ist jedoch die große Anzahl theoretisch möglicher Datenmuster (z.B. sind bei 20 Items  $2^{20} \approx 10^6$  verschiedene Datenmuster möglich), deren empirische Beobachtung unrealistisch große Stichproben erfordert. Aus diesem Grund werden durch einen parametrischen Bootstrap künstliche Datensätze simuliert, in denen das Rasch-Modell

<sup>7</sup>Entgegen der in der Literatur üblichen Konvention werden die globalen Prüfgrößen ( $LR_{\text{sat}}$  und  $\xi^2$ ) nicht mit  $\chi^2$  bezeichnet, um Verwechslungen zwischen Prüfgröße und Verteilung zu vermeiden.

mit den empirisch geschätzten Modellparametern per Definition gilt. Anhand dieser Daten wird anschließend die Prüfgröße

$$\xi^2 = \sum_{\mathbf{x}} \frac{(o_{\mathbf{x}} - e_{\mathbf{x}})^2}{e_{\mathbf{x}}} \quad (20)$$

berechnet, wobei  $o_{\mathbf{x}}$  bzw.  $e_{\mathbf{x}}$  die Anzahl beobachteter bzw. erwarteter Antwortpattern  $\mathbf{x}$  bezeichnet. Die  $\xi^2$ -Werte folgen einer  $\chi^2$ -Verteilung, sodass der  $\xi^2$ -Wert der „echten“ Daten auf Signifikanz geprüft werden kann.

### 7.1.5 Itemfit-Statistiken

Letztlich werden die empirischen ICCs betrachtet und die Itemfit-Werte bestimmt. In diesem Zusammenhang stellen die Infit- und Outfit-Werte hilfreiche Entscheidungs- und Gütekriterien für die Modellpassung der ICCs dar. Outfit-Werte berechnen sich durch das Aufsummieren der Quadrate standardisierter Residuen (= Differenz zwischen beobachtetem und erwartetem Wert einer Person  $i$  bei Item  $j$ ), während Infit-Werte eine mit der Modellvarianz gewichtete Quadratsumme der Residuen darstellen (Bond & Fox, 2007). Die MSQ-Werte dieser beiden Parameter liegen idealerweise zwischen 0.5 und 1.5 (Linacre, 2003). Liegen beide Werte unter 0.8, liegt ein Overfit vor, d.h. das Item ist zu deterministisch (zu hohes Maß an Vorhersagbarkeit, zu wenig Variation in den Daten). Liegen beide Werte über 1.2, liegt ein Underfit vor, d.h. die Daten weisen zu viel Variation auf. Ein Mixed-Fit liegt vor, wenn nur einer der beiden Werte außerhalb des Intervalls  $[0.8, 1.2]$  liegt.

## 7.2 Rasch-Analyse des Leistungstests KiRC

Nachdem die notwendigen methodischen Grundlagen des Rasch-Modells erläutert wurden, werden in diesem Abschnitt die empirischen Daten des KiRC-Tests an das Rasch-Modell angepasst. Dabei ist die zentrale Fragestellung, ob sich das Rasch-Modell für eine hinreichend große Itemgruppe als gültig erweist oder nicht. Es folgt zunächst eine Planung der Testanalysen, in der die Kriterien zur dichotomen Aufteilung der Personen- und Itemgruppen festgelegt, Ablehnungsniveaus  $\alpha$  bestimmt und das Vorgehen im Falle lokaler Modellverletzungen beschrieben werden. Es schließen sich die Ergebnisse der Modellprüfung an.

### 7.2.1 Spezifikationen der Testanalysen

Die Analyse basiert auf der Eich-Stichprobe (vgl. Kap. 6.2) und beginnt mit den 22 Items der optimierten Kern-Itemgruppe (vgl. Kap. 6.4.4). Damit sind aus der Itemschnittmenge beider Hauptstudien bereits diejenigen Items aus der Analyse ausgeschlossen, die sich nach der KTT als zu einfach oder als nicht testkonform herausgestellt haben. Mit dieser initialen Itemgruppe startet die erste Iteration der Rasch-Analyse mit einem Datensatz von 22 Items, die von je 165 Personen beantwortet wurden.

#### Teilungskriterien

Die in Kap. 7.1.3 genannten Modell-Geltungstests beruhen auf Teilungskriterien des Datensatzes, die a posteriori festzulegen sind. Die Auswahl externer Teilungskriterien wird durch die erhobenen Stichprobendaten eingeschränkt und erfolgt nach zwei Kriterien: Erstens sollen die beiden resultierenden Gruppen annähernd gleich groß sein und zweitens sollen Modellverletzungen durch das gewählte Kriterium plausibel erklärbar sein. Für die Teilung der Personengruppen bieten sich folgende sieben Kriterien an, wobei die Begründung für die Kriterienwahl in Klammern gegeben wird: 1. Geschlecht (offensichtlich), 2. Teilnahme an Hauptstudie 1 oder 2 (sprachliche Unterschiede der Items, verschiedene Dozenten), 3.

Vorleistungen<sup>8</sup> (Kontrolle verschiedener Problemlösungsstrategien; gilt auch für 4.–6.), 4. Besuch eines Physik- bzw. 5. Mathematik-Leistungskurses, 6. Studienfachwahl und 7. Teilung am Median als internes Teilungskriterium (Itemhomogenität; sog. „Median-Split“).

Für die Teilung der Itemgruppen wurden sechs Kriterien gewählt: 1. Itemformat (unterschiedliche Ratewahrscheinlichkeiten), 2. Anzahl verwendeter Repräsentationen (schwierigkeitserzeugende Elemente, zusammen mit den Folgenden), 3.–5. Item enthält graphische, bildhafte oder algebraisch-formale Elemente und 6. Median-Split (Homogenität). Wie sich die Items gemäß den genannten Kriterien zu den Teilgruppen zusammensetzen, kann aus Tabelle 4 erschlossen werden. Beispielsweise unterteilt das Kriterium „Bildhafte Items“ den Itempool in Items mit und ohne Kennung „B“ in der Spalte RF. Auch das Skript im Anhang enthält die Einteilung der Items in die jeweiligen Gruppen (S. 214).

### Signifikanzniveaus

Da an einem Datensatz mehrere Tests durchgeführt werden, muss die Inflation des  $\alpha$ -Fehlers kontrolliert werden, indem das lokale Signifikanzniveau ( $\alpha = 0.05$ ) nach unten korrigiert wird. Wird der konservativsten Korrekturmethode gefolgt und das lokale Signifikanzniveau durch die Anzahl der durchgeführten Tests geteilt (Bonferroni-Korrektur), ergibt sich eine sehr kleine untere Schranke und die Modellgeltungstests sind nicht mehr sensitiv genug gegenüber Verletzungen der Modellannahmen. Dies bedeutet eine erhöhte Gefahr für einen Fehler 2. Art, die Modellgültigkeit fälschlicherweise anzunehmen. Koller et al. (2012) erarbeiteten einen Kompromiss zwischen gar keiner und der zu konservativen Bonferroni-Korrektur, indem sie einzelne Tests als eigenständig betrachten (z.B. den Andersen LR-Test), innerhalb denen die Korrektur erfolgt. Die Autoren begründen das Vorgehen damit, dass die Teilung stets einen neuen Datensatz erzeugt. Die angepassten Signifikanzniveaus werden bei den Ergebnisberichten angegeben.

### Vorgehen bei Modellverletzung

Bei lokalen Modellverletzungen einzelner Items wird eine zweite Iteration der Modellgeltungstests unter Ausschluss der auffälligen Items durchgeführt. Streng genommen wirkt sich schon das Entfernen eines einzigen Items auf alle Modellparameter und Prüfgrößen aus, sodass auch die Reihenfolge des Itemausschlusses Auswirkungen auf die Analyseergebnisse nehmen kann. Um den Aufwand überschaubar zu halten, werden problematische Items gruppenweise ausgeschlossen, wenn die Ursache der statistischen Auffälligkeit auch inhaltlich plausibel ist. Die vorangegangene Optimierung der Itemgruppe durch die Klassische Testanalyse erbrachte einen hinreichend großen Itempool, sodass durch eventuelles Streichen einiger nicht Rasch-konformer Items die inhaltliche Validität der Skala nicht gefährdet ist.

### Software, Ergebnisberichte und Darstellung

Die Analysen der Invarianzeigenschaften, die Bestimmung der Infit- und Outfit-Werte, sowie die Berechnung der ICCs wurden mit der kommando-basierten Open-Source-Programmiersprache R und der zugehörigen Bibliothek eRm durchgeführt (Mair & Hatzinger, 2007). Das erstellte Skript zur Steuerung des Programms befindet sich im Anhang (Kap. 12.5). Für die globalen Modellgeltungstests und das Bootstrapping wurde das Programm WINMIRA genutzt (von Davier, 2001). Im digitalen Anhang befinden sich die z.T. sehr umfangreichen Ergebnisberichte der Rasch-Analyse, die im hier Abgedruckten zusammengefasst sind.

<sup>8</sup>Aus den durchschnittlichen Physik- und Mathematiknoten aus der Schulzeit sowie aus der Abiturnote wurde eine Vorleistungskennzahl berechnet, die anschließend am Mittelwert dichotomisiert wurde, um zwei Leistungsgruppen zu erhalten.

## 7.2.2 Ergebnisse I: Invarianzeigenschaften

### Erste Iteration

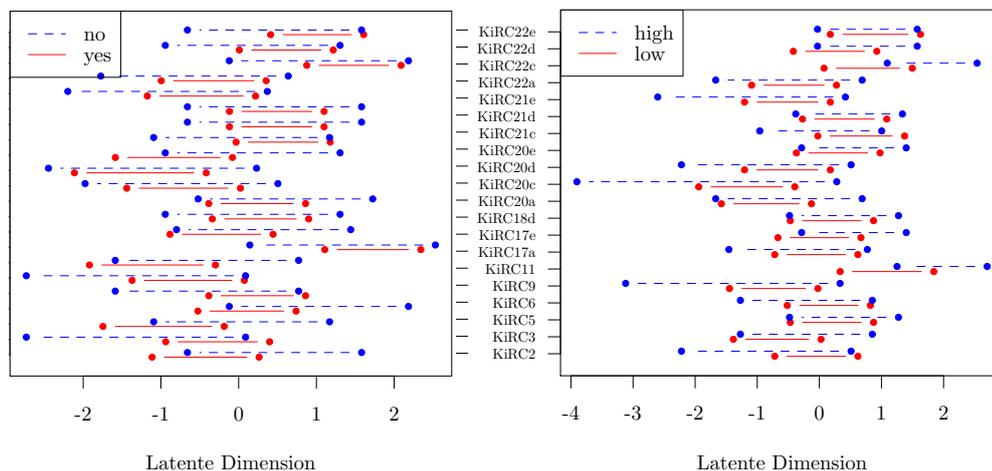
Ausgehend von den 22 Items der optimierten Kern-Itemgruppe wurde zunächst die Invarianz der Itemparameter innerhalb der Personengruppen überprüft. Der Datensatz wurde für jeden der sieben Teilungskriterien nacheinander in zwei Personengruppen aufgeteilt, die Parameterschätzung für die Gruppen getrennt durchgeführt und die Resultate anschließend mit dem Andersen LR-Test verglichen. Der Modellgeltungstest wurde für kein Kriterium auf dem gewählten  $\alpha$ -Niveau von 0.007 (= lokales  $\alpha/7$ ) signifikant, siehe Tab. 23 (linke Hälfte).

**Tab. 23:** Ergebnisse des Andersen LR-Tests auf Subgruppeninvarianz

Teilungskriterium	Erste Iteration <sup>a</sup>		Zweite Iteration <sup>b</sup>	
	Testwert	<i>p</i> -Wert	Testwert	<i>p</i> -Wert
Geschlecht	27.70	.149	24.54	.105
Mathe-LK	38.15	.012	31.36	.018
Physik-LK	12.07	.938	8.91	.943
Vorleistung	21.78	.412	17.86	.398
Semester	19.46	.556	13.22	.722
Hauptfach	26.03	.205	21.63	.199
Median	32.14	.030	33.00	.011

<sup>a</sup>22 Items,  $df = 21$ ,  $\alpha = 0.007$ . <sup>b</sup>Ausschluss von K6, K10, K21d, K22e;  $df = 17$ .

Da die Teilung nach Mathematik-Leistungskurs ebenso wie die Teilung am Median zu auffällig geringen  $p$ -Werten führte, wurden alle Items hinsichtlich diesen Kriterien noch einzeln mithilfe des Wald-Tests überprüft. Wie die im Anhang dargestellten Ergebnisse des Wald-Tests zeigen, unterscheiden sich die Parameterwerte der einzelnen Items innerhalb der beiden Gruppen nicht signifikant voneinander (Tab. A-5, S. 212). Graphisch kann dieses Resultat auch anhand des *DIFplots* nachvollzogen werden (Abb. 20). Der *DIFplot* stellt die Konfidenzintervalle der geschätzten Schwierigkeitsparameter innerhalb der beiden Gruppen pro Item übereinander dar. Diese überschneiden sich bei jedem Item.



**Abb. 20:** (*DIFplot*) Konfidenzintervalle der getrennten Itemparameterschätzung in zwei Subgruppen nach Teilung des Datensatzes gemäß „Mathe-LK“ (links) und „Median“ (rechts). Bemerkung: Im linken Teilbild erfolgt der Vergleich der Schätzparameter zwischen Studenten, die einen Mathematik-Leistungskurs besuchten („yes“) und die, die keinen besuchten. Rechts werden die Parameterschätzungen nach Aufteilung der Stichprobe am „Median“ zwischen oberer („high“) und unterer Hälfte verglichen.

Im nächsten Schritt wurde die Invarianz der Schätzwerte innerhalb verschiedener Itemgruppen mit dem Martin-Löf-Test überprüft. Wie Tabelle 24 (linke Hälfte) zeigt, liegen die  $p$ -Werte für jedes Teilungskriterium weit von dem Signifikanzniveau entfernt, sodass die Items als homogen angenommen werden können.

**Tab. 24:** Ergebnisse des Martin-Löf-Tests auf Itemhomogenität

Teilungskriterium	Erste Iteration <sup>a</sup>			Zweite Iteration <sup>b</sup>		
	LR	$df$	$p$ -Wert	LR	$df$	$p$ -Wert
Itemformat	89.07	111	.938	61.73	71	.776
Anzahl Repräsentationen	116.57	116	.468	78.38	79	.499
Graphische Items	100.76	116	.842	78.06	76	.413
Bildhafte Items	115.48	119	.574	61.73	71	.068
Formale Items	93.22	119	.961	93.22	119	.776
Itemschwierigkeit (Median)	85.05	120	.993	62.32	80	.928

*Bemerkungen:* LR = Likelihood Ratio; Signifikanzniveau  $\alpha = 0.008$ .

<sup>a</sup>Likelihood des gesamten Modells gemäß Gl. (18):  $L \approx 1480$ . <sup>b</sup> $L \approx 1178$ .

Um die Items auf lokale stochastische Unabhängigkeit (LSU) zu prüfen, wurde zunächst die tetrachorische Korrelationsmatrix berechnet, die in Anbetracht der Größe ausschließlich im digitalen Anhang dargestellt ist. Aus dieser Matrix wurden 29 Itempaare  $(j, l)$  mit Korrelationen  $r_{jl} \geq 0.5$  identifiziert. Zwar ist zu erwarten, dass die Items untereinander korrelieren (Eindimensionalität), aber zu hohe Item-Item-Korrelationen geben einen Anhaltspunkt auf Verletzungen der LSU. Eine rein explorative Vorgehensweise, d.h. das Prüfen aller möglichen Itempaare auf LSU, würde aufgrund der hohen Anzahl durchzuführender Einzeltests  $((22 \cdot 22 - 22)/2 = 231$  Tests) zu einem extrem niedrigen  $\alpha$ -Niveau führen, sodass die Gefahr für einen Fehler zweiter Art steigt. Ausgehend von den vorselektierten 29 Itempaaren wurde der Datensatz wie in Kap. 7.1.3 beschrieben schrittweise an Item  $j$  geteilt und die Parameter des Items  $l$  anschließend für die Gruppen „ $j$  falsch gelöst“ und „ $j$  richtig gelöst“ getrennt berechnet. Die  $z$ -Statistiken und  $p$ -Werte des Waldtests auf Itemebene für Item  $l$  sind in Tab. 25 (linke Hälfte) zusammengefasst. Zur Signifikanzbestimmung wurde die Benjamini-Hochberg-Korrektur angewandt. Der Wald-Test wurde für acht Itempaare signifikant, die anschließend nach inhaltlichen Gründen auf Verletzung der LSU untersucht wurden. In der Tat ist die Verletzung der LSU bei den Items K21c und K21d inhaltlich erklärbar: Das Item K21 stellt einen Beschleunigung-Zeit-Graphen dar und die Aussage K21c lautet „Der Körper erreicht seine maximale Beschleunigung zum Zeitpunkt  $t_1$ “ (vgl. Testbogen im Anhang, ab S. 205). In Item K21d wird nach der Richtungsumkehr während der Bewegung gefragt. Wenn aber erkannt wird, dass der Körper die Maximalgeschwindigkeit nicht zum Zeitpunkt  $t_1$ , sondern zum Zeitpunkt  $t_2$  erreicht, dann ist es naheliegend, dass die Aussage in K21d falsch sein muss. Auch Item K21e („Der Körper befindet sich zum Zeitpunkt  $t_2$  in Ruhe“) stellt sich in diesem Zusammenhang als problematisch dar, da eine logische Abhängigkeit zu Item K21d gegeben ist. Folglich wird Item K21d für die nächste Iteration eliminiert. In der Itemgruppe K22 korrelieren alle Aussagen hoch miteinander, denn sie erfordern allesamt eine komponentenweise Zerlegung der vektoriellen Bewegungsgrößen. Die höchsten Korrelationen weisen die Items K22a und K22e auf ( $r_{jl} = 0.69$ ). Diese Items bauen zwar nicht logisch aufeinander auf, aber es ist plausibel, dass eine Zerlegung des Geschwindigkeitsvektors (K22a) dann gelingt, wenn auch der Ortsvektor zerlegt werden kann (K22e) und umgekehrt. Ähnliche Argumente erklären die Verletzung der LSU bei den Itempaaren 6 und 9 sowie 5 und 10. Die Items beider Paare nutzen jeweils den gleichen Inhalt, die gleiche Repräsentationsform und ähneln sich stark im Wortlaut (vgl. Tab. 4 und Fragebogen). Da die Items K6 und K10 auch hohe Korrelationen zu einigen anderen Items aufweisen, werden sie ebenso wie Item K22e für die nächste Iteration entfernt.

**Zweite Iteration**

Die Analyseschritte wurden mit den 18 verbleibenden KiRC-Items wiederholt, nachdem die Items K6, K10, K21d und K22e aus dem Datensatz eliminiert wurden. Auch die geschätzten Parameter der reduzierten Itemgruppe weisen keine Verstöße gegen die Invarianzeigenschaften auf, wenn obige Teilungskriterien angewandt werden (Tabellen 23 und 24). Zur Prüfung der LSU wurde wiederum eine Vorselektion anhand der tetrachorischen Korrelationen vorgenommen. So wurden 12 Itempaare mit auffällig hohen Korrelationen identifiziert, die nach den Ergebnissen der Wald-Tests aber als lokal stochastisch unabhängig betrachtet werden können, vgl. rechte Hälfte in Tab. 25. Die verbleibende Itemgruppe gilt somit hinsichtlich der Invarianzeigenschaften als Rasch-konform und wird weiter analysiert (*Rasch-Itemgruppe*).

**Tab. 25:** Prüfung auf lokale stochastische Unabhängigkeit

Item $j$	Item $l$	$r_{jl}$	Erste Iteration		Zweite Iteration	
			$z$	$p$ -Wert ( $\alpha$ ) <sup>a</sup>	$z$	$p$ -Wert ( $\alpha$ ) <sup>a</sup>
K2	K4	.51	1.30	.194 (.043)	1.51	.132 (.013)
	K6	.53	1.38	.167 (.040)	-	-
	K10	.71	3.14	.002 (.005)*	-	-
	K17a	.60	2.10	.036 (.019)	2.30	.021 (.006)
K3	K21c	.51	2.79	.005 (.012)*	2.89	.004 (.004)
K4	K10	.53	1.77	.077 (.029)	-	-
	K17a	.52	1.56	.120 (.038)	1.57	.115 (.010)
	K21c	.52	1.35	.176 (.041)	1.39	.166 (.025)
K5	K10	.62	2.93	.003 (.009)*	-	-
K6	K9	.67	2.89	.004 (.010)*	-	-
	K10	.59	2.07	.039 (.021)	-	-
	K17a	.66	3.00	.003 (.007)*	-	-
	K22e	.55	1.72	.085 (.031)	-	-
K9	K10	.52	1.60	.119 (.036)	-	-
	K17a	.59	2.26	.024 (.017)	2.01	.005 (.004)
K10	K17a	.59	1.79	.073 (.028)	-	-
K17a	K21c	.51	1.07	.285 (.048)	1.32	.188 (.050)
	K21e	.53	1.22	.221 (.047)	1.46	.145 (.017)
K20c	K20d	.50	1.23	.219 (.045)	-	-
	K21c	.57	1.68	.092 (.033)	1.87	.062 (.008)
	K22a	.54	0.21	.835 (.050)	-	-
K20d	K21e	.60	2.46	.016 (.016)	2.43	.015 (.005)
K21c	K21d	.75	3.73	.000 (.002)*	-	-
	K21e	.60	1.92	.054 (.022)	1.94	.052 (.006)
K21d	K21e	.60	2.61	.009 (.014)*	-	-
	K22a	.51	1.92	.055 (.024)	-	-
K22a	K22c	.50	1.61	.108 (.034)	1.94	.052 (.007)
	K22d	.51	1.90	.058 (.026)	2.33	.020 (.005)
	K22e	.69	3.37	.001 (.003)*	-	-

*Bemerkungen:*  $z$ -Werte wurden durch Wald-Tests ermittelt. Ausschluss der Items K6, K10, K21d, K22e führt zu fehlenden Einträgen (-) für die zweite Iteration.

<sup>a</sup>In Klammern ist das Signifikanzniveau nach Benjamini-Hochberg angegeben.

\* $p < \alpha$  (lokal).

### 7.2.3 Ergebnisse II: Globale Modellpassung

Um die globale Modellpassung der 18 Items zu testen, wurden die Likelihoods auf Grundlage der geschätzten Parameter und des saturierten Modells berechnet. Die Testung der bei Modellgeltung zu erwartenden Antwortpattern am saturierten Modell nach Gl. (19) führt zwar zur Annahme des Modells ( $\chi^2 = 1612$ ,  $df \approx 2.6 \cdot 10^5$ ,  $p > 0.05$ ), aber streng genommen sind die Voraussetzungen zur Interpretation dieses Ergebnisses nicht gegeben (zu geringe Anzahl beobachteter Pattern, d.h. zu viele freie Parameter, vgl. Methodenkapitel 7.1.4). Unter Ausnutzung eines parametrischen Bootstraps (durchgeführt mit WINMIRA) wurden 400 Datensätze simuliert, die per Definition dem Rasch-Modell genügen. Die festgelegten Modellparameter dieser simulierten Daten entsprechen den geschätzten Werten aus der empirischen Datenmatrix. Die  $\xi^2$ -Werte der 400 Datensätze wurden gemäß Gl. (20) ermittelt und sind in Abb. 21 in Form eines Histogramms dargestellt. Aus dieser Abbildung wird deutlich, dass der  $\xi^2$ -Wert der „echten“ empirischen Daten ( $\xi^2 = 3.8 \cdot 10^5$ ) innerhalb des 5% Quantils aller simulierten  $\xi^2$ -Werte liegt, wodurch die globale Modellgültigkeit abgesichert ist.

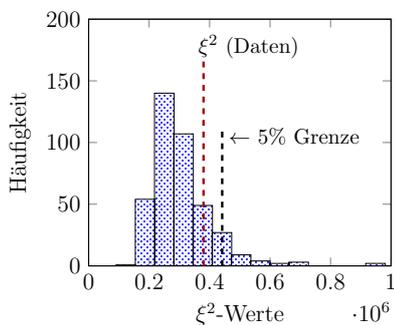


Abb. 21: Verteilung der  $\xi^2$ -Werte nach Simulation 400 Rasch-konformer Datensätze. Der empirische  $\xi^2$ -Wert des „echten“ Datensatzes fällt in das 5%-Quantil der  $\chi^2$ -Verteilung.

### 7.2.4 Ergebnisse III: Itemparameter, Item-Fit und ICCs

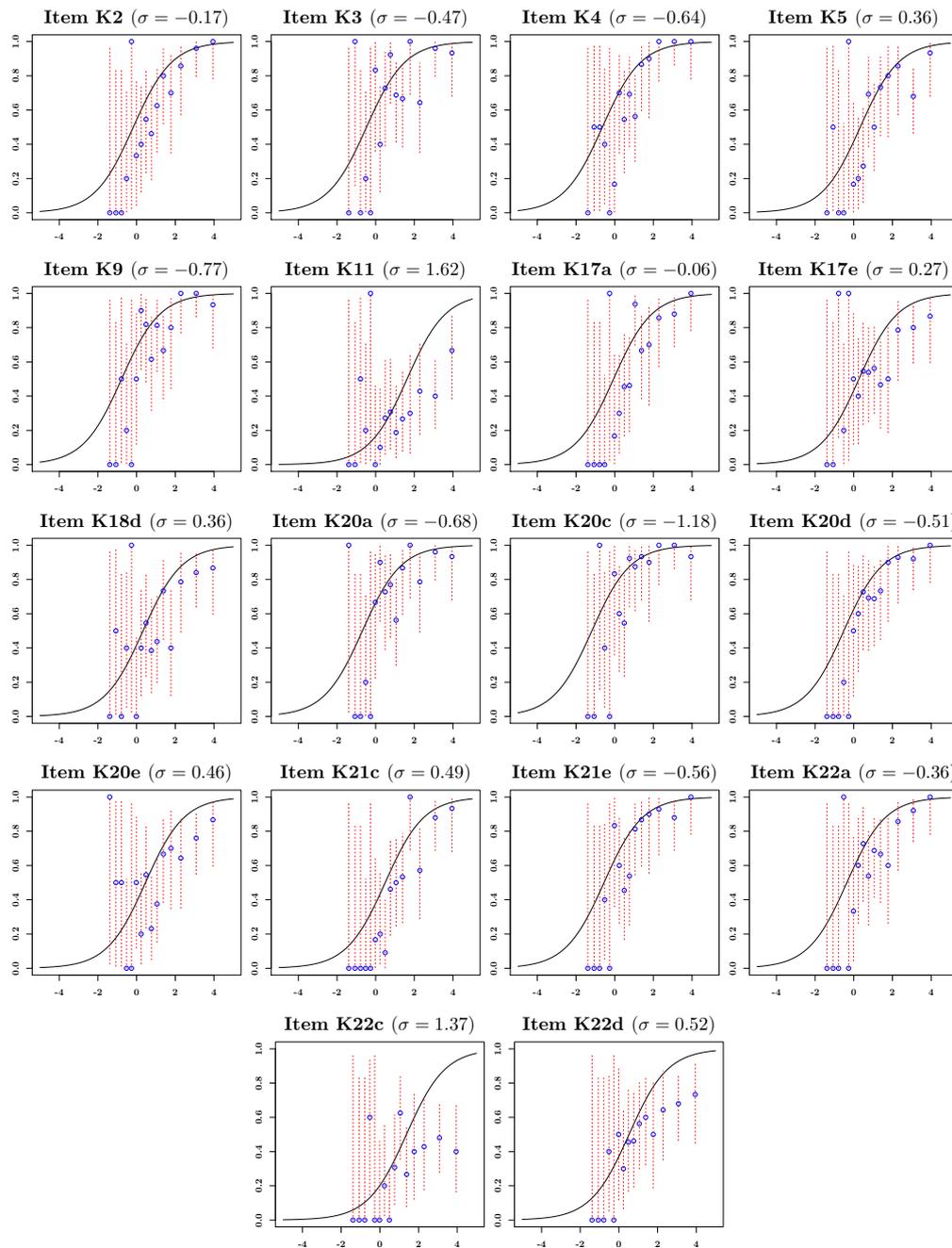
Zuletzt werden die Infit- und Outfit-Statistiken der Items zusammen mit den Aufgabencharakteristischen Kurven (ICCs) betrachtet. Wie aus Tabelle 26 hervorgeht, liegen sowohl die Infit- als auch Outfit-Statistiken der 18 Items allesamt im Zielbereich zwischen 0.5 und 1.5. Zwar sind die Items demnach prinzipiell Rasch-konform, doch weisen die Outfit-Werte einiger Items Tendenzen zum Overfit bzw. Underfit auf (Kennwerte außerhalb des Intervalls [0.8, 1.2]). Bei den Items K3, K11 und K17 liegt der Outfit-MSQ jeweils über 1.2, was auf einen

Tab. 26: Itemparameter  $\sigma$  und Fitstatistiken der Rasch-Itemgruppe

Item	$\sigma$	Outfit	Infit	Fit-Typ <sup>a</sup>	Item	$\sigma$	Outfit	Infit	Fit-Typ <sup>a</sup>
K2	-.17	.72	.87	(-)	K20a	-.68	1.19	1.00	ok
K3	-.47	1.39	1.08	(+)	K20c	-1.18	.76	.87	(-)
K4	-.64	.65	.88	(-)	K20d	-.51	.75	.88	(-)
K5	.36	1.11	1.00	ok	K20e	.46	1.09	1.05	ok
K9	-.77	.80	.91	ok	K21c	.49	.74	.82	(-)
K11	1.62	1.27	1.12	(+)	K21e	-.56	.80	.91	ok
K17a	-.06	0.74	.84	(-)	K22a	-.36	.89	1.03	ok
K17e	.27	1.26	1.16	(+)	K22c	1.37	1.17	1.13	ok
K18d	.36	1.07	1.04	ok	K22d	.52	1.09	1.12	ok

<sup>a</sup>Angaben in Klammern indizieren einen Mixed-Fit mit Tendenz zum Under- (+) bzw. Overfit (-).

Underfit deutet, d.h. es liegt mehr Varianz in den Daten, als es das Modell vorhersagt. Die ICCs dieser Items (Abb. 22) machen deutlich, dass Personen im mittleren Fähigkeitsbereich erwartungswidrig hohe Lösungswahrscheinlichkeiten besitzen, d.h. die Items besitzen eine geringe Trennschärfe. In der Tat liegen die klassisch berechneten Diskriminationsindizes dieser Items mit 0.32, 0.44 und 0.39 unterhalb des Skalenmittelwerts  $D = 0.52$  (vgl. Tabelle 15 sowie Tab. 14). Die Items K2, K4, K17a, K20c, K20d und K21c weisen hingegen Outfit-Werte unterhalb 0.8 auf, zeigen also eine Tendenz zum Overfit mit zu wenig Variation in den Daten. Durch Betrachtung der ICCs wird deutlich, dass die Items K17a und K21c

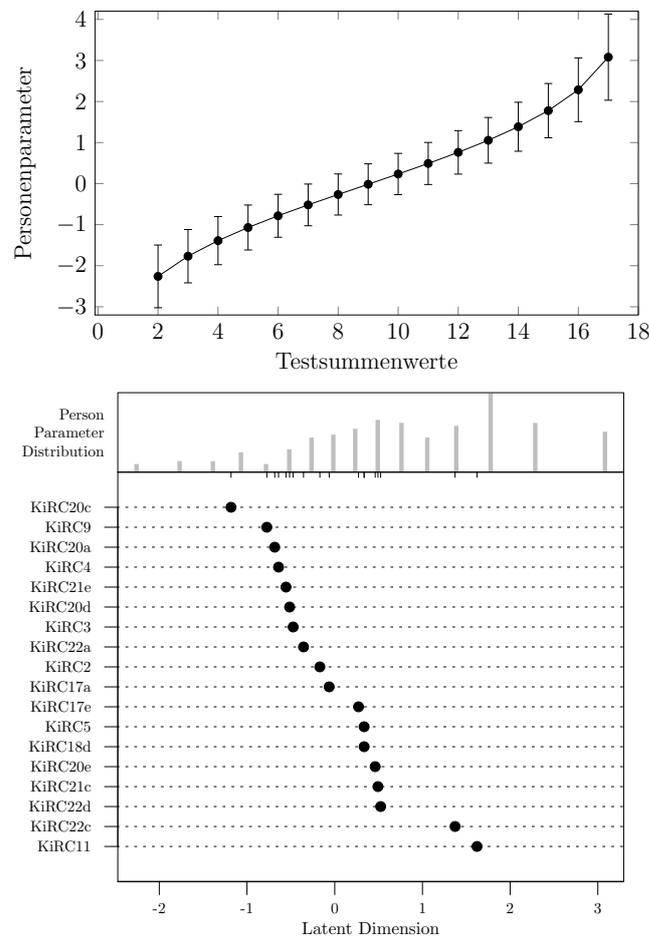


**Abb. 22:** Aufgabencharakteristische Kurven  $\mathcal{P}_\sigma(\eta)$  (ICCs) der Rasch-Itemgruppe des KiRC-Instruments mit empirischen Daten ( $\circ$ ) und Konfidenzintervallen ( $\cdots$ ).

in den extremen Fähigkeitsbereichen fast deterministisches Antwortverhalten zeigen, was durch die sehr hohen klassischen Diskriminationswerte (0.73 und 0.76) unterstrichen wird (vgl. Tab 14). Bei den Items K20c und K20d weichen vor allem die empirisch beobachteten Lösungswahrscheinlichkeiten im unteren bis mittleren Fähigkeitsbereich von der Theoriekurve ab. Die restlichen Items zeigen einen guten Modellfit und eine gute Übereinstimmung der empirischen Daten mit den theoretischen ICCs.

### 7.2.5 Ergebnisse IV: Graphische Darstellung der Item- und Personenparameter

Abschließend werden die Verteilungen der Personen- und Itemparameter betrachtet. Dies stellt kein Prüfkriterium der Modellgültigkeit dar, sondern dient dazu einzuschätzen, wie gut die Items auf die Personenstichprobe abgestimmt sind. Im unteren Teil der Abb. 23 sind die Items nach ihrem Schwierigkeitsindex auf dem latenten Kontinuum angeordnet. Durch die Striche am oberen Rand wird der direkte Bezug der Itemparameter zu den Personenparameter deutlich; bei Drehung um  $90^\circ$  entspricht diese Darstellung einer Wright-Map: Im unteren bis mittleren Fähigkeitsbereich liegen die Items dicht beieinander und messen die Fähigkeiten dort besonders genau. Im oberen Fähigkeitsbereich stehen hingegen nur wenige Items zur Verfügung, obwohl sich dort viele Personen verorten.



**Abb. 23:** Zusammenhang zwischen Testsumme und Personenparameter (oben) sowie Verteilung der Personen- und Itemparameter auf einer gemeinsamen latenten Dimension (unten).

In der oberen Abbildung ist der Zusammenhang zwischen den erreichten Testsummenwerten und den Personenparametern dargestellt. Der Kurvenverlauf ist erwartungsgemäß nicht-linear (vgl. Rost, 1996, S. 127), aber es besteht eine hohe Korrelation zwischen beiden Größen ( $R^2 = 0.98$ ). Die Genauigkeit der Schätzung nimmt zu den Rändern des Fähigkeitskontinuums ab, da zur Schätzung der Parameter dieser Personen weniger Items zur Verfügung stehen.

### 7.3 Rasch-Analyse des Konzepttests

Zur Prüfung der Rasch-Konformität der Konzept-Items wurde analog zu den KiRC-Items vorgegangen. Da die Skala jedoch nur fünf Items beinhaltet, wurde davon abgesehen, eine bessere Modellanpassung durch Itemausschluss zu erzielen. Die zentrale Frage ist also, ob die bestehenden Items Rasch-konform sind oder nicht. Die in Kapitel 7.2.1 festgelegten Teilungskriterien der Stichprobe wurden für die Rasch-Analyse der Konzept-Items beibehalten. Da alle Konzept-Items vom gleichen Frageformat (MC) und vom gleichen Repräsentationstyp (bildhaft) sind (vgl. Tab. 4), wird nur das interne Teilungskriterium (Median-Split) für den Martin-Löf-Test verwendet. Im Fließtext werden nur die Ergebnisse benannt, während die ausführlichen Testberichte im Anhang zu finden sind (ab S. 212).

Die Prüfung der *Invarianzeigenschaften* ergibt eine Auffälligkeit hinsichtlich des Teilungskriteriums „Geschlecht“ ( $\chi^2 = 15.12$ ,  $df = 4$ ,  $p = 0.004$ ). Der Waldtest grenzt diese Verletzung der Invarianzeigenschaft auf Item Kon2 („Scheibe bekommt einen Stoß“) ein ( $z = 3.08$ ,  $p = 0.002$ ), vgl. Tab. A-6. Offenbar lösen Frauen das Item mit einer viel geringeren Wahrscheinlichkeit (klassischer Schwierigkeitsindex  $P = 0.26$ ) als Männer ( $P = 0.72$ ). Eine Kontextabhängigkeit kann vermutet werden: Das im Item abgebildete Szenario erinnert an Stöße, die in Sportarten vorkommen, die vorwiegend Männer ansprechen (z.B. Billard, Fußball, Eishockey, usw.). Der Martin-Löf-Test auf Itemhomogenität grenzt an Signifikanz ( $\chi^2 = 10.67$ ,  $df = 5$ ,  $p = 0.06$ ). Die tetrachorischen Korrelationen zwischen den Items fallen niedrig aus, was mit dem Ergebnis des Martin-Löf-Tests wenig homogener Items in Einklang steht. Die lokale stochastische Unabhängigkeit der Items Kon1 und Kon2 wurde geprüft ( $r_{jt} \geq 0.50$ ); eine Modellverletzung kann ausgeschlossen werden ( $z = 1.23$ ,  $p = 0.22$ ), vgl. Tab A-7.

Der *globale Modellgeltungstest* gegen das saturierte Modell fällt knapp signifikant aus ( $\chi^2 = 39.21$ ,  $df = 25$ ,  $p = 0.04$ ), und auch das Resultat des parametrischen Bootstrap-Verfahrens verortet den empirischen  $\xi^2$ -Wert knapp unterhalb des 5%-Quantils der Verteilung (siehe Abb. A-2).

Die *Itemfitstatistiken* liegen allesamt im Zielbereich zwischen 0.5 und 1.5 — einzig Item Kon1 weist eine Tendenz zum Overfit auf, was auf die niedrige Diskrimination im unteren Fähigkeitsbereich zurückzuführen ist. Die Theoriekurven der ICCs spiegeln die empirischen Daten insgesamt sehr gut wieder (vgl. Abb. A-3). Die Abstände der Itemparameter auf der Logit-Skala sind nahezu äquidistant, streuen aber zu wenig, um das Fähigkeitspektrum hinreichend abzubilden.

### 7.4 Diskussion der Ergebnisse und Zusammenfassung

Die Analyse der KiRC-Items auf Rasch-Skalierbarkeit schloss zunächst vier Items aufgrund von Verletzungen der lokalen stochastischen Unabhängigkeit aus. Durch den Vergleich zwischen getrennten und gemeinsamen Parameterschätzungen nach Aufteilung des Datensatzes gemäß verschiedenen Teilungskriterien konnte gezeigt werden, dass die verbleibenden 18 Items den geforderten Invarianzeigenschaften des Rasch-Modells genügen. Auch die globalen Modellgeltungstests zeigen keine Verletzung der Modellgültigkeit an. Da der  $\chi^2$ -Statistik auf Grundlage des Tests am saturierten Modell nicht vertraut werden kann, wurde dieser Befund durch einen parametrischen Bootstrap abgesichert. Die Itemfit-Indizes zeigen ebenfalls keine

Verletzungen der Modellanforderungen, weisen aber auf Tendenzen des Over- bzw. Underfits hin. Der Overfit der Items konnte auf die hohe Diskrimination der Items zurückgeführt werden und ist für die Qualität der Messung nicht von Bedeutung. Die Items mit Underfit weisen zu niedrige Diskriminationsindizes auf und sollten für die Verbesserung der Messeigenschaften in einer revidierten Fassung des Instruments überarbeitet werden. Streng genommen stehen die unterschiedlichen Diskriminationsindizes im Widerspruch zur Annahme parallel verlaufender ICCs, wie sie im Rasch-Modell gefordert werden. Die Differenz der Diskriminationsparameter zwischen den Items ist jedoch tolerierbar, sodass das Rasch-Modell nicht abgelehnt wird. Wenn eine größere Datenbasis vorhanden ist, bietet sich die Testung eines zweiparametrischen Item-Response-Modells an, um die Modellpassung mit dem Rasch-Modell zu vergleichen.

Wie durch die Gegenüberstellung der Item- und Personenparameter gezeigt wurde, decken die bisherigen Items vor allem den unteren und mittleren Fähigkeitsbereich ab, sodass ein weiterer Optimierungsschritt des Instruments in der Entwicklung schwieriger Items für den oberen Fähigkeitsbereich besteht.

Zusammenfassend sprechen die Ergebnisse der Rasch-Analyse für eine Modellpassung und weisen auf Optimierungsmöglichkeiten für die Weiterentwicklung des Tests hin. Die Testsummen der Rasch-Itemgruppe stellen somit valide Indikatoren der Leistungsfähigkeit dar und werden in späteren Wirksamkeitsanalysen neben der (klassisch) optimierten Kern-Itemgruppe eingesetzt. Auf eine Analyse einzelner Antwortmuster kann verzichtet werden, da aus der Rasch-Konformität folgt, dass die Testsummen die gesamte Information über die Fähigkeit einer Person tragen. Folglich ist es zur Beurteilung der Repräsentationskompetenz von Studenten unerheblich, *welche* Aufgaben richtig gelöst wurden. Wie gezeigt wurde, korrelieren die Testsummen mit den geschätzten Personenparametern sehr hoch ( $R^2 = 0.98$ ), sodass es für die Ergebnisse der Vergleichsanalysen keinen praxisrelevanten Unterschied macht, ob von einer Person die Testsumme  $T_i$  oder die Fähigkeit  $\eta_i$  betrachtet wird.

Die Ergebnisse der Konzepttest-Analyse warnen davor, die Gültigkeit des Rasch-Modells anzunehmen. Item Kon2 verletzt gegen die Annahme der Stichprobenunabhängigkeit. Da das Item für die Vergleichsanalyse aber inhaltlich wertvoll ist und die Entfernung aus dem Instrument die ohnehin kurze Skala auf vier Items verkürzen würde, wird davon abgesehen eine potentielle Modellverbesserung durch den Itemausschluss zu erzeugen. Die globalen Prüfkriterien sind unter Einschluss aller fünf Items an der Signifikanzschwelle zur Ablehnung des Modells. Entgegen diesen Befunden zeigen die empirischen Lösungswahrscheinlichkeiten eine hervorragende Passung an die aus dem Modell abgeleiteten theoretischen ICCs. Insgesamt wird die Skala des Konzepttests unverändert für die Vergleichsanalysen im Sinne einer zur Repräsentationskompetenz (KiRC) komplementären, ergänzenden Leistungsmessung beibehalten. Der Verletzung der Invarianzeigenschaft wird dadurch entgegnet, indem die Itemstatistiken von Kon2 bei der Untersuchung von Einflüssen des Geschlechts auf das Konzeptverständnis explizit berichtet werden (zusätzlich zu den Ergebnissen der Gesamtskala).

## Kapitel 8

# Analyse leistungsbezogener Antwortsicherheiten

In Kapitel 5.3 wurde dargestellt, dass zu manchen Items des Leitungstests Antwortsicherheiten erhoben wurden, um Rückschlüsse auf die metakognitiven Fähigkeiten der Studierenden zu ziehen. In diesem Kapitel werden zunächst die methodischen Grundlagen zur Analyse von diesen Sicherheitsangaben im Zusammenhang mit der gegebenen physikalisch-inhaltlichen Antwort dargestellt (Kap. 8.1). Mit Blick auf die in Kap. 5.3 formulierte Zielsetzung, die mit der Erhebung von Antwortsicherheiten im Rahmen dieser Arbeit verfolgt wurde, werden in Kap. 8.2 Fragestellungen formuliert, die auf Basis des Datensatzes der vorliegenden Studie beantwortet werden.

Die Analyse-Ergebnisse sind bereits veröffentlicht (Klein et al., 2016), weshalb sich die Ergebnisdarstellung in Kap. 8.3 auf eine Auswahl der wichtigsten Resultate beschränkt. Darüber hinaus werden die eigenen Daten in die Ergebnisse bisheriger themenverwandter Arbeiten eingeordnet.

### 8.1 Methodische Grundlagen

Im Folgenden wird mit  $S_j$  die Antwortsicherheit und mit  $P_j$  (wie gewohnt) der Item-Schwierigkeitsindex einer Aufgabe  $j$  bezeichnet. Da es sich um Item-Indizes handelt, erfolgt die Bildung des Mittelwerts über alle Personen. Kenngrößen ohne Index beziehen sich auf Testmittelwerte.

Zur Analyse fragebezogener Antwortsicherheiten wurden in der Naturwissenschaftsdidaktik verschiedene methodische Ansätze gewählt, die im Wesentlichen auf die Berechnung von Korrelationen und anderen Kenngrößen zurückzuführen sind (Aslanides & Savage, 2013; Pieschl, 2008; Planinic et al., 2006; Potgieter et al., 2010).<sup>1</sup> Diese werden im Folgenden zusammengefasst und erläutert.

#### 8.1.1 Berechnung und Vergleich absoluter Größen

Einige Physikdidaktiker vertreten die Ansicht, dass ein tiefes Verständnis der Testinhalte erst dann vorliegt, wenn Lernende hohe Testleistungen ( $P$ ) erreichen und hohe Antwortsicherheiten ( $S$ ) angeben (Aslanides & Savage, 2013; Lindsey & Nagel, 2015). Die absoluten Ausprägungen der beiden Variablen werden isoliert voneinander betrachtet und die Veränderung nach einer Lernphase diagnostiziert (Aslanides & Savage, 2013; Sharma & Bewes, 2011; Struck & Yerrick,

<sup>1</sup>Für eine ausführliche Darstellung über die bisherige naturwissenschaftsdidaktische Forschung zu Antwortsicherheiten siehe Klein et al. (2016).

2010). Die beiden Größen werden durch den bias-Index (zu dt. Tendenz), definiert als

$$b_j = S_j - P_j, \quad (21)$$

unmittelbar in Bezug zueinander gesetzt. Überschätzungen der eigenen Leistung gehen mit positiven Werten ( $b_j > 0$ ), und Unterschätzungen mit negativen Werten ( $b_j < 0$ ) einher (vgl. z.B. Sharma & Bewes, 2011). Der bias-Index ist sowohl für jedes Item als auch für den Gesamttest definiert ( $b = S - P$ ). Kritisch betrachtet ist sowohl die Interpretation absoluter Größen als auch deren Verrechnung gemäß Gl. (21) unsachgemäß. Es lässt sich aus messtheoretischer Sicht keine Rechtfertigung für eine derartige Rechenoperation erkennen, wenn beide Größen zuvor nicht auf einer gemeinsamen Skala kalibriert wurden (z.B. mittels Rasch-Modellen). Ebenso ist eine Veränderung der Antwortsicherheit ohne die Information über die tatsächliche Leistung wenig informativ — insbesondere auf Ebene des Gesamttests. Auch wenn viele Forscher für die Verrechnung absoluter Größen argumentieren (Schraw, 1995), wird in dieser Arbeit von dieser Methode aus genannten Gründen Abstand genommen.

### 8.1.2 Zusammenhangsmaße auf Stichprobenebene

Ein übliches Maß zur Beurteilung der Genauigkeit der Selbsteinschätzung ist der *CAQ*-Index (*confidence-judgement accuracy quotient*), definiert durch

$$CAQ = \frac{S_r - S_f}{SD_p}. \quad (22)$$

Dabei bedeuten  $S_{r/f}$  die mittlere Antwortsicherheit bei richtigen bzw. falschen Antworten. Im Nenner steht die gepoolte Standardabweichung  $SD_p$  aus den Antwortsicherheiten für richtige bzw. falsche Lösungen, was der Tatsache Rechnung trägt, dass einzelne Personen die Sicherheitsskala unterschiedlich breit ausschöpfen. Der Index geht zurück auf Shaughnessy (1979), wird für jede Person individuell berechnet, ist  $t$ -verteilt und besitzt die formale Struktur einer Effektstärke. Er kann als Analogie zum Sensitivitätsmaß  $d'$  der Signalentdeckungstheorie betrachtet werden („Häufigkeit Treffer minus Häufigkeit falscher Alarm“, Velden, 1982) und wird genau dann Null, wenn keine adäquate Selbsteinschätzung erfolgt, d.h. wenn es bei richtigen und falschen Antworten keinen Sicherheitsunterschied gibt. Eine Stichprobe besitzt definitionsgemäß eine angemessene Selbsteinschätzung, wenn sich der mittlere *CAQ*-Index signifikant von Null unterscheidet.

In diesem Zusammenhang wird oft auch die Korrelation zwischen mittlerer Aufgabenschwierigkeit und Antwortsicherheit auf Stichprobenebene berichtet. Eine Population mit guter „metakognitiver Kalibrierung“ verspürt im Durchschnitt eine hohe Antwortsicherheit bei einfachen Aufgaben (hohe Lösungswahrscheinlichkeit) sowie eine niedrige Sicherheit bei schwierigen Aufgaben (niedrige Lösungswahrscheinlichkeit), vgl. Lindsey und Nagel (2015). Der Korrelationskoeffizient zwischen mittlerer Sicherheit und mittlerem Schwierigkeitsindex ist signifikant größer Null.

### 8.1.3 Zusammenhangsmaße auf Itemebene

Analysen auf Itemebene betreffen die Frage, ob sich die mittlere Sicherheit, die mit richtigen Antworten assoziiert wird, von der mittleren Sicherheit bei falschen Antworten signifikant unterscheidet. Dies ist äquivalent zur Berechnung von Korrelationen zwischen Sicherheit und Schwierigkeit auf der Ebene eines einzelnen Items (da Items nur entweder richtig oder falsch beantwortet werden können).

Während eine Analyse auf der Stichprobenebene eine Aussage über die metakognitive Kalibrierung einer Population bzgl. eines Lerninhalts im Allgemeinen trifft, dient die Analyse auf der Itemebene zur Identifikation von Lernschwierigkeiten und Fehlvorstellungen bzgl. konkreter Lerninhalte. In der Praxis wird angenommen, dass sich Fehlvorstellungen durch

niedrige Lösungswahrscheinlichkeiten bei vergleichsweise hohen Antwortsicherheiten manifestieren, auch wenn die Stichprobe im Allgemeinen (d.h. auf der Stichprobenebene) eine gute metakognitive Kalibrierung zeigt (Aslanides & Savage, 2013; Planinic et al., 2006; Potgieter et al., 2010). Items, die Fehlvorstellungen adressieren, heben sich von anderen Items also dadurch ab, dass zutreffende Selbsteinschätzungen seltener gelingen.

## 8.2 Fragestellungen und Vorgehen

Im Rahmen dieser Arbeit werden zwei Fragestellungen aufgeworfen, die mithilfe der zur Verfügung stehenden Daten aus dem ersten Testzeitpunkt beantwortet werden (Eich-Stichprobe).

1. Wie groß ist die Genauigkeit der Selbsteinschätzung (metakognitive Kalibrierung) bzgl. repräsentationsbezogenen bzw. konzeptorientierten Aufgaben auf Stichprobenebene und gibt es einen Zusammenhang zum Leistungsniveau/zum Geschlecht der Studenten?
2. Unterscheiden sich die Zusammenhänge zwischen Antwortsicherheit und Lösungswahrscheinlichkeit einzelner Items und lassen sich daraus Rückschlüsse auf vorliegende Fehlkonzepte oder Verständnisschwierigkeiten ziehen?

Die Ergebnisse zu diesen Fragestellungen wurden bereits veröffentlicht und diskutiert (Klein et al., 2016), sodass sich in den folgenden Kapiteln auf die wichtigsten Resultate bzgl. den KiRC-Items beschränkt wird; weitere Ergebnisse sowie die Analyse der Konzept-Items können in dem zitierten Artikel nachgeschlagen werden.

## 8.3 Ergebnisse

An die Resultate aus der klassischen Testkalibrierung (Kapitel 6.4.6) anknüpfend wurden die Antwortsicherheiten zu dem KiRC-Test und zu dem Konzept-Test getrennt voneinander untersucht, da sich die Sicherheitsskalen faktoranalytisch trennten. Als Datengrundlage dienten die Testergebnisse und Sicherheitsangaben der Eich-Stichprobe ( $N = 165$ ).

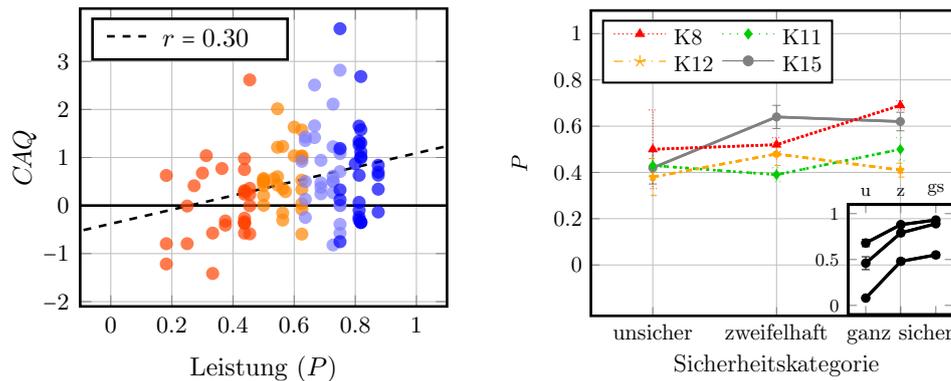
### 8.3.1 Ergebnisse auf Stichprobenebene

Der mittlere CAQ-Index nach Gl. (22) beträgt für den KiRC-Test  $CAQ = 0.52$  ( $SD = 0.08$ ). Der Wert unterscheidet sich signifikant von Null ( $t(111) = 6.37$ ,  $p < 0.001$ ). Innerhalb von Leistungsquartilen ergaben sich mit absteigender Leistung die Indizes 0.72, 0.74, 0.59 und 0.05, die sich mit Ausnahme des Letzten alle signifikant von Null unterscheiden. Leistungsschwache Studenten besitzen folglich keine adäquate Selbsteinschätzung. Der Zusammenhang zwischen Leistung und CAQ-Index wird in Abb. 24 (links) verdeutlicht und durch einen signifikanten Korrelationskoeffizient statistisch bestätigt ( $r(112) = 0.30$ ,  $p = 0.001$ ).

Es wurden keine signifikanten Unterschiede bezüglich den CAQ-Indizes zwischen Studierenden männlichen und weiblichen Geschlechts festgestellt ( $t(110) = 0.56$ ,  $p = 0.58$ ).

### 8.3.2 Ergebnisse auf Ebene einzelner Items

Durch Korrelationsanalysen wurden alle Items auf einen linearen Zusammenhang zwischen Lösungswahrscheinlichkeit und gegebener Antwortsicherheit getestet. Es wurden vier KiRC-Items identifiziert, die nur eine geringe, nicht signifikant von Null verschiedene Korrelation aufweisen ( $r < 0.11$ ,  $p > 0.10$ ): K8, K11, K12 und K15. Die Lösungswahrscheinlichkeiten dieser Items sind in Abb. 24 (rechts) kategorienweise dargestellt, wobei auf die  $S = 0$  Kategorie („geraten“) verzichtet wird. Es wird deutlich, dass höhere Sicherheiten nicht mit größeren Lösungswahrscheinlichkeiten verbunden werden. Obwohl die Studierenden insgesamt eine gute metakognitive Kalibrierung zeigen, schätzen sie ihre Leistung bei den genannten Items nicht adäquat ein.



**Abb. 24:** Links: Zusammenhang zwischen Leistung und Genauigkeit der Selbsteinschätzung ( $CAQ$ -Index). Jeder Datenpunkt entspricht einem Studenten, wobei die vier Leistungsquartile farblich unterschieden werden. Der  $CAQ$ -Index kann für extreme  $P$ -Werte nicht berechnet werden, da zur Berechnung der Standardabweichung in Gl. (22) mindestens zwei Sicherheitsangaben mit  $P_j = 0$  (bzw.  $P_j = 1$ ) vorliegen müssen. Rechts: Lösungswahrscheinlichkeiten  $P_j$  pro Sicherheitskategorie für Items ohne signifikante Korrelation zwischen Itemschwierigkeit und Antwortsicherheit. Zum Vergleich zeigt die eingebettete Grafik drei Items mit stark linearem Zusammenhang zwischen Schwierigkeit und Sicherheit (K1, K4, K9; von unten nach oben).

## 8.4 Diskussion und Interpretation der Ergebnisse

Zur Beantwortung der ersten Fragestellung wurde der  $CAQ$ -Index als Genauigkeitsmaß der Selbsteinschätzung auf Stichprobenebene bestimmt. Unabhängig vom Geschlecht beurteilten die Studenten ihre Leistung zutreffend ( $CAQ > 0$ ,  $p < 0.001$ ). Es wurde allerdings festgestellt, dass leistungsschwache Studenten ihre Fähigkeiten überschätzten. Die festgestellte Abhängigkeit der Selbsteinschätzung von der Fähigkeit ist in der Literatur als *Dunning-Kruger-Effekt* bekannt (Kruger & Dunning, 1999), und wurde auch in anderen Arbeiten aus der Physikdidaktik nachgewiesen (Lindsey & Nagel, 2015; Sharma & Bewes, 2011) — wenngleich nicht für repräsentationsbezogene Aufgaben. Wegen der großen Bedeutung der Repräsentationskompetenz für das Verstehen von Physik und der Sonderstellung der Kinematik innerhalb der Fachsystematik fallen diese Ergebnisse besonders ins Gewicht. Verfügen leistungsschwache Studenten nur über ein inadäquates Reflexionsvermögen der eigenen Fähigkeiten, besteht die Gefahr, Lerndefizite nicht aufholen zu können. Gerade in den ersten Vorlesungswochen, aus denen die Daten stammen, ist nicht auszuschließen, dass sich solche Defizite auf die Leistung auswirken. Aus dieser Überlegung wird die Forschungsfrage generiert, ob der  $CAQ$ -Index ein Prädiktor für die Leistungsentwicklung darstellt (vgl. Hypothesen der Hauptuntersuchung, Kap. 9). Perspektivisch könnten an dieser Stelle Förderungsmaßnahmen zur Steigerung der metakognitiven Kalibrierung ansetzen.

Die zweite Fragestellung wurde durch die Berechnung der Korrelationen zwischen Antwortsicherheit und Lösungswahrscheinlichkeit auf Itemebene beantwortet. Bei vier KiRC-Items bestand kein Zusammenhang zwischen Leistung und Antwortsicherheit. In Klein et al. (2016) wird begründet, dass inhaltliche Schwierigkeiten und existierende Fehlvorstellungen dazu führten, dass Studierende ihre Leistung bei diesen Items nicht adäquat einschätzen konnten. Die in Klein et al. (2016) geführte Diskussion wird an dieser Stelle exemplarisch für Item 11 aufgeführt. Dieses Item thematisiert den Verlauf der Beschleunigungsvektoren bei einer beschleunigten Kurvenfahrt und damit ein Konzept der Rotationskinematik. Ein häufig gewählter Distraktor zeigt gleich lange, ins Zentrum gerichtete Beschleunigungspfeile, was bei einer gleichmäßigen Kurvenfahrt mit konstanter Bahngeschwindigkeit korrekt wäre. Vermutlich erinnern sich manche Studierende an dieses bekannte Bild aus dem Physikunterricht

der Schule und bringen Kreisbewegungen generell in Verbindung mit einer ausschließlich radialen Beschleunigungskomponente, sodass sie ihrer (falschen) Antwort eine hohe Sicherheit zuschreiben. In einer Pilotierungsphase des KiRC-Instruments wurden die Studierenden gebeten, die von ihnen gewählte Antwortalternative schriftlich zu begründen. Beispielsweise argumentierten einige Studenten, dass eine konstante Länge der Beschleunigungspfeile zu einer gleichmäßigen Beschleunigung führt.<sup>2</sup> Dieser Begründung liegt ein Verständnis der eindimensionalen Kinematik zugrunde, die den Richtungscharakter außer Acht lässt. Andere Studenten argumentierten, dass die Geschwindigkeit (Winkelgeschwindigkeit) nur dann größer wird, wenn die Beschleunigung (Winkelbeschleunigung) größer wird.<sup>3</sup> Hinter beiden Vorstellungen stecken Fehlkonzepte, die ihren Ursprung in der linearen Kinematik haben (McDermott et al., 1987), und die sich auf die Rotationskinematik übertragen (Mashood & Singh, 2012). Diese Begründungen untermauern die Hypothese, dass sich hinter den hohen Sicherheiten bei falschen Antworten fehlerbehaftete Vorstellungen über den Lerninhalt verbergen, die nur schwer zu überwinden sind.

Aus der in Klein et al. (2016) geführten Diskussion lässt sich zusammenfassen, dass es (1) Studierenden trotz guter metakognitiver Kalibrierung innerhalb einer bekannten Domäne bei einzelnen Aufgaben nicht gelingt, zwischen richtigen und falschen Lösungen zu unterscheiden, (2) inadäquates Beurteilungsvermögen auf Lern- und Verständnisschwierigkeiten der Repräsentation und der zugrunde liegenden Konzepte zurückgeführt werden kann und sich somit (3) die Analyse eignet, Items mit Lernschwierigkeiten zu identifizieren.

Abschließend werden die gefundenen Zusammenhänge in die Ergebnisse der Literatur eingeordnet. Dazu sind in Tab. 27 die Korrelationen zwischen den Variablen Leistung ( $P$ ) und Antwortsicherheit ( $S$ ) aufgeführt. Die Berechnung der mittleren Korrelation zwischen diesen beiden Variablen über alle Studien (eigene ausgenommen) erfolgte nach Anwenden des Transformationsverfahrens nach Fisher (vgl. Leonhart, 2013) und ergibt den Wert  $\bar{r} = 0.45$ . Der Wert stimmt mit dem Ergebnis der vorliegenden Studie gut überein.

Zusammenfassend lässt sich festhalten, dass die Analyse von Antwortsicherheiten wichtige

<sup>2</sup>Zitat Student A: „[...] Der Betrag [muss] konstant sein, da  $|\vec{v}|$  in etwa konstant größer wird“. Die qualitativen Original-Daten befinden sich im digitalen Anhang.

<sup>3</sup>Zitat Student B: „[Die] Beschleunigung wird größer, da die Geschwindigkeitspfeile größer werden.“

**Tab. 27:** Studienübergreifender Vergleich von Zusammenhängen zwischen Leistungs- ( $P$ ) und Sicherheitsindikatoren ( $S$ )

Referenz	Inhaltsdomäne	Variablen	$N$	$r$
Dowd et al., 2015	Physik I (Uni)	$S, P$	89	0.39***
Aslanides & Savage, 2013	SRT (Uni)	$S, P$	70	0.11**
			63	0.17**
Lindsey & Nagel, 2015	Mech. (Uni)	$S, P$	58	0.51**
	Chemie I (Uni)	$S, P$	58	0.47**
Sharma & Bewes, 2011	Mech. (Sek. II)	$b, P$	140	-0.17
			227	-0.31**
			123	-0.48**
Potgieter et al., 2010	Mechanik (Uni)	$S, P$	33	0.56 <sup>k.A.</sup>
Struck & Yerrick, 2010	Mech. (Sek. II)	$S, P$	$n = 33^a$	0.85***
Caleon & Subramaniam, 2010	Mech. (Sek. II)	$S, P$	$n = 14^b$	0.40
		$P, CAQ$	$n = 14^b$	0.11
Diese Arbeit (Mechanik, Uni)	Repräsentationen	$S, P$	157	0.49***
		$P, CAQ$	112	0.30***
		$S, P$	124	0.42***

*Bemerkungen:* CAQ = Confidence-judgement Accuracy Quotient,  $b$  = Bias,  $n$  = Anzahl Items (Zusammenhangsmaße beziehen sich auf Item-Mittelwerte), SRT = Spezielle Relativitätstheorie.

<sup>a</sup> $N = 675$ , <sup>b</sup> $N = 243$  (Berechnung durch den Autor dieser Arbeit aus den Daten der Publikation).

\*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

Hinweise sowohl über die metakognitive Kalibrierung der Stichprobe als auch über den Charakter spezifischer Lerninhalte und Gegenstände der Testinstrumente liefert. Diese Aspekte sind vor dem Hintergrund der Physikdidaktik in zweierlei Hinsicht interessant: Erstens spielen metakognitive Fähigkeit eine große Rolle für den selbstgesteuerten Lernprozess, der insbesondere in der Hochschule von großer Bedeutung für den Lernerfolg ist (Gerholz, 2012). Die Diagnose der metakognitiven Kalibrierung ist ein notwendiger Schritt für gezielte Interventionsmaßnahmen. Zweitens hat die physikdidaktische Forschung in den letzten Jahrzehnten große Anstrengungen unternommen, Lernschwierigkeiten und Fehlkonzepte bei Lernenden zu identifizieren. Das Nutzen von Antwortsicherheiten scheint vielversprechend, einen ökonomischen methodischen Beitrag zu diesen Arbeiten leisten zu können.

## Teil III

# Hauptuntersuchung



## Kapitel 9

# Hypothesen und Forschungsfragen

In diesem Kapitel werden die Fragestellungen und deren untergeordneten Hypothesen (Hyp) sowie Forschungsfragen (FF) formuliert, die der Hauptuntersuchung zugrunde liegen. Die gerichteten Hypothesen 1.1 bis 1.4 (S. 128) resultieren aus den theoriegestützten Annahmen über den Einfluss des Instruktionsmaterials (*videobasierte Aufgaben*, d.h. V- und mV-Aufgaben) auf die Zielvariablen Leistung, Motivation und Neugierde. Diese Annahmen werden im Folgenden mit Bezug auf die Theorie (Kap. 3) repetiert:

- Zur Bearbeitung videobasierter Aufgaben setzen sich die Studierenden intensiv mit zwei unterschiedlichen Medien, nämlich textbasierten Instruktionen und Experimentalvideos, auseinander. Die dadurch erreichte Multikodierung des Lerninhalts, der den auditiven Kanal (schriftliche Instruktion) und den visuellen Kanal anspricht, fördert das Verständnis der physikalischen Inhalte nach Annahmen der CATLM besser als ausschließlich textbasierte Instruktionen.
- Durch die Verwendung multipler Repräsentationsformen ermöglichen videobasierte Aufgaben eine größere Verarbeitungstiefe und eine größere Vernetzungsdichte des Lerninhalts im Vergleich zu traditionellen Aufgaben. Entsprechend werden positive Effekte auf Konzeptverständnis und Repräsentationskompetenz erwartet.
- Videobasierte Instruktionsmaterialien besitzen einen hohen Grad an Authentizität (Realitätsbezug und Wissenschaftsbezug), denn sie bilden wahre Experimente ab und akzentuieren das Wechselspiel zwischen Theorie und Experiment. Dies wirkt sich positiv auf die Motivation der Studierenden aus.
- Studierende erheben eigene Messdaten, wodurch der Wissenschaftsbezug verstärkt wird. Sie erhalten experimentelle Rückmeldungen über theoretische Berechnungen, was motivationsfördernd wirkt (Steigerung des Selbstkonzepts).
- Mobile Videoanalyse-Aufgaben gewähren den Lernenden mehr Freiraum beim Problemlösen, wodurch sich wahrgenommene Autonomie, Neugierde und Interesse steigern lassen.

Unter Annahme einer positiven Wirkung videobasierter Aufgaben auf Repräsentationskompetenz und Konzeptverständnis können auch positive Effekte bzgl. der Klausur-Gesamtnote erwartet werden, da diese Variablen unmittelbar in Zusammenhang mit Problemlösefähigkeit stehen (siehe Kap. 3.1.3). Auf die Formulierung einer entsprechenden Hypothese wird verzichtet, da diese in der vorliegenden Untersuchung aufgrund des Studiendesigns nicht untersucht werden konnte (vgl. Kap. 10.2). Jedoch sei angemerkt, dass Gruppenunterschiede bzgl. einzelnen Klausuraufgaben untersucht werden, die sich inhaltlich zu verschiedenem

Instruktionsmaterial zuordnen lassen. Außerdem werden Zusammenhänge zwischen Klausur-Gesamtnote und anderen Leistungsindikatoren analysiert.

Die Gleichheitshypothesen 2.1 bis 2.3 beziehen sich auf die Absicht, belastungsähnliche Aufgabenpaare zu konstruieren (vgl. Kap. 10.3) und fußen auf dem Prinzip des fairen Vergleichs. Zwar reduziert (nach der Theorie) die Multikodalität des Lerninhalts und der verkürzte Aufgabentext die kognitive Belastung der Studierenden, doch muss untersucht werden, ob der Effekt nicht durch eine erhöhte Belastung durch den Umgang mit neuen Medien und dem Verwenden einer höheren Anzahl an Repräsentationen überkompensiert wird. Auch die Bearbeitungszeit stellt in diesem Zusammenhang ein wichtiger Indikator zur Abschätzung des Interventionserfolgs dar.

In ähnlichem Sinne wird mit dem „Manipulation Check“ (Hyp 3.1) überprüft, ob der im Studiendesign intendierte Unterschied zwischen den Aufgaben auch tatsächlich, d.h. statistisch signifikant und mit relevanter Effektstärke, durch die Lernenden wahrgenommen wird. Da die eingesetzten V und mV-Aufgaben eine Brücke zum Physikpraktikum herstellen, ist der Experimentbezug auch als abhängige Variable aufzufassen, die den Interventionserfolg quantifiziert.

Wie in Kapitel 5.1.3 erwähnt wurde, besteht bei der Analyse des Leistungstest eine Wahl-freiheit erstens in der Zusammenstellung der Itemgruppen und zweitens in der Möglichkeit zur Verrechnung von Teilsummen zu einer Gesamtsumme. Die methodischen Fragestellungen 3.2 – 3.4 greifen diesen Sachverhalt auf und entgegen dem pragmatischen Vorgehen einer festen Testsummenbildung (z.B. Gleichgewichtung beider Testkomponenten) mit einer parametrischen Komposition der Testkomponenten, wodurch die empirischen Ergebnisse durch Angabe von Effektstärkebereichen abgesichert werden. Die Forschungsfragen 4.1 bis 4.5 zielen auf die Untersuchung von stichprobenbedingten Einflussparametern der Resultate ab. Es wird untersucht, ob vorhandene Unterschiedseffekte von Persönlichkeitseigenschaften, Leistungsindikatoren oder Vorleistungen abhängen oder dagegen robust sind.

## 1. Hauptfragestellung: Besitzen videobasierte Aufgaben (d.h. V-Aufgaben und mV-Aufgaben) zur Experimentalphysik 1 eine größere Wirksamkeit als inhaltsähnliche traditionelle Aufgaben?

Hyp 1.1 Videobasierte Aufgaben fördern die Repräsentationskompetenz (Kinematik).

Die *Repräsentationskompetenz* in der Interventionsgruppe (IG) ist größer als in der Kontrollgruppe (KG), d.h. videobasierte Aufgaben führen zu einer größeren Repräsentationskompetenz im Vergleich zu traditionellen Aufgaben.

Hyp 1.2 Videobasierte Aufgaben fördern das Konzeptverständnis (Kinematik).

Das *konzeptionelle Verständnis* in der IG ist größer als in der KG, d.h. videobasierte Aufgaben führen zu einem größeren konzeptionellen Verständnis im Vergleich zu traditionellen Aufgaben.

Hyp 1.3 Videobasierte Aufgaben fördern die Motivation\*.

Die *Motivation* in der IG ist größer als in der KG, d.h. videobasierte Aufgaben führen zu einem größeren Motivationsgrad verglichen mit traditionellen Aufgaben.

\*Motivation umfasst die Subdimensionen Realitätsbezug, Interesse / Engagement, Selbstkonzept, disziplinäre Authentizität, Experimentbezug und Autonomie.

Hyp 1.4 Videobasierte Aufgaben fördern die Neugierde.

Die *Neugierde* in der IG ist größer als in der KG, d.h. videobasierte Aufgaben führen zu größerer Neugierde verglichen mit traditionellen Aufgaben.

## 2. Fragestellungen zur aufgabeninduzierten Belastung: Führen videobasierte Aufgaben zu einer größeren Belastung als inhaltsähnliche traditionelle Aufgaben?

- Hyp 2.1 Die empfundene *Aufgabenschwierigkeit* ist in beiden Untersuchungsgruppen gleich groß.
- Hyp 2.2 Die wahrgenommene *Beanspruchung* ist in beiden Untersuchungsgruppen gleich groß.
- Hyp 2.3 Die *Bearbeitungszeit der Aufgaben (Time-on-Task)* ist in beiden Untersuchungsgruppen gleich groß.

### 3. Methodische Fragestellungen:

- Hyp 3.1 Videobasierte Aufgaben besitzen für die Lernenden einen größeren *Experimentbezug* als traditionelle Aufgaben. (Manipulation Check)
- FF 3.2 Welchen Einfluss hat die *relative Gewichtung beider KiRC-Testhälften* auf das Ergebnis der Lernwirksamkeitsuntersuchung?
- FF 3.3 Welchen Einfluss hat die *Wahl der KiRC-Itemgruppe* auf das Ergebnis der Lernwirksamkeitsuntersuchung?
- FF 3.4 Welchen Einfluss hat die Verwendung *ratekorrigierter Testsummen* („Formula Scoring“) auf das Ergebnis der Lernwirksamkeitsuntersuchung?

### 4. Fragestellungen zur Robustheit: Wie stark hängt der Interventionserfolg von verschiedenen Einflussparametern ab?

- FF 4.1 Welchen Einfluss hat das *Geschlecht* auf die abhängigen Variablen der Hauptfragestellung? (Gender-Effekte)
- FF 4.2 Welchen Einfluss haben die *Vorleistungen* auf die abhängigen Variablen der Hauptfragestellung? (Matthäus-Effekte)
- FF 4.3 Hängt der Lernzuwachs von *metakognitiven Fähigkeiten* ab? (Dunning-Kruger-Effekt)
- FF 4.4 Hängt der Lernzuwachs von dem *Selbstkonzept* ab?
- FF 4.5 Hängt der Lernzuwachs von der *empfundenen Beanspruchung, Aufgabenschwierigkeit* oder *Time-on-Task* ab?

Die beiden Forschungsfragen aus Kap. 8.2 komplettieren die Fragestellungen, die in dieser Arbeit untersucht werden. Da diese nicht im unmittelbaren Bezug zur Wirksamkeitsfrage videobasierter Aufgaben stehen, werden sie an dieser Stelle nicht weiter aufgeführt.



# Kapitel 10

## Material und Methode

In diesem Kapitel wird beschrieben, mit welchen Methoden die Hypothesen und Forschungsfragen des vorangegangenen Kapitels beantwortet werden. Dazu wird zunächst die Stichprobe anhand deskriptiver Daten charakterisiert (Kap. 10.1), das Untersuchungsdesign dargestellt, begründet und methodisch diskutiert (Kap. 10.2). Aufgrund des Designs, der Gruppenzuweisung von Studierenden und deren Anwesenheit zu verschiedenen Testzeitpunkten wird eine Kern-Stichprobe definiert, die die Datengrundlage der meisten Analysen bildet. Da das Instruktionsmaterial und die Testinstrumente bereits in Kap. 4 und in Teil II dieser Arbeit ausführlich dargestellt wurden, werden an dieser Stelle nur die relevanten Informationen in Bezug auf den Einsatz in der Studie spezifiziert (Kap. 10.3 und 10.4). Dazu zählen zum Beispiel der effektive Anteil videobasierter Aufgaben in der Interventionsgruppe und die Einsatzzeitpunkte der Instrumente.

Schließlich werden die statistischen Methoden zur Auswertung der Ergebnisse erläutert (Kap. 10.5) und eine Übersicht über das Vorgehen zur Beantwortung der Fragestellungen gegeben (Kap. 10.6).

### 10.1 Gesamtstichprobe

Die Hauptuntersuchung fand in den beiden Wintersemestern 2013/14 (WS 13/14) und 2014/15 (WS 14/15) im Rahmen der wöchentlichen Übungen zur Experimentalphysik 1 statt. Insgesamt nahmen 165 Studierende (127 männlich / 38 weiblich) am ersten Testzeitpunkt zum Semesterbeginn teil, die mit folgender Häufigkeit die jeweiligen Studiengänge belegten:

- 53 × Diplomphysik (32%)
  - 34 × Bachelor Physik (21%)
  - 17 × Lehramt Physik (10%)
  - 14 × Biophysik (8%)
  - 47 × Physik als Nebenfach (28%)
- } 118 Hauptfach-Studierende

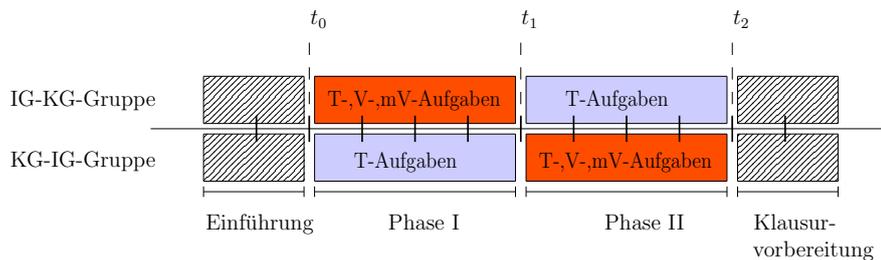
Die Mehrzahl der Studierenden (62%) besuchte in ihrer Schulzeit einen Physik-Leistungskurs; 79% besuchten einen Mathematik-Leistungskurs. Diese hohen Werte stehen in Einklang mit den Ergebnissen fachmonographischer Studien (Ramm, 2008) und sind ein Indikator für ein hohes Fachinteresse. Die Studierenden weisen im Durchschnitt hohe Vorleistungen in diesen Fächern auf: Die durchschnittliche MSS-Punktzahl in Physik beträgt  $M = 11.8$  Punkte ( $SD = 2.4$  Punkte) und in Mathematik  $M = 11.7$  Punkte ( $SD = 2.6$  Punkte) auf einer Skala von 0 (ungenügend) bis 15 Punkten (sehr gut).

Die Hochschulzugangsberechtigung wurde von den meisten Studierenden in Rheinland-Pfalz erworben (65%), gefolgt vom Saarland (9%), Baden-Württemberg und Hessen (jeweils

6%). Die Abitur-Gesamtnote als Grad der Qualifikation beträgt durchschnittlich  $M = 2.1$  ( $SD = 0.7$ ), und ist im rheinland-pfälzischen Vergleich ( $M = 2.6$ ) überdurchschnittlich.<sup>1</sup> Alle Angaben zur Vorleistung beziehen sich auf einen Zeitraum von etwa 2 Jahren und beruhen auf Selbstauskünften der Studierenden zum ersten Testzeitpunkt.

## 10.2 Studiendesign

Das Design der Untersuchung ist in Abb. 25 dargestellt. Die Studierenden wurden zu Beginn randomisiert in zwei Gruppen eingeteilt (für Details zur Gruppeneinteilung siehe Unterabschnitt 10.2.1). Die Bezeichnungen „IG-KG-Gruppe“ bzw. „KG-IG-Gruppe“ beziehen sich auf die chronologische Rolle der Gruppen im Rotationsdesign (zunächst Interventionsgruppe, dann Kontrollgruppe bzw. umgekehrt). Im Folgenden werden nur die Bezeichnungen KG und IG verwendet, wenn klar ist, auf welche Interventionsphase sich bezogen wird.



**Abb. 25:** Studienverlaufsplan: Zeitstrahl des ersten Semesters von links nach rechts mit den Testzeitpunkten  $t_0$ ,  $t_1$  und  $t_2$ . Vertikale Striche deuten Übungseinheiten bzw. Semesterwochen an. KG = Kontrollgruppe, IG = Interventionsgruppe.

Die Studierenden durchliefen in einem Untersuchungssemester mehrere Phasen:

- In einer zweiwöchigen *Einführungsphase* erhielten alle Studenten im Rahmen zweier Präsenzübungen eine standardisierte Einführung in das Videoanalyse-Programm. Dazu zählten die Erklärung des Messprinzips, die Installation der Software auf dem eigenen Notebook<sup>2</sup> sowie das Einarbeiten in die Programmfunktionen durch einen Wechsel zwischen Demonstrationen und angeleitetem selbsttätigem Arbeiten. Alle Studenten erhielten außerdem eine Kopie der Installationssoftware sowie das in der Einführungsphase verwendete Videomaterial mit einer ersten fakultativen Übungsaufgabe zur Wiederholung der Programmbedienung im Selbststudium. In der Einführungsphase wurden außerdem ein „nulltes“ Übungsblatt zur Auffrischung mathematischer Kernkompetenzen (Inhalte der Sekundarstufe II) sowie ein erstes Übungsblatt mit ausschließlich T-Aufgaben ausgeteilt.
- Die *erste Interventionsphase* beginnt in der dritten Vorlesungswoche mit der Bearbeitung des zweiten verpflichtenden Übungsblattes. Während dieser vierwöchigen Phase unterschieden sich die Übungsblätter zwischen den Gruppen. Die untere Gruppe in Abb. 25 war in dieser Phase Kontrollgruppe (KG) und bearbeitete ausschließlich T-Aufgaben, die obere Gruppe war Interventionsgruppe (IG) und bearbeitete eine Mischung aus T-, V- und mV-Aufgaben (siehe Kap. 10.3). Alle Übungsaufgaben wurden im Selbststudium innerhalb einer Woche bearbeitet, zur Korrektur eingereicht und in der darauffolgenden Übung mit den Übungsgruppenleitern besprochen.

<sup>1</sup>Als Bezugspunkt wurden Daten zur durchschnittlichen Abiturnote im Jahr 2013 herangezogen (<http://de.statista.com/statistik/daten/studie/36277/umfrage/durchschnittliche-abiturnoten-im-vergleich-der-bundeslaender/>; 03/2016).

<sup>2</sup>Studierenden ohne PC-Zugang erhielten die Möglichkeit, Notebooks auszuleihen.

- Nach der ersten Interventionsphase tauschten die Übungsgruppen ihre Rollen und weitere vier Interventionswochen schlossen sich an (*Zweite Interventionsphase*). Diese Phase eignet sich aus mehreren Gründen nur bedingt dazu, Erkenntnisse über die Wirksamkeit der videobasierten Aufgaben zu gewinnen, weshalb der Interpretationsschwerpunkt auf der ersten Phase liegt (vgl. Diskussion in Kap. 10.2.2).
- Wegen der Klausur am Semesterende wurde davon abgesehen, die Studie bis zum Semesterende auszudehnen. In der Phase der *Klausurvorbereitung* enthielten die letzten beiden Übungsblätter für *beide Gruppen* sowohl T- als auch V-Aufgaben. Das Klausurergebnis kann nur bedingt als Indikator des Interventionserfolgs ausgewertet werden, da zum Semesterende jeder Student mit Videoanalyse-Aufgaben gearbeitet hat. Möglichkeiten, wie das Klausurergebnis dennoch berücksichtigt werden kann, werden in Kap. 10.4 diskutiert.

Das Rotationsdesign impliziert drei Testzeitpunkte zur Erfassung der Untersuchungsvariablen, die unmittelbar vor der ersten Interventionsphase ( $t_0$ ), direkt danach ( $t_1$ ) sowie nach der zweiten Phase liegen ( $t_2$ ). Die Testung nahm jeweils eine halbe Übungsveranstaltung (45–50 min) in Anspruch.

### 10.2.1 Gruppeneinteilung und Kern-Stichprobe

In den beiden Untersuchungssemestern wurden jeweils fünf Übungsgruppen angeboten, in die sich die Studierenden unmittelbar nach der ersten Vorlesung ohne Einflussnahme der Versuchsleiter verteilten. In jedem Semester standen zwei Übungsleiter zur Verfügung, die parallel eine Kontroll- und eine Interventionsgruppe betreuen konnten. Nachdem sich die Studierenden auf die Übungsgruppen verteilten, wurde unter Berücksichtigung gleicher Gesamtgruppengrößen entschieden, welche Übungsgruppe zunächst Kontroll- bzw. Interventionsgruppe war. Tabelle 28 zeigt die Aufteilung der Gesamtstichprobe in die Übungsgruppen zum Beginn der Studie.

Die vier Betreuer B1, B2, B4 und B5 instruierten aus Gründen der fairen Vergleichbarkeit jeweils eine KG und eine IG. Es wurde darauf geachtet, dass diese Gruppen durch möglichst viele Hauptfach-Studierende und möglichst wenige Nebenfach-Studierende besetzt waren, da letztere für die Auswertung der Ergebnisse nicht relevant sind.<sup>3</sup>

So gelang es, dass sich 94% der Hauptfach-Studierenden (111 von 118) in parallelen Übungsgruppen befanden. Die 7 Hauptfach-Studenten, die von dem Übungsleiter B3 betreut wurden, bleiben mangels fehlender methodischer Kontrolle in der Vergleichsanalyse unberücksichtigt.

<sup>3</sup>Da Nebenfach-Studenten weder Präsenz- noch Bearbeitungspflicht der Übung haben, werden diese in der Analyse nicht berücksichtigt.

**Tab. 28:** Gruppengrößen zu Beginn der Intervention ( $t_0$ )

Semester	Betreuer	KG	IG	Nebenfach-Studenten <sup>a</sup>
WS 13/14	B1	12	7	7
	B2	19	16	2
	B3	-	(7)	12
WS 14/15	B4	11	13	6
	B5	11	22	4
	B6	-	0	16
Gesamt		53	58	47

*Bemerkungen:* KG und IG beziehen sich auf die erste Interventionsphase. Hauptfach-Studenten in der Gruppe von Betreuer B3 sowie Nebenfach-Studenten werden in der Auswertung nicht berücksichtigt.

<sup>a</sup>Anzahl an Nebenfach-Studenten in KG und IG zusammen.

Zur Überprüfung, ob sich die Übungsleiter zwischen den betreuten Parallelgruppen unterschiedlich verhielten, wurde das Betreuerverhalten von den Studierenden beurteilt (Betreuerfragebogen, vgl. Kap. 5.7).

### Kern-Stichprobe und Studien-Dropout

Die ursprüngliche Stichprobe aus 111 relevanten Hauptfach-Studenten dezimierte sich aufgrund von Krankheitsausfällen sowie Studienfachwechslern/-abbrechern zu den späteren Testzeitpunkten auf 78 ( $t_1$ ) bzw. 68 ( $t_2$ ) Studierende. In der verbundenen Stichprobe der Zeitpunkte  $t_0$  und  $t_1$  befinden sich 76 Studierende (40 in der KG, 36 in der IG). Wie im nächsten Abschnitt diskutiert wird, stellt diese *Kern-Stichprobe* die Grundlage vieler Analysen dar. Die 35 Hauptfach-Studenten, die in dieser Schnittmenge zwar zum Zeitpunkt  $t_0$ , aber nicht zum Zeitpunkt  $t_1$  in der Übung anwesend waren, werden als Studienabbrecher (Dropout) bezeichnet. In Kapitel 11.1.1 werden systematische Unterschiede zwischen diesen beiden Studierendengruppen analysiert und in Kap. 12.4 diskutiert.

## 10.2.2 Begründung des Studiendesigns und kritische Stellungnahme

Das Studiendesign wurde mit den Vorlesungsdozenten der Experimentalphysik 1 unter Berücksichtigung von zeitlichen und organisatorischen Rahmenbedingungen im Vorfeld der Untersuchung abgestimmt. So stellte sich die Verwendung eines Rotationsdesigns als guter Kompromiss zwischen ausreichender Interventionsstärke (jeweils vier Wochen) und Fairness gegenüber den Studierenden dar (zum Semesterende bearbeitete jeder Student den gleichen Anteil videobasierter Aufgaben). In einem einfachen Vergleichsgruppen-Design ohne Rotation der Gruppen könnten zwar höhere Interventionsstärken realisiert werden, aber Studierende könnten sich benachteiligt fühlen, nicht mit videobasierten Aufgaben arbeiten zu dürfen (oder — je nach subjektiver Sichtweise — mit ihnen arbeiten zu müssen). Das Rotationsdesign hat prinzipiell die Stärke, dass eine größere Datengrundlage gesammelt wird (drei Testzeitpunkte), jedoch stellt sich bei kritischer Betrachtung heraus, dass beide Interventionsphasen aus methodischer Sicht nicht gleichwertig interpretierbar sind:

- Die Studierenden der beiden Gruppen (KG-IG bzw. IG-KG) unterscheiden sich zu Beginn der zweiten Interventionsphase *systematisch* voneinander, da sie zuvor mit unterschiedlichen Aufgabenformaten arbeiteten.
- Es ist zu vermuten, dass die Studierenden beider Gruppen den Materialwechsel (von T-Aufgaben zu videobasierten Aufgaben bzw. umgekehrt) unterschiedlich wahrnehmen und sich dies auf Motivation (und Leistung) während der zweiten Phase auswirkt.
- Die Leistungstests sind inhaltlich auf die erste Interventionsphase abgestimmt.

Außerdem unterscheiden sich wegen des zeitlichen Voranschreitens der Vorlesung die Instruktionsmaterialien zwischen den Interventionsphasen zwangsläufig, sodass die Ergebnisse beider Phasen nur bedingt in Bezug zueinander gesetzt werden können. Trotz diesen methodischen Schwächen der zweiten Interventionsphase werden die Ergebnisse mit gleicher statistischer Sorgfalt ausgewertet wie die der ersten Phase. Der Interpretationsschwerpunkt liegt aus den genannten Gründen jedoch auf dem ersten Interventionszeitraum.

In diesem Zusammenhang soll auch die Notwendigkeit einer Einführungsphase zu Beginn des Semesters herausgestellt und die Konsequenz für die Ergebnisinterpretation diskutiert werden: In der Einführungsphase tritt auch die (zeitlich erste) Kontrollgruppe schon in Kontakt mit dem Medium Videoanalyse. Dies ist aus praktischer Sicht unverzichtbar, da für diese Gruppe wegen der strikten zeitlichen Regelung des Übungsbetriebs keine Möglichkeit besteht, die Einführung in das Medium unmittelbar vor dem eigentlichen Gebrauch — also beim Gruppenwechsel — nachzuholen. Aus methodischer Sicht stellt die Einführungsphase kein

Problem dar, da sich der erste Testzeitpunkt ( $t_0$ ) an diese Phase anschließt und eventuelles Lernen während der Einführungsphase im Prä-Test erhoben wird und somit keinen Einfluss auf den Gruppenvergleich hat. Ein möglicher positiver Lerneffekt vor der eigentlichen Untersuchung wird damit kontrolliert. Falls in der Einführungsphase schon Lerneffekte durch die Videoanalyse generiert werden, dann ist die Untersuchung konsequenterweise weniger sensitiv gegenüber unterschiedlichen Lernzuwächsen. Alle erzielten Ergebnisse sind als konservative Abschätzung zu betrachten, die ohne Einführungsphase vermutlich größer ausfallen würden. Der verzögerte Start der Intervention in der dritten Vorlesungswoche hat hingegen auch Vorteile: Die Mechanik des Massenpunktes (insbesondere die Kinematik), die wesentlicher Inhalt der Leistungsinstrumente ist, wurde zum ersten Testzeitpunkt schon in der Vorlesung behandelt. Bei der Prä-Testung kann also davon ausgegangen werden, dass grundlegende Kenntnisse zur Kinematik vorhanden sind, sodass das fehlende Grundwissen nicht mit den Hauptvariablen Repräsentationskompetenz bzw. Konzeptverständnis konfundiert (vgl. Kap. 5.1.1). Außerdem gewinnt die Stichprobe an Homogenität bzgl. der Leistung und die Untersuchung damit an Kontrolle.

### 10.3 Instruktionsmaterial

Das Instruktionsmaterial bestand aus den Aufgaben, die im Rahmen der Übungen zur Bearbeitung standen, wurden in Kap. 4 ausführlich beschrieben und sind im Anhang dargestellt (ab S. 220).<sup>4</sup> Das wesentliche Unterscheidungsmerkmal zwischen videobasierten und traditionellen Aufgaben besteht in dem Umgang mit experimentellen Daten und dem Nutzen multipler Repräsentationen. Bei videobasierten Aufgaben kann der Lernende jederzeit zwischen realem Videobild und Datendarstellung wechseln, nutzt Diagramme, Tabellen und ins Videobild eingeblendete Vektoren zur Aufgabebearbeitung und verfügt dadurch über Kontrollmöglichkeiten der eigenen Ergebnisse. Diese Tätigkeiten finden während der Aufgabebearbeitung entweder implizit statt oder werden durch die Aufgabenstellung explizit angeleitet.

#### 10.3.1 Interventionsstärke (Aufgabendosis)

Ein wöchentlicher Übungszettel umfasste typischerweise *vier* Aufgaben mit einem engen inhaltlichen Bezug zu aktuellen Vorlesungsthemen, um diese systematisch zu wiederholen und mit Anwendungsbeispielen zu verknüpfen.

IG Die Interventionsgruppe bearbeitete pro Übungsblatt zwei T-Aufgaben und zwei videobasierte Aufgaben.

Über eine gesamte Interventionsphase bearbeitete die IG sechs V-Aufgaben, zwei mV-Aufgaben und acht T-Aufgaben, wobei die beiden mV-Aufgaben gegen Ende der Interventionsphase lagen.

KG Die Kontrollgruppe bearbeitete ausschließlich T-Aufgaben (vier pro Aufgabenblatt). Zwei T-Aufgaben waren identisch mit den beiden T-Aufgaben der IG, die anderen beiden Aufgaben korrespondierten zu den videobasierten Aufgaben. Über eine gesamte Interventionsphase bearbeitete die KG also 16 T-Aufgaben.

→ Aufgrund des Rotationsdesigns bearbeitete jeder Student über die gesamte Vorlesungszeit die gleiche Anzahl von T-, V- und mV-Aufgaben.

Obwohl gemäß den Instruktionsmaterialien de facto eine Mischung aus videobasierten und traditionellen Aufgaben gegen rein traditionelle Aufgaben verglichen wurde, können in der Studie festgestellte Gruppenunterschiede bzgl. den untersuchten Variablen auf den Einsatz

<sup>4</sup>Da eine hohe Redundanz zwischen den Aufgaben aus den beiden Studien (WS 13/14 und WS 14/15) besteht, sind lediglich die Aufgaben der zweiten Hauptstudie (WS 14/15) abgebildet. Im digitalen Anhang befinden sich alle Übungsblätter beider Studien.

videobasierter Aufgaben gegenüber korrespondierenden T-Aufgaben zurückgeführt werden, da die übrigen T-Aufgaben in beiden Gruppen identisch waren.

### 10.3.2 Faire Vergleichbarkeit der Aufgaben

Um das Instruktionsmaterial in beiden Gruppen vergleichbar zu halten, wurden T-V- bzw. T-mV-Aufgabenpaare konstruiert, die sich auf den gleichen physikalischen Inhalt beziehen, auf den identischen Kontext zurückgreifen und einen vergleichbaren Schwierigkeitsgrad aufweisen. Bevor das entwickelte Instruktionsmaterial in den Übungen eingesetzt wurde, fanden wöchentliche Absprachen mit den lehrenden Professoren, den Kollegen innerhalb der Didaktik-Arbeitsgruppe und der Übungsordination statt. In dieser Diskussionsrunde wurden zum Vorlesungsfortschritt passende Aufgabenstellungen aus einem bestehenden Aufgabenpool ausgewählt, wobei auf eine sinnvolle Streuung der Anforderungen geachtet wurde. Bei der Auswahl der beiden Aufgabenpaare wurden die jeweiligen Teilaufgaben aufeinander abgestimmt, um den Arbeitsaufwand für beide Gruppen etwa gleich zu halten. Das Bearbeiten der Übungsaufgaben war für die Hauptfach-Studenten verpflichtend. Um zur Abschlussklausur zugelassen zu werden, mussten die Studierenden mindestens 50% aller erreichbaren Punkte erzielen.<sup>5</sup>

## 10.4 Zielvariablen und Erhebungsinstrumente

Die Testinstrumente wurden in Teil II dieser Arbeit ausführlich dargestellt, innerhalb der Eich-Stichprobe analysiert und sind in Tab. 29 mit Verweisen auf die jeweiligen Kapitel zusammenfassend dargestellt. Die primären Zielvariablen sind die Repräsentationskompetenz (kognitiv), das Konzeptverständnis von Bewegungen (kognitiv), die Motivation (verschiedene Subskalen) und die Neugierde (als aktueller Zustand). Begleitvariablen, die für die Einschätzung des Erfolgs der Intervention essenziell sind, umfassen die durch das Studium empfundene Beanspruchung, die wahrgenommene Aufgabenschwierigkeit und die Bearbeitungszeit der Übungsaufgaben. Die Beanspruchung und die Aufgabenschwierigkeit stellen dabei einerseits Kontrollvariablen dar und sind andererseits selbst Zielvariablen gemäß den Hypothesen 2.1 und 2.2. Das Verhalten des Übungsgruppenleiters (Betreuerskala) wurde rein zu methodischen Kontrollzwecken erhoben. Darüber hinaus wurden die Vorleistungen (Abiturnote, Mathematik- und Physikleistungen in der Schule) abgefragt, der Besuch von schulischen Leistungskursen sowie das Geschlecht. Die Klausur-Gesamtnote eignet sich aufgrund des Studiendesigns nicht zur Abschätzung des Interventionserfolgs. Es können jedoch einzelne Klausuraufgaben, die sich im Wesentlichen auf die Inhalte und Konzepte der Kinematik beschränken, der ersten Interventionsphase zugeordnet werden (Kap. 11.6). Außerdem trägt die Untersuchung des Zusammenhangs zwischen dem Klausurergebnis, den Kovariaten und den kognitiven Zielvariablen der Intervention (Konzeptverständnis und Repräsentationskompetenz) dazu bei, Prädiktoren für den Studienerfolg zu identifizieren.

Um die Qualität der Instrumente zu den späteren Zeitpunkten einzuschätzen, wurde eine Neubestimmung der wichtigsten Skalengrößen in der Kern-Stichprobe zu allen Testzeitpunkten durchgeführt, deren Ergebnis in Kap. 11.1.3 berichtet wird. Bis auf folgende Ausnahmen wurden alle in Tab. 29 genannten Instrumente zu jedem der drei Testzeitpunkte in beiden Hauptstudien eingesetzt (vgl. auch Tab. 6 auf S. 85):

- Neugierde als Personeneigenschaft: Die Items zur Erfassung dieser Variable standen erst in der zweiten Hauptstudie zur Verfügung. Aufgrund fehlender Werte für Studierende im WS 13/14 wird diese Variable in den Vergleichsanalysen gesondert behandelt.
- Neugierde als aktueller Zustand, Bearbeitungszeit der Aufgaben und Beurteilung des Übungsleiters: Diese Variablen wurden nicht zum Prä-Zeitpunkt erhoben, da sie erst

<sup>5</sup>Im ersten Teil der Hauptstudie (WS 13/14) mussten mindestens 50% der videobasierten Aufgaben und 50% der traditionellen Aufgaben erfolgreich bearbeitet werden. Im WS 14/15 wurde auf diese Regel verzichtet.

nach mehrfacher Teilnahme an den Übungen und Bearbeitung der Instruktionsmaterialien valide eingeschätzt werden können.

- Test zur Erfassung der Repräsentationskompetenz (KiRC): Es wurde darauf verzichtet, KiRC zum Testzeitpunkt  $t_2$  einzusetzen, da die Testsummen zum ersten Post-Testzeitpunkt schon Deckeneffekte zeigten und der Testinhalt (Repräsentationen in der Kinematik) kein wesentlicher Gegenstand der zweiten Interventionsphase war.
- Die Vorleistungen und die Abiturnote wurden nur zum Prä-Zeitpunkt erhoben.

**Tab. 29:** Variablen und Instrumente der Hauptuntersuchung

	(Sub-)Skalen	Beschreibung	Analyse	Zeitdauer (min)
<i>Abhängige Variablen</i>				
Leistung	KiRC	5.1	6.4.4	25
	Konzepttest	5.2	6.4.5	5
Motivation	SK, IE, RA, dA, Atnm, Exp	5.4	6.4.3	10
Neugierde	NP, NZ	5.5	6.4.2	3
Beanspruchung	PSQ	5.6	6.4.1	1
Aufgabenschwierigkeit (Klausur)	CL	5.6	6.4.1	2
	-	-	-	-
<i>Unabhängige Variablen</i>				
Gruppenzugehörigkeit	-	-	-	-
Geschlecht	-	-	-	-
<i>Kovariaten / Kontrollvariablen &amp; Moderatorvariablen</i>				
Beanspruchung	PSQ	5.6	6.4.1	s.o.
Aufgabenschwierigkeit	CL	5.6	6.4.1	s.o.
Abiturnote	-	-	-	-
Vorleistungen Ma / Phy	-	-	-	-
Bearbeitungszeit	-	-	-	-
Betreuerverhalten	Betreuerskala	5.7	6.4.7	1
<i>Manipulation Check</i>				
Experimentbezug	Exp	5.4.4	6.4.3	s.o.

*Bemerkungen:* KiRC = Repräsentationskompetenz (Kinematik), SK = Selbstkonzept, IE = Interesse / Engagement, RA = Realitätsbezug, dA = disziplinäre Authentizität, Atnm = Autonomie, Exp = Experimentbezug, NP / NZ = Neugierde als Personeneigenschaft / als aktueller Zustand, PSQ = Beanspruchung (Perceived Stress Questionnaire), CL = Aufgabenschwierigkeit (Cognitive Load).

## 10.5 Statistische Methoden und Kenngrößen

Die verwendeten Auswertemethoden umfassen primär (Ko-)Varianzanalysen, an wenigen Stellen (einfache) Regressionsanalysen und Korrelationsanalysen. Diese Verfahren werden als bekannt vorausgesetzt (Bortz & Schuster, 2010), weshalb nur die grundsätzliche Idee der Kovarianzanalyse sowie deren notwendigen Voraussetzungen benannt werden (Kap. 10.5.1). Nicht etabliert hingegen ist die Idee, die Auswirkung der Art der Testsummenbildung auf die Ergebnisse der Vergleichsanalysen, also insbesondere den Effektstärken  $d$ , zu untersuchen (Forschungsfrage 3.2). Die variable relative Gewichtung zweier Teilsommen zu einer Gesamtsumme führt zu einigen methodischen Vorteilen und Besonderheiten gegenüber der aus pragmatischen Gründen häufig benutzten Gleichgewichtung von Testkomponenten. Wie sich herausstellt, ist diese *Testmischungsanalyse* im varianzanalytischen Sinne eine Verallgemeinerung der sogenannten Ratekorrektur, die sich zur Verrechnung zweier Testhälften mit unterschiedlichen Ratewahrscheinlichkeiten eignet. Da Ratekorrekturen in der Praxis häufig verwendet werden und im Rahmen dieser Arbeit der Ausgangspunkt zur Entwicklung der Testmischungsanalyse waren, werden die methodischen Grundlagen in Kap. 10.5.3 dargestellt.

### 10.5.1 Kovarianzanalyse (ANCOVA)

Varianzanalysen werden genutzt, um die Bedeutsamkeit einer unabhängigen Variable (im vorliegenden Fall beispielsweise des Faktors „Gruppe“ oder „Geschlecht“) für eine abhängige Variable (zum Beispiel Leistung) zu untersuchen (Bortz & Döring, 2006; Bortz & Schuster, 2010). Im einfachsten Fall eines einfaktoriellen Versuchsdesigns wird die totale Quadratsumme, die die Variabilität aller Messwerte enthält, in eine Varianzkomponente zerlegt, die auf die Aufteilung in die Untersuchungsgruppen zurückzuführen ist („Treatmentquadratsumme“), und in einen übrigen Teil, der zusammengefasst als Fehlerquadratsumme bezeichnet wird und den Einfluss von *Störvariablen* enthält. Unter Berücksichtigung der Freiheitsgrade beider Varianzkomponenten wird der  $F$ -Bruch gebildet, der die mittleren Treatment- und Fehlerquadratsummen ins Verhältnis setzt. Falls die Mittelwertsunterschiede zwischen den Gruppen statistisch bedeutsam sind, dann unterscheidet sich die mittlere Quadratsumme des Treatments signifikant von der mittleren Fehlerquadratsumme und der  $F$ -Bruch überschreitet einen kritischen Wert, der von den Freiheitsgraden und des gewählten  $\alpha$ -Niveaus abhängt.

Die Kovarianzanalyse (ANCOVA) basiert auf diesem Prinzip und versucht die Fehlervarianz durch die Kontrolle von Störvariablen zu reduzieren. Dies gelingt dadurch, neben einer abhängigen Variable und dem Faktor zusätzlich theoretisch begründete *Kovariaten* in das Modell einzubeziehen, die einen Teil der Fehlervarianz erklären und diese damit reduzieren können. Um den Einfluss dieser Kovariaten auf die abhängige Variable zu neutralisieren, werden die Mittelwerte der abhängigen Variable durch regressionsanalytische Techniken korrigiert. Neben diesem Ziel, eine abhängige Variable von einer kontrollierbaren Störgröße „zu bereinigen“, liefert die ANCOVA zudem die Information, wie bedeutsam einzelne Kovariaten für die Untersuchung sind.

Neben einer möglichst hohen Reliabilität der gemessenen (metrisch skalierten) Kovariaten müssen mehrere Voraussetzungen gegeben sein, um eine Variable als Kovariate in die Varianzanalyse einbeziehen zu können (Bortz & Schuster, 2010, S. 311–313). Erstens muss die Kovariate mit der abhängigen Variable korreliert sein, da die Fehlervarianzreduktion ansonsten entweder nicht auftritt oder rein zufällig ist. Zweitens dürfen sich die Regressionssteigungen zwischen Kovariate und abhängiger Variable zwischen den Gruppen nicht signifikant voneinander unterscheiden, da ansonsten eine Interaktion zwischen der Kovariate und der unabhängigen Variable vorliegt (Hollingsworth, 1980). Eine dritte Voraussetzung, gegen deren Verletzung die Kovarianzanalyse allerdings robust reagiert (Stevens, 2012), betrifft die Normalverteilung der Kovariaten innerhalb der beiden Untersuchungsgruppen. Viertens sollten sich die Gruppen nicht signifikant bezüglich den Kovariaten unterscheiden (Gruppenunabhängigkeit). Eine Verletzung gegen die Gruppenunabhängigkeit stellt aber genau dann kein Ausschlusskriterium der Kovariaten dar, wenn die Ursache der Unterschiede nicht systematischer Natur ist, sondern glaubhaft auf Zufall zurückgeführt werden kann (unglückliche Randomisierung, Jamieson, 2004). Unabdingbar aber ist die fünfte Bedingung, dass (zeitlich veränderliche) Kovariaten *vor* der Intervention erhoben werden müssen, da ansonsten die Gefahr besteht keine Fehlervarianz (sondern eine systematische Varianz) zu korrigieren (Mayers, 2013, S. 368). Neben den fünf genannten Bedingungen zur Verwendung von Kovariaten im Analysemodell gelten zudem die Voraussetzungen der Varianzanalyse, die an dieser Stelle als bekannt vorausgesetzt werden dürfen (Normalverteilung der abhängigen Variable, Varianzhomogenität, vgl. Bortz & Schuster, 2010).

### 10.5.2 Effektstärken

Zur Bewertung der praktischen Relevanz eines Ergebnisses aus einer statistischen Analyse wird die Effektstärke herangezogen. Ein übliches Maß ist Cohen's  $d$ , definiert durch

$$d = \frac{\mu_A - \mu_B}{\sigma}, \quad (23)$$

d.h. der Mittelwertunterschied  $\mu_A - \mu_B$  zwischen zwei Vergleichsgruppen (oder zwischen zwei Testzeitpunkten) A und B wird ins Verhältnis zur Standardabweichung gesetzt (Cohen, 1992).

Während als Schätzer für  $\mu_A$  und  $\mu_B$  üblicherweise die Mittelwerte  $M$  bzw. die Testsummen  $T$  aus den empirischen Daten verwendet werden, gibt es zur Schätzung der Standardabweichung  $\sigma$  verschiedene Ansätze. Neben der gepoolten Standardabweichung der beiden betrachteten Gruppen (oder Testzeitpunkten) schlägt Glass (1976) beispielsweise vor, ausschließlich die empirische Standardabweichung der Kontrollgruppe  $SD_{KG}$  zu nutzen, d.h.

$$d = \frac{M_{IG} - M_{KG}}{SD_{KG}}. \quad (24)$$

Beim Vergleich von Testzeitpunkten rückt die Standardabweichung zum Prä-Zeitpunkt an die Stelle von  $SD_{KG}$ . Durch das Nutzen der Standardabweichung der Kontrollgruppe / des Prä-Zeitpunktes wird eine konservative Schätzung der Effektstärke realisiert, da diese Standardabweichungen i.A. größer ausfallen als die der Interventionsgruppe / des Post-Zeitpunktes. Eine mögliche Varianzreduktion in Folge der Intervention schlägt sich damit nicht in der Effektstärke nieder, wodurch eine Überschätzung von  $d$  vermieden wird.

In einem Prä-Post-Design mit zwei Gruppen kann Gleichung (24) erweitert werden, indem die Effekte zunächst gruppenweise für beide Testzeitpunkte geschätzt und danach voneinander subtrahiert werden (Furtak et al., 2012):

$$d_{\text{Prä/Post, KG/IG}} = \frac{(M_{IG, \text{Post}} - M_{IG, \text{Prä}}) - (M_{KG, \text{Post}} - M_{KG, \text{Prä}})}{SD_{KG, \text{Prä}}} \quad (25)$$

Auch hier fließt bei der Division durch die Standardabweichung der Kontrollgruppe die Annahme ein, dass die Intervention zu einer Varianzreduktion führt, womit die Schätzung von  $d$  gemäß Gl. (25) ebenfalls als konservativ gilt. Nach Cohen (1992) indiziert  $0.2 \leq d < 0.5$  einen kleinen,  $0.5 \leq d < 0.8$  einen mittleren und  $d \geq 0.8$  einen großen Effekt und nach Hattie (2009) sind Effektstärken  $d \geq 0.3$  von praktischer Relevanz („*zone of desired effects*“).

Im Rahmen von Varianzanalysen werden häufig auch andere Effektstärkemaße berichtet, die den Anteil der aufgeklärten Varianz in Relation zur Gesamtvarianz setzen und die in ein  $d$  nach Cohen umgerechnet werden können. Das partielle Eta-Quadrat ( $\eta_p^2$ ) gibt in diesem Zusammenhang den Anteil der aufgeklärten Variabilität der Messwerte relativ zur Summe aus systematischer Effekt-Quadratsumme und Fehlerquadratsumme auf Stichprobenebene an,

$$\eta_p^2 = \frac{QS_{\text{Effekt}}}{QS_{\text{Effekt}} + QS_{\text{Fehler}}}. \quad (26)$$

Davon zu unterscheiden ist das totale Eta-Quadrat  $\eta^2$ , welches sich auf die totale Quadratsumme bezieht,

$$\eta^2 = \frac{QS_{\text{Effekt}}}{QS_{\text{total}}}. \quad (27)$$

Im einfaktoriellen Fall einer Varianzanalyse ohne Kovariaten ist  $\eta_p^2 = \eta^2$ . Da sich beide Größen auf die Variabilität der Messwerte in der gezogenen Stichprobe beziehen, wird die Umrechnung des totalen Eta-Quadrats  $\eta^2$  in den Populationsschätzer  $\omega^2$  empfohlen, welcher dann schließlich in ein  $d$  umgerechnet werden kann (Wolf, 1998):

$$\omega^2 = \frac{\eta^2 \cdot (df_{\text{Fehler}} + df) - df}{1 - \eta^2 + df_{\text{Fehler}}} \quad (28)$$

$$d = 2 \cdot \sqrt{\frac{\omega^2}{1 - \omega^2}}. \quad (29)$$

Dabei bezeichnen  $df_{\text{Fehler}}$  und  $df$  die Fehler- bzw. Effekt-Freiheitsgrade.

### 10.5.3 Ratekorrekturen (Formula Scoring)

Unter einer Ratekorrektur wird eine Korrektur der Testsumme bei Multiple-Choice-Tests verstanden (Frary, 1988). Da bei geschlossenen Antwortformaten durch bloßes Raten bereits eine

Mindestpunktzahl per Zufall erreicht wird, werden die Testsummen unter Berücksichtigung der Anzahl dargebotener Antwortalternativen korrigiert. Bezeichnet  $T$  die (mittlere)<sup>6</sup> Summe der richtigen Antworten bei einem Test, der insgesamt  $n$  Items umfasst, die jeweils  $q$  Antwortalternativen anbieten, dann ergibt sich die korrigierte Testsumme  $T'$  durch eine lineare Transformation der ursprünglichen Testsumme gemäß

$$T' = \left(T - \frac{n}{q}\right) \cdot \frac{q}{q-1}. \quad (30)$$

Anschaulich wird die originäre Testsumme  $T$  zunächst um die Punktzahl reduziert, die durch bloßes Raten zustande kommt (Annahme: jede Alternative wird im Mittel mit der gleichen Wahrscheinlichkeit  $1/q$  gewählt), und anschließend derart gestreckt, dass für  $T = n$  auch  $T' = n$  gilt. Da eine Ratekorrektur eine lineare Transformation der Testsumme ist, hat sie keine Auswirkungen auf die Ergebnisse statistischer Vergleichsanalysen oder Effektstärken.<sup>7</sup> Weniger offensichtlich ist, dass eine Ratekorrektur auch keinen Einfluss auf den normierten Lernzuwachs (sogenannter Hake-Index) hat, siehe Kap. 10.5.4.

Die Ratekorrektur dient primär dazu, Tests (oder Testteile) mit einer unterschiedlichen Anzahl von Antwortalternativen miteinander vergleichbar zu machen und ist deshalb für die Berechnung der KiRC-Testsummen relevant. Wird die Ratekorrektur auf zwei Testhälften 1 und 2 mit unterschiedlicher Anzahl von Items  $n_1$  und  $n_2$  sowie verschiedener Anzahl von Antwortalternativen  $q_1$  und  $q_2$  angewandt, ergibt sich für die Summe der beiden Testhälften

$$T' = T'_1 + T'_2 = \left(T_1 - \frac{n_1}{q_1}\right) \cdot \frac{q_1}{q_1-1} + \left(T_2 - \frac{n_2}{q_2}\right) \cdot \frac{q_2}{q_2-1} \quad (31)$$

$$= \alpha \left(T_1 + \frac{\beta}{\alpha} T_2\right) + \gamma \quad (32)$$

mit

$$\alpha = \frac{q_1}{q_1-1} \quad (33)$$

$$\beta = \frac{q_2}{q_2-1} \quad (34)$$

$$\gamma = -\frac{n_1}{q_1-1} - \frac{n_2}{q_2-1}. \quad (35)$$

Die Gesamttestsumme  $T'$  lässt sich also als gewichtete Testsumme aus den Teilkomponenten darstellen, wobei die Gewichte durch die Anzahl von Antwortalternativen und Testitems festgelegt sind.

### 10.5.4 Normierter Lernzuwachs (Hake-Index)

Der Hake-Index  $g_H$  (im Jargon auch Hake-Gain genannt) ist definiert als normierter mittlerer Lernzuwachs,

$$g_H = \frac{P_{\text{post}} - P_{\text{prä}}}{1 - P_{\text{prä}}}, \quad (36)$$

und wurde von Hake als Maß zur Beurteilung des Lernerfolgs in Physik-Einführungsveranstaltungen eingeführt (Hake, 1998). Der Hake-Index wird aus dem mittleren Anteil richtiger Lösungen zum Prä- und Post-Zeitpunkt  $P_{\text{prä}}$  bzw.  $P_{\text{post}}$  innerhalb einer Gruppe berechnet

<sup>6</sup>Eine Unterscheidung zwischen der Testsumme einer Person und der mittleren Testsumme einer Gruppe von Personen ist an dieser Stelle nicht notwendig. Wenn nicht anders beschrieben, werden im Folgenden durch „ $T$ “ stets Gruppenmittelwerte bezeichnet, wobei Indizes näher spezifizieren, über welche Personengruppen (KG/IG) oder Itemgruppe (z.B. 1/2) die Testsumme gemittelt wird.

<sup>7</sup>Dies gilt nur für die Ratekorrektur einer homogenen Testkomponente — nicht, wenn zwei Testkomponenten mit unterschiedlicher Anzahl von Alternativen gemischt werden — und folgt sofort aus den Rechenregeln für Varianzen, vgl. auch Fußnote 8 (S. 141).

und setzt den erzielten Lernzuwachs ins Verhältnis zum maximal möglichen Lernzuwachs. Werden Testsummen an Stelle von mittleren Lösungswahrscheinlichkeiten verwendet, lautet die Formel entsprechend

$$g_H = \frac{T_{\text{post}} - T_{\text{prä}}}{n - T_{\text{prä}}}, \quad (37)$$

wenn  $n$  die maximal mögliche Testsumme bezeichnet. Hinter der Konstruktion des Index steckt die Annahme, dass bei hohem Vorwissen ein geringerer Lernzuwachs möglich ist als bei wenig Vorwissen. Diese Art der Konstruktion einer individuellen Bezugsnorm ist aus theoretischer Sicht plausibel; allerdings wird sie den Erkenntnissen der Lernpsychologie nicht gerecht, dass Lernende mit hoher Vorleistung oftmals größeres Steigerungspotential haben als schwächere (sogenannter *Matthäus-Effekt*, „Erfolg führt zu Erfolg“, Merton, 1968). Auch wenn die mit dem Index verbundene pädagogische Wertung daher fragwürdig erscheint, ermöglichen die Hake-Indizes studienübergreifende Vergleiche des Lernzuwachses zu zahlreichen Arbeiten in einem ähnlichen (universitärem) Lernszenario. Typischerweise erzielen konventionelle Lehrmethoden normierte Lernzuwächse zwischen  $g_H = 0.2$  und  $0.3$  (Coletta et al., 2007), was Hake als kleinen Lernzuwachs bezeichnet. Ab einem Wert von  $g_H = 0.7$  wird von einem großen, dazwischen von einem mittleren Lernzuwachs gesprochen (Hake, 1998).

Interessanterweise ist der Hake-Index invariant unter der Ratekorrektur, d.h. wird der Hake-Index als Funktion von Prä- und Post-Testvariablen aufgefasst, dann gilt

$$g_H(T_{\text{prä}}, T_{\text{post}}) = g_H(T'_{\text{prä}}, T'_{\text{post}}), \quad (38)$$

wie unter Verwendung von Gl. (30) und Gl. (37) leicht nachgeprüft werden kann. Umgekehrt gehört die Ratekorrektur zu einer besonderen Klasse nicht-trivialer Lineartransformation, die  $g_H$  invariant lässt (Beweis im Anhang, ab S. 219).

### 10.5.5 Gewichtete Testsummenbildung (Testmischungsanalyse)

Wie in Kap. 5.1.3 beschrieben wurde, ist es zielführend, die erzielten Testsummen in zwei Testhälften eines Instruments zu einer gemeinsamen Testsumme zusammenzuführen:

$$T = T_1 + \kappa T_2 \equiv T(\kappa). \quad (39)$$

Der Index 1 bezeichnet hierbei die Testhälfte mit MC3-Items, der Index 2 die Testhälfte mit MTF-Items, und  $\kappa \in [0, \infty)$  bestimmt die relative Gewichtung der beiden Testsummen. Diese (beliebige) Testmischung knüpft an das Konzept zusammengesetzter Testsummen an (Crocker & Algina, 1986), stellt eine Erweiterung der Ratekorrektur dar und enthält diese als Spezialfall: Gleichung (32) ist im varianzanalytischen Sinne<sup>8</sup> äquivalent zu Gl. (39), wenn  $\kappa = \beta/\alpha$  gesetzt wird.

Während sich innerhalb der einzelnen Testhälften 1 und 2 jeweils Effektstärken  $d_1$  und  $d_2$  berechnen lassen, wird im Folgenden untersucht, wie sich die Komposition der einzelnen Testsummen nach Gl. (39) auf die (zusammengesetzte) Effektstärke  $d(\kappa)$  und (weiter unten) auf die Reliabilität des Gesamttests auswirkt. Das primäre Ziel in dieser Behandlung besteht nicht darin, diejenige Gewichtung  $\kappa$  zu finden, die zu extremalen Effekten führt, sondern zu bestimmen, welchen Einfluss die Testsummengewichtung auf den Wertebereich von  $d$  nimmt. Der Unzulänglichkeit, keine Gewichtung a priori festlegen zu können, wird damit begegnet, den Einfluss der Gewichtung auf das Ergebnis des Gruppenvergleichs zu quantifizieren.

Mit Blick auf die Anwendung in dieser Arbeit wird auf die Definition der Effektstärke gemäß Gleichung (24) für den Vergleich zweier Gruppen (KG / IG) zu einem Testzeitpunkt zurückgegriffen. Die Effektstärke  $d$  lässt sich demnach schreiben als

$$d(\kappa) = \frac{T_{\text{IG}}(\kappa) - T_{\text{KG}}(\kappa)}{SD_{\text{KG}}(\kappa)}, \quad (40)$$

<sup>8</sup>Additive Konstanten und multiplikative Faktoren haben keine Auswirkungen auf varianzanalytische Ergebnisse oder Effektstärken: Denn für Zufallsvariablen  $X$  und  $Y$  gilt ( $c$  sei eine Konstante):  $\text{VAR}(cX) = c^2\text{VAR}(X)$ ,  $\text{VAR}(X + c) = \text{VAR}(X)$  und  $\text{VAR}(X + Y) = \text{VAR}(X) + \text{VAR}(Y) + 2\rho(X, Y)$ .

wobei mit den gewichteten Testsummen und Standardabweichungen die folgenden Ausdrücke gemeint sind:

$$T_{IG}(\kappa) = T_{1,IG} + \kappa T_{2,IG} \quad (41)$$

$$T_{KG}(\kappa) = T_{1,KG} + \kappa T_{2,KG} \quad (42)$$

$$SD_{KG}(\kappa) = \sqrt{SD_{1,KG}^2 + \kappa^2 SD_{2,KG}^2 + 2\kappa \cdot \rho(T_{1,KG}, T_{2,KG})}. \quad (43)$$

Die Formel für die gewichtete Standardabweichung folgt aus den Regeln über das Rechnen mit Varianzsummen (vgl. Fußnote 8, S. 141). Dabei bezeichnet  $\rho(T_{1,KG}, T_{2,KG})$  die Kovarianz zwischen den beiden Testsummen in der Kontrollgruppe, die im Allgemeinen nicht verschwindet, da der Test ein eindimensionales Konstrukt misst (und somit die Subkomponenten korreliert sind). Einsetzen von Gl. (41) und (42) in (40) und umschreiben führt zu

$$d(\kappa) = \frac{T_{1,IG} - T_{1,KG}}{SD_{1,KG}} \cdot \frac{SD_{1,KG}}{SD_{KG}(\kappa)} + \kappa \frac{T_{2,IG} - T_{2,KG}}{SD_{2,KG}} \cdot \frac{SD_{2,KG}}{SD_{KG}(\kappa)} \quad (44)$$

$$= d_1 \frac{SD_{1,KG}}{SD_{KG}(\kappa)} + \kappa d_2 \frac{SD_{2,KG}}{SD_{KG}(\kappa)}. \quad (45)$$

Dabei wurden die Standardabweichungen der Kontrollgruppe in den beiden Testhälften sowohl im Zähler als auch im Nenner ergänzt, um die Definitionen der Effektstärken  $d_1$  und  $d_2$  zu nutzen. Durch weiteres Umformen ergibt sich schließlich

$$d(\kappa) = d_1 \frac{1}{\sqrt{1 + \kappa^2 \phi^2 + 2\kappa \theta_1}} + d_2 \frac{1}{\sqrt{1 + \kappa^{-2} \phi^{-2} + 2\kappa^{-1} \theta_2}} \quad (46)$$

mit den Varianzverhältnissen der beiden Testhälften  $\phi$ , sowie den Varianzverhältnissen  $\theta_{1/2}$ , die sich aus der Kovarianz beider Testhälften ergeben,

$$\phi^2 = \frac{SD_{2,KG}^2}{SD_{1,KG}^2} \quad (47)$$

$$\theta_1 = \frac{\rho(T_{1,KG}, T_{2,KG})}{SD_{1,KG}^2} \quad (48)$$

$$\theta_2 = \frac{\rho(T_{1,KG}, T_{2,KG})}{SD_{2,KG}^2}. \quad (49)$$

Gleichung (46) führt die Effektstärke  $d$  auf die beiden Effektstärken der Testhälften 1 und 2 zurück, wobei der Gewichtungsparemeter  $\kappa$  als Veränderliche aufgefasst wird. Die Grenzfälle  $\kappa = 0$  bzw.  $\kappa \rightarrow \infty$  führen konstruktionsgemäß zu den Effektstärken  $d_1$  bzw.  $d_2$ . Da die Testmischungsanalyse eine Erweiterung der Ratekorrektur darstellt, deckt der Wertebereich von  $d(\kappa)$  auch die Effektstärke ab, die unter einer Ratekorrektur erreicht wird. Mit anderen Worten lässt sich eine Gewichtung für nicht rate-korrigierte Werte finden, sodass die Effektstärke ratekorrigierter Testsummenwerte reproduziert wird.

Gleichung (46) bildet die Grundlage zur Anpassung der empirischen Daten an eine analytische Funktion, wobei die Varianzverhältnisse als freie Parameter behandelt werden. Es sei darauf hingewiesen, dass sich die hier vorgestellte Methodik nicht nur auf die Unterscheidung von Testhälften gemäß den Item-Formaten beschränkt, sondern auch für beliebige andere Unterteilungen eines Gesamttests in zwei Subkomponenten gültig ist. Eine Erweiterung der Formel auf die Unterteilung eines Tests in  $2 < k \leq n$  Testhälften ist erstrebenswert und könnte beispielsweise dann Anwendung finden, wenn mehrere Subdimensionen zu einem übergeordneten Konstrukt zusammengeführt werden.

### Reliabilitäten zusammengesetzter Testsummen

Aus ähnlichen Überlegungen kann eine geschlossene analytische Funktion für die Reliabilität ( $\alpha_c$  nach Cronbach) in Abhängigkeit des Gewichtungsparemters  $\kappa$  hergeleitet werden (siehe

Anhang, S. 219), wobei an dieser Stelle darauf verzichtet wurde, die Reliabilität  $\alpha_C(\kappa)$  explizit auf die Reliabilitäten  $\alpha_{C,1}$  und  $\alpha_{C,2}$  der beiden Testhälften zurückzuführen. Die Bestimmung der  $\alpha_C(\kappa)$ -Daten gibt Aufschluss darüber, wie die beiden Messinstrumente (bzw. die beiden Hälften) zu kombinieren sind, um eine optimale Messschärfe des Gesamtinstruments zu erhalten. Diese Frage ist unabhängig von der Vergleichsanalyse beantwortbar und behält für ähnliche Stichproben Gültigkeit. Die  $\alpha_C(\kappa)$ -Daten stellen deshalb eine inhaltlich wertvolle Ergänzung der  $d(\kappa)$ -Daten dar.

## 10.6 Strukturierung der Analysen

In Tabelle 30 sind die statistischen Analyseverfahren aufgelistet, die zur Klärung der Forschungsfragen und -hypothesen angewendet werden.

Die Voranalysen betreffen einerseits die Festlegung der Kovarianzmodelle. Zur Modellspezifizierung der Kovarianzanalysen werden systematisch diejenigen Kovariaten selektiert, die die Voraussetzungen zur Reduktion der Fehlervarianz bzgl. den abhängigen Variablen erfüllen (vgl. Kap. 10.5.1) und aus theoretischer Perspektive sinnvoll erscheinen. Andererseits werden methodische Fragestellungen beantwortet und Ergebnisse von Routine-Prozeduren berichtet (Ausreißerkorrektur, Neubestimmung der Testgüte in der Kern-Stichprobe und Analyse des Betreuerverhaltens). Auch die Ursache für den Studienausstieg (Dropout) von 35 Studierenden zwischen den beiden Testzeitpunkten  $t_0$  und  $t_1$  wird als untergeordnete Fragestellung an dieser Stelle analysiert.

Die Hauptfragestellungen (Hypothesengruppe 1) werden hauptsächlich mithilfe von (univariaten) Kovarianzanalysen untersucht, wobei das Modell nur einen Faktor („Gruppe“) enthält. Aufgrund des Designs mit drei Messzeitpunkten werden typischerweise drei Kovarianzanalysen durchgeführt: Die erste Kovarianzanalyse vergleicht die abhängige Variable zum Post1-Zeitpunkt zwischen den Gruppen. Ein signifikantes Ergebnis des Gruppen-Faktors spricht für die Ablehnung der Nullhypothese für den ersten Interventionszeitraum. In diesem Fall geben die Effektstärke und die deskriptiven Daten einen Aufschluss über die Größe und Richtung des Gruppenunterschieds. Die zweite Kovarianzanalyse vergleicht die abhängige Variable zum Post2-Zeitpunkt zwischen den Gruppen. Infolge des Rotationsdesigns werden hier keine gruppenspezifischen Unterschiede erwartet. Die dritte Kovarianzanalyse vergleicht die Veränderung der abhängigen Variablen zwischen den beiden Post-Zeitpunkten miteinander. Dieser Vergleich kann als Interaktionseffekt zwischen dem Faktor „Gruppe“ und dem Zeitverlauf der abhängigen Variablen während der zweiten Interventionsphase interpretiert werden und führt bei signifikantem Ergebnis zur Ablehnung der Nullhypothese für den zweiten Interventionszeitraum. Wie oben werden in diesem Fall Effektstärken und deskriptive Daten zur Interpretation des Ergebnisses betrachtet. Zusätzlich vermitteln Streudiagramme der Variablen zwischen je zwei Zeitpunkten einen Eindruck über die Variation der Variablen auf Gruppenebene und über extreme Ausprägungsänderungen auf Individualebene. Zusätzlich zu den Kovarianzanalysen werden bei den Leistungsvariablen die normierten Lernzuwächse berechnet, um die erzielten Lernerfolge vor dem Hintergrund bestehender physikdidaktischer Forschungsergebnisse vergleichen zu können (siehe Kap. 10.5.4). Außerdem werden die Ergebnisse von Items, die sich gemäß psychometrischen Kriterien als nicht skalenkonform herausstellten, als Komplementierung der Hauptergebnisse berichtet.

Auch die Hypothesen zur instruktionsbedingten Beanspruchung (Hypothesengruppe 2) sowie zum Manipulation Check (Hyp 3.1) werden varianzanalytisch beantwortet. Die übrigen methodischen Fragestellungen, die sich auf die Bildung der Testsumme als Indikator der Repräsentationskompetenz beziehen, werden durch die in Kap. 10.5.5 erarbeitete Testmischungsanalyse beantwortet.

Durch die Bestimmung der Kovarianzmodelle wird implizit ein Teil der Robustheitsfrage 4.2 beantwortet: Trägt eine Kovariate signifikant zur Varianzaufklärung bei, dann besteht ein Zusammenhang zwischen ihr und der abhängigen Variable. Darüber hinaus bedeutet die Homogenität der Regressionssteigung, die ebenfalls im Rahmen der Voranalyse

geprüft wird, dass sich der Zusammenhang zwischen Kovariate und abhängiger Variable nicht gruppenspezifisch unterscheidet. Mit der Voranalyse wird also implizit überprüft, ob Unterschiedseffekte in den abhängigen Variablen von Persönlichkeitsfaktoren abhängen oder dagegen robust sind. Bezüglich den Robustheitsfragen, die nicht durch die oben genannten Kovarianzanalysen implizit beantwortet werden, beschränkt sich die Analyse auf den ersten Interventionszeitraum.

Zur Bestimmung von Gender-Effekten (FF 4.1) werden zweifaktorielle Varianzanalysen (Faktoren „Gruppe“ und „Geschlecht“) genutzt. Dabei stellen die Terme des Haupteffekts „Geschlecht“ bzw. des Interaktionseffekts „Geschlecht  $\times$  Gruppe“ die relevanten Prüfkriterien dar, die berichtet werden. In diesem Zusammenhang wird das Ergebnis aus Kap. 7.3 aufgegriffen, wonach sich das Item Kon2 als nicht Rasch-konform herausstellte. Das Item verletzte gegen die Stichprobenhomogenität bei Teilung des Datensatzes nach Geschlecht. Die deskriptiven Daten und Lernzuwächse werden für dieses Item einzeln berichtet.

Die Forschungsfrage 4.3 wird regressionsanalytisch beantwortet, wobei der Lernzuwachs, definiert als Differenz zwischen Post1- und Prä-Variablen, als Regressand und die metakognitive Fähigkeit als Regressor in das Modell einfließen und die Regression gruppenweise durchgeführt wird. Damit wird überprüft, ob Unterschiedseffekte von metakognitiven Fähigkeiten abhängen und welchen Einfluss diese auf den Lernzuwachs haben. Auf multiple Regressionsanalysen zur Aufklärung der Beziehungen einer AV und *mehreren* KV wird aufgrund der dafür zu geringen Stichprobengröße verzichtet. Die Forschungsfragen 4.4 und 4.5 werden auf äquivalente Weise beantwortet. Schließlich werden die Ergebnisse der Klausur unter den in Kap. 10.4 genannten Aspekten analysiert.

Tab. 30: Strukturierung der Analysen

Ziel bzw. Hyp / FF und Zielvariable	stat. Analysemethode	Bemerkung	Kapitel
<b>Voranalysen und untergeordnete Fragestellungen</b>			
Untersuchung des Dropouts	<i>t</i> -Test	Analyse systematischer Unterschiede in relevanten Variablen zwischen Weiterstudierenden und Studienabbrechern	11.1.1
Kontrolle des Betreuers	<i>t</i> -Test	Prüfung, ob sich Übungsgruppenleiter zwischen ihren Parallelgruppen unterschiedlich verhielten	11.1.2
Testgüte in Kern-Stichprobe	Skalenanalyse	Re-Analyse der psychometrischen Eigenschaften der Instrumente innerhalb der Kern-Stichprobe	11.1.3
Korrektur von Ausreißern	Ausreißeranalyse	Korrektur von Ausreißerwerten aller Variablen innerhalb der Kern-Stichprobe	11.1.4
Festlegung der ANCOVA-Modelle	NV-Analyse, <i>t</i> -Test, Korrelationsanalyse, Regressionsanalyse	siehe Kap. 10.5.1	11.1.5
<b>Hauptanalyse</b>			
1.1 Repräsentationskompetenz	ANCOVA	Zusatz: Normierter Lernzuwachs Zusatz: Gruppenvergleich auf Itemebene (deskriptiv)	11.2.1
1.2 Konzeptverständnis	ANCOVA	Zusatz: Normierter Lernzuwachs Zusatz: Distraktorenanalyse (deskriptiv)	11.2.2
1.3 Motivation	ANCOVA	Zusatz: Auswertung ausgeschlossener Items (gemäß Kap. 6.4.3)	11.2.3
1.4 Neugierde (Zustand)	ANCOVA	Zusatz: Stabilität der Neugierde als Personeneigenschaft	11.2.4
2.1 Aufgabenschwierigkeit	ANCOVA	-	11.3.1
2.2 Beanspruchung	ANCOVA	-	11.3.2
2.3 Time-on-Task	ANOVA	-	11.3.3
3.1 Experimentbezug	ANOVA	-	11.4.1
3.2-3.4 Repräsentationskompetenz	Testmischungsanalyse	Reliabilität als Funktion der Testsummengewichtung	11.4.2
4.1 siehe 1.1 – 1.4	zweifaktorielle ANCOVA	Zusatz: Item Kon2 einzeln	11.5.1
4.2 siehe 1.1 – 1.4	enthalten in ANCOVA (1.1–1.4)	-	11.5.2
4.3 Leistung (siehe 1.1 und 1.2)	Regressionsanalyse	-	11.5.3
4.4 Leistung (siehe 1.1 und 1.2)	Regressionsanalyse	-	11.5.3
4.5 Leistung (siehe 1.1 und 1.2)	Regressionsanalyse	-	11.5.3
- Klausurergebnis	Korrelationsanalyse, <i>t</i> -Tests	Zusammenhangsanalyse zwischen dem Klausurergebnis und den Variablen der Intervention. Zuordnung einzelner Aufgaben zu Interventionsphasen.	11.6



# Kapitel 11

## Ergebnisse

### Vorbemerkungen zur Notation und Terminologie

- Die deskriptiven Daten (Gruppenmittelwerte und Standardabweichungen) aller Variablen sind in Tab. A-10 im Anhang (S. 216) gesammelt.
- Im Ergebnisbericht werden mit den Indizes 0, 1 und 2 die erfassten Variablen zum Prä-Zeitpunkt  $t_0$ , Post1-Zeitpunkt  $t_1$  und Post2-Zeitpunkt  $t_2$  bezeichnet (z.B. meint  $KiRC_1$  die Variable Repräsentationskompetenz zum Zeitpunkt Post1).
- Die Differenz einer abhängigen Variable  $X$  zwischen den beiden Zeitpunkten  $t_1$  und  $t_2$  ist selbst eine abhängige Variable (vgl. Kap. 10.6) und wird mit  $X_{2-1}$  bezeichnet.
- In Streudiagrammen werden Wertepaare einer abhängigen Variablen  $X$  zu zwei Zeitpunkten auf Personenebene dargestellt. Auf diese Weise können (extreme) Veränderungen von einzelnen Studenten oder Studentengruppen veranschaulicht werden. Die Änderung auf Personenebene zwischen zwei Zeitpunkten wird mit  $\Delta X$  bezeichnet (z.B. entspricht  $\Delta KiRC$  der Änderung der Repräsentationskompetenz zwischen den Zeitpunkten, die im Streudiagramm dargestellt sind).
- Im Symbol- und Abkürzungsverzeichnis (S. 12) sind die Kurzbezeichnungen aller Variablen aufgeführt.
- Die Studierendekohorte, die in der ersten Interventionsphase mit videobasierten Aufgaben arbeitete (Interventionsgruppe) und in der zweiten Phase mit traditionellen Aufgaben (Kontrollgruppe) wird als „IG-KG“-Gruppe bezeichnet, die andere als „KG-IG“-Gruppe. Die Bezeichnungen KG und IG werden verwendet, wenn bzgl. der betreffenden Interventionsphase keine Verwechslungen anzunehmen sind.
- Die Bezeichnung Kern-Stichprobe meint die Hauptfach-Studierenden, die zum Zeitpunkt  $t_0$  und  $t_1$  in parallel instruierten Übungsgruppen anwesend waren (vgl. Kap. 10.2.1). Alle Vergleichsanalysen, die mehrere Zeitpunkte einschließen, beruhen stets auf verbundenen Stichproben; ein Bias aufgrund von Studienausstieg wird somit vermieden.
- Die (Ko)Varianz- und Regressionsanalysen wurden mit SPSS durchgeführt, Testmischungsanalysen mit EXCEL und die Fitkurven wurden mit MATLAB bestimmt. Output-Dateien sind im digitalen Anhang einzusehen; die Tabellen und Abbildungen im Fließtext zeigen aus Übersichtsgründen nur Endergebnisse oder Resultate, die zur Beantwortung der Fragestellungen relevant sind.

## 11.1 Voranalysen und untergeordnete Fragestellungen

### 11.1.1 Unterschiede zwischen Studienabbrechern und Weiterstudierenden

Zunächst wurde versucht, eine plausible Ursache für den Schwund von 35 Hauptfach-Studierenden zwischen den Zeitpunkten  $t_0$  und  $t_1$  auf Grundlage der zur Verfügung stehenden Daten zu finden. Dazu wurden Vorleistungen und relevante Indikatoren der Studienmotivation (Interesse, Selbstkonzept, Neugierde), empfundene Belastung sowie kognitive Fähigkeiten zwischen den Weiterstudierenden (Kern-Stichprobe) und der Abbrechergruppe miteinander verglichen, siehe Tab. 31. Die Daten zeigen, dass die Abbrechergruppe durchschnittlich niedrigere Abiturnoten ( $p < 0.01$ ,  $d = 0.61$ ) und Mathematik-Vorleistungen ( $p < 0.001$ ,  $d = 0.77$ ) als die Kern-Stichprobe aufweist. Beide Unterschiede sind hochsignifikant und von großer praktischer Bedeutung. Aufgrund der schwächeren Vorleistung ist es nicht verwunderlich, dass die Studienabbrecher die bis zum Zeitpunkt der Erhebung bearbeiteten Aufgaben<sup>1</sup> schwerer einschätzten als die Kern-Stichprobe ( $p < 0.05$ ,  $d = 0.44$ ) und in dem Test zur Repräsentationskompetenz schlechter abschnitt ( $p < 0.05$ ,  $d = 0.51$ ). Zudem sind auch Interesse ( $p < 0.01$ ,  $d = 0.66$ ) und physikbezogenes Selbstkonzept ( $p < 0.01$ ,  $d = 0.61$ ) in der Abbrechergruppe signifikant geringer als bei den Weiterstudierenden.

Außerdem wurde festgestellt, dass die Dropout-Quote in der Kontrollgruppe 25% und in der Interventionsgruppe 34% beträgt. Dieser Unterschied ist nicht signifikant ( $t(118) = 1.10$ ,  $p = 0.27$ ).

**Tab. 31:** Vergleich zwischen Studienaussteigern und Weiterstudierenden

Variable	Kern-Stichprobe ( $N = 76$ )		Dropout ( $N = 35$ )		$t$	$p$	$d$
	$M$	$SD$	$M$	$SD$			
Abiturnote <sup>a</sup>	2.1	.7	2.5	.4	2.88	.005	.61
Mathe-Vorleistung <sup>b</sup>	11.7	2.3	9.9	2.1	3.62	.000	.77
Physik-Vorleistung <sup>b</sup>	11.6	3.0	11.0	2.1	.91	.364	-
Interesse	.75	.11	.67	.12	3.23	.001	.66
Selbstkonzept	.61	.11	.53	.15	3.00	.004	.61
Beanspruchung	.57	.15	.56	.18	.25	.805	-
Aufgabenschwierigkeit	.43	.17	.50	.18	2.13	.036	.44
Neugierde (NP)	.79	.10	.76	.13	.83	.411	-
Repräsentationskomp.	.67	.20	.54	.20	2.49	.014	.51
Konzeptverständnis	.54	.30	.49	.33	.96	.340	-
<hr/>							
Geschlecht $\sigma$ (%)	73.7		82.9				
Physik LK (%)	64.5		65.7				
Mathematik LK (%)	78.9		74.3				

*Bemerkung:* Die Daten beziehen sich auf den ersten Testzeitpunkt  $t_0$ .

<sup>a</sup>Wertebereich von 1.0 (sehr gut) bis 4.0 (ausreichend). <sup>b</sup>Wertebereich von 0 (ungenügend) bis 15 (sehr gut).

### 11.1.2 Verhalten des Übungsleiters in den Gruppen

Vor der Hauptanalyse der Daten wurden die Beurteilungen der Übungsgruppenleiter durch die Studierenden ausgewertet, um unterschiedliches Verhalten einzelner Betreuer in KG und IG auszuschließen. Die Studierenden bewerteten am Ende der jeweiligen Interventionsphase das Betreuerverhalten retrospektiv (Engagement und Motivationsbereitschaft). Wie die Auswertung der Ergebnisse in Tabelle 32 zeigt, unterscheiden sich die Beurteilungen der

<sup>1</sup>Zum Zeitpunkt der Datenerhebung wurde bereits ein Mathematik-Übungsblatt und ein erstes Experimentalphysik-Übungsblatt (ohne videobasierte Aufgaben) bearbeitet, vergleiche Kap. 10.2.

Übungsgruppenleiter nicht zwischen den Interventions- und Kontrollgruppen. Dies gilt für die Parallelgruppen jedes einzelnen Betreuers und für die Gesamtbeurteilung der Betreuer (Mittelwerte über alle Studierenden) jeweils in beiden Interventionsphasen.

**Tab. 32:** Vergleich der Einschätzungen des Übungsleiters zwischen den Gruppen

Übungsleiter (Betreuer)	Erste Phase					Zweite Phase				
	KG		IG		<i>p</i>	KG		IG		<i>p</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
B1	.82	.09	.74	.03	.30	.82	.03	.83	.13	.95
B2	.65	.15	.69	.14	.50	.72	.13	.69	.17	.42
B4	.68	.15	.78	.09	.18	.83	.20	.74	.04	.39
B5	.74	.09	.66	.15	.16	.69	.16	.69	.11	.95
Gesamt	.71	.14	.69	.14	.93	.73	.16	.73	.14	.84

*Bemerkungen:* Gruppenvergleiche anhand zweiseitiger *t*-Tests. Auf Bericht der Teststatistik *t* wird verzichtet, da  $p \gg 0.05$ .

### 11.1.3 Kenngrößen der Instrumente innerhalb der Kern-Stichprobe

Zur Kontrolle der zeitlichen Stabilität der psychometrischen Skaleneigenschaften wurden die wichtigsten Kenngrößen innerhalb der Kern-Stichprobe zu allen drei Testzeitpunkten erneut bestimmt, siehe Tab. 33. Die Itemzusammenstellung gemäß der Skalenbildung (Kap. 6) bleibt dabei unverändert, d.h. die faktorielle Struktur der Instrumente wurde nicht erneut überprüft. Die Motivations-Subskalen (IE, SK, dA und Atnm) zeigen sich zu allen drei Testzeitpunkten als zufriedenstellend konsistent ( $\alpha_c = 0.70 - 0.81$ ); die RA-Skala weist darüber hinaus eine sehr gute Konsistenz auf ( $\alpha_c = 0.83 - 0.91$ ). Die Konsistenz der Exp-Skala verbessert sich im Laufe der Intervention und liegt im zufriedenstellenden Bereich. Auch die Reliabilitäten der Neugierde- und Belastungsskalen sind durchgehend gut bis exzellent. Das KiRC-Inventar zur Erfassung der Repräsentationskompetenz besitzt ebenfalls gute Reliabilitäten; die Indizes des Konzept-Tests sind hingegen — wie schon innerhalb der Eich-Stichprobe — kritisch (vgl. Kap. 6.4.5). Ein Vergleich der in Tab. 33 angegebenen Reliabilitäten zum Prä-Zeitpunkt zu den korrespondierenden Werten innerhalb der Eich-Stichprobe (Tab. 22, S. 101) ergibt eine maximale Differenz in den  $\alpha_c$ -Werten von 0.01 (KiRC). Zusammengefasst sprechen die Ergebnisse also dafür, dass die Skalenstrukturen erstens zeitlich stabil sind und sich zweitens zwischen der Eich- und der Kern-Stichprobe nicht wesentlich unterscheiden.

**Tab. 33:** Skalenkenngrößen innerhalb der Kern-Stichprobe zu verschiedenen Testzeitpunkten

Skala	Prä-Zeitpunkt			Post1-Zeitpunkt			Post2-Zeitpunkt		
	<i>M/P</i>	<i>D</i>	$\alpha_c$	<i>M/P</i>	<i>D</i>	$\alpha_c$	<i>M/P</i>	<i>D</i>	$\alpha_c$
PSQ	.57	.39	.81	.67	.38	.92	.65	.38	.84
CL	.43	.43	.93	.55	.31	.81	.51	.25	.79
NP	.79	.27	.80	.75	.27	.85	.74	.29	.87
NZ	-	-	-	.62	.34	.82	.62	.32	.82
Exp	.48	.37	.72	.49	.40	.75	.51	.43	.79
RA	.65	.33	.83	.62	.35	.90	.62	.34	.91
IE	.75	.28	.76	.79	.32	.81	.71	.31	.71
SK	.61	.28	.74	.62	.33	.74	.66	.32	.73
dA	.73	.23	.77	.69	.27	.80	.68	.28	.72
Atnm	.69	.38	.71	.60	.41	.70	.63	.41	.77
KiRC	.67	.52	.83	.80	.42	.76	-	-	-
Kon	.54	.74	.51	.67	.64	.38	.73	.63	.57

*Bemerkungen:* *M/P* = Skalenmittelwert, *D* = mittlerer Diskriminationsindex,  $\alpha_c$  = Reliabilität (nach Cronbach). Für Abkürzungen der Skalen siehe globales Symbol- und Abkürzungsverzeichnis.

### 11.1.4 Ausreißeranalysen

Die Skalenmittelwerte aller Studenten wurden variablenweise — wie in Kap. 6.3 beschrieben — systematisch nach Ausreißern überprüft. Im gesamten Datensatz (41 verschiedene Variablen mit insgesamt 2675 Einträgen) traten 7 Ausreißer auf (0.3%), die auf den jeweiligen Skalenmittelwert plus / minus (je nach Richtung des Ausreißers) 3-fache Standardabweichung korrigiert wurden.

### 11.1.5 Festlegung der ANCOVA-Modelle

Wie in den methodischen Grundlagen zur Kovarianzanalyse erläutert, werden nur diejenigen Variablen als Kovariaten in das Modell einbezogen, die mit der abhängigen Variablen korrelieren und theoretisch sinnvoll sind. Außerdem müssen Kovariaten zum Prä-Zeitpunkt erhoben worden sein. Mit Blick auf den umfangreichen Variablenplan und die Modellkomplexität wird zunächst eine Auswahl getroffen, um die Anzahl zu prüfender Kovariaten gering zu halten. Die aus theoretischer Sicht bedeutendsten Kovariaten sind

- die Prä-Variablen  $X_0$ , die vermutlich in Zusammenhang mit den zugehörigen abhängigen Variablen  $X_1$ ,  $X_2$  und  $X_{2-1}$  steht,
- die Vorleistungen (Abinote, Mathe- und Physikpunktzahl) zur Reduktion der Fehlervarianz bzgl. den kognitiven Variablen KiRC und Kon,
- die zum Prä-Zeitpunkt erhobenen Belastungsvariablen  $PSQ_0$  und  $CL_0$  für die kognitiven Variablen KiRC, Kon sowie für die kognitiv-affektive Motivationskomponente SK,
- Neugierde als Personeneigenschaft (NP) für Neugierde als aktueller Zustand (NZ)<sup>2</sup>.

Die Zusammenhänge zwischen diesen (potentiellen) Kovariaten und den abhängigen Variablen werden in einem ersten Schritt korrelationsanalytisch überprüft. Kovariaten, die diese notwendige Bedingung zur Aufnahme in das jeweilige Kovarianzmodell nicht erfüllen, werden nicht weiter auf Modelleignung getestet. Um festzustellen, ob sich die Untersuchungsgruppen in ihren Ausprägungen der Kovariaten bzw. der Untersuchungsvariablen vor der Intervention unterscheiden, werden in einem zweiten Schritt unabhängige *t*-Tests durchgeführt. Für die endgültige Aufnahme der Kovariaten in das Modell wird schließlich die Homogenität der Regressionsteigungen überprüft. Da die Normalverteilung der Variablen bereits innerhalb der Eich-Stichprobe nachgewiesen wurde (Kap. 6.5) wird auf die Darstellung der Überprüfung innerhalb der Untergruppen verzichtet. Gegen geringfügige Abweichungen zur Normalverteilung, die aus statistischen Gründen auftreten könnten, ist das Verfahren ohnehin robust (vgl. Kap. 10.5.1).

### Korrelationsanalysen

Aus der ersten Korrelationstabelle (Tab. 34) geht hervor, dass die betrachteten Kovariaten keinen Zusammenhang mit den Post2-Variablen und den Differenzvariablen (Post2-Post1) aufweisen. Die Mathematik-, nicht aber die Physik-Vorleistung, steht mit den Leistungen des Konzept- und Repräsentationstests zum ersten Testzeitpunkt in Verbindung. Die Abiturnote korreliert ebenso mit dem Repräsentationstest zum ersten Post-Zeitpunkt, wobei sich das negative Vorzeichen aus der Polung der Abiturnote ergibt (bessere Leistung entspricht geringerer Note). Außerdem besteht ein umgekehrter Zusammenhang zwischen der Aufgabenschwierigkeit und dem Selbstkonzept zum ersten Postzeitpunkt, d.h. hohe Schwierigkeiten stehen mit geringen Ausprägungen des Selbstkonzepts in Verbindung.

Wie die Ergebnisse der zweiten Korrelationstabelle (Tab. 35) zeigen, bestehen starke Zusammenhänge zwischen allen wiederholt gemessenen Variablen zu den Zeitpunkten Prä und Post1, Prä und Post2 sowie Post1 und Post2. Letztere sind für die Modellspezifikation

<sup>2</sup>Die Variable NP bleibt infolge des zu kleinen NP-Datensatzes unberücksichtigt, vgl. 10.4.

**Tab. 34:** Korrelationen zwischen KV und AV

KV	SK <sub>1</sub>	Kon <sub>1</sub>	KiRC <sub>1</sub>	SK <sub>2</sub>	Kon <sub>2</sub>	SK <sub>2-1</sub>	Kon <sub>2-1</sub>
Abinote	-.20	-.21	-.33*	-.11	-.06	-.05	.19
Mathe-Vorleistung	.07	.25*	.38*	.00	.06	-.06	-.22
Physik-Vorleistung	.07	.14	.12	.03	.03	.01	-.04
PSQ <sub>0</sub>	-.11	.05	.11	-.04	.07	.04	.03
CL <sub>0</sub>	-.34*	-.14	-.12	-.06	-.03	.07	.04

\* $p < \alpha$  gemäß Benjamini-Hochberg Korrektur.

nicht relevant<sup>3</sup>, werden aber für Interpretationszwecke dennoch berichtet. Zwischen den Prä-Variablen und den Änderungsvariablen (Post2-Post1) besteht kein Zusammenhang, d.h. die Studierenden haben unabhängig ihrer Vorleistung die gleichen Entwicklungschancen. Auffällig sind die sehr hohen Zusammenhänge zwischen den wiederholt gemessenen Neugierde-Variablen (als Personeneigenschaft). Die intraindividuelle zeitliche Stabilität lässt zu Recht eine Identifikation dieser Neugierde-Komponente als „Trait“ zu.

**Tab. 35:** Korrelationen der AV zwischen den Testzeitpunkten

AV (X)	X <sub>0</sub> , X <sub>1</sub>	X <sub>0</sub> , X <sub>2</sub>	X <sub>0</sub> , X <sub>2-1</sub>	X <sub>1</sub> , X <sub>2</sub>
PSQ	.52*	.49*	.00	.70*
CL	.60*	.51*	-.08	.69*
NZ	-	-	-	.65*
RA	.68*	.70*	-.09	.75*
IE	.59*	.74*	.20	.72*
SK	.63*	.42*	-.05	.73*
dA	.61*	.66*	.00	.71*
Exp	.59*	.68*	.07	.46*
Atnm	.29*	.29*	-.02	.50*
KiRC	.72*	-	-	-
Kon	.65*	.56*	-.04	.65*
NP	.84*	.86*	.19	.85*

\* $p < \alpha$ . Das Signifikanzniveau wurde nach dem Benjamini-Hochberg-Verfahren bestimmt (siehe Fußnote 2, S. 92).

### Vergleichbarkeit der Gruppen zum Prä-Zeitpunkt

Wie in Tab. 36 dargestellt ist, zeigen sich mit Ausnahme des Interesses und des Experimentbezugs keine signifikanten Gruppenunterschiede in den erhobenen Variablen zum Prä-Zeitpunkt. Es kann insbesondere davon ausgegangen werden, dass die Kontroll- und Interventionsgruppe in ihrer Vorleistung (fachbezogene Schulnoten, Abiturnote) und in der Ausprägung der Repräsentationskompetenz / des Konzeptverständnisses vor Beginn der Studie vergleichbar sind. Das geringfügig höhere Interesse / Engagement (IE) der Interventionsgruppe im Vergleich zur Kontrollgruppe ist entweder auf eine unglückliche Randomisierung oder auf eine systematische Beeinflussung der Interventionsgruppe vor dem Messzeitpunkt zurückzuführen: Die beiden Untersuchungsgruppen waren vor dem Testzeitpunkt — also während der Einführungsphase — schon darüber informiert, ob sie mit videobasierten Aufgaben arbeiten werden (siehe Kap. 10.2.2). Es ist nicht auszuschließen, dass sich dies auf das Interesse / Engagement auswirkte. Wie sich nach Prüfung der anderen Voraussetzungskriterien herausstellt, eignet sich diese Kovariate prinzipiell zur Reduktion der Fehlervarianz bezüglich des Interesses zum Post1- und Post2-Zeitpunkt (vgl. Tab. A-9). Deshalb wurde entschieden, IE<sub>0</sub> als Kovariate beizubehalten, aber die Analyse zudem auch ohne Kovariate durchzuführen. Die Gruppenunterschiede

<sup>3</sup>Nur zum Prä-Zeitpunkt erhobene Variablen dürfen zur Reduktion der Fehlervarianz in das Modell einbezogen werden.

bzgl. des Experimentbezugs (Manipulation-Check-Variable) sind mit dem gleichen Argument begründbar und bekräftigen den Verdacht, dass es sich um eine systematische Beeinflussung handelte. Auch für die Kovarianzanalysen bzgl. den Exp-Variablen werden die Modelle sowohl mit als auch ohne die Kovariate  $Exp_0$  definiert.

**Tab. 36:** Homogenität der Untersuchungsgruppen

AV / KV	KG ( $N = 40$ )		IG ( $N = 36$ )		$t$	$p$	$d$
	$M$	$SD$	$M$	$SD$			
Abiturnote <sup>a</sup>	2.2	.9	2.0	.7	1.66	.101	-
Mathe-Vorleistung <sup>b</sup>	11.5	2.2	11.9	2.4	.90	.371	-
Physik-Vorleistung <sup>b</sup>	11.6	2.7	11.7	3.0	.18	.862	-
PSQ <sub>0</sub>	.55	.17	.59	.13	1.23	.224	-
CL <sub>0</sub>	.43	.16	.42	.19	.11	.915	-
NP <sub>0</sub>	.79	.11	.79	.10	.21	.839	-
RA <sub>0</sub>	.63	.11	.68	.14	1.68	.100	-
IE <sub>0</sub>	.72	.08	.78	.12	2.28	.026	.63
SK <sub>0</sub>	.60	.12	.62	.10	.61	.541	-
dA <sub>0</sub>	.73	.08	.74	.10	.33	.743	-
Exp <sub>0</sub>	.45	.14	.51	.15	2.10	.041	.51
Atnm <sub>0</sub>	.68	.15	.72	.15	1.12	.265	-
KiRC <sub>0</sub>	.67	.18	.67	.22	.15	.973	-
Kon <sub>0</sub>	.54	.30	.53	.29	.52	.861	-
Geschlecht $\sigma$ (%)	72.5		75.0				
Physik LK (%)	57.5		72.2				
Mathematik LK (%)	82.0		75.0				

*Bemerkung:* Vergleich durch zweiseitige  $t$ -Tests ( $df = 74$ ).

<sup>a</sup>Wertebereich von 1.0 (sehr gut) bis 4.0 (ausreichend). <sup>b</sup>Wertebereich von 0 (ungenügend) bis 15 (sehr gut).

### Homogenität der Regressionssteigungen

Zur Prüfung der Regressionssteigungen auf Homogenität wurde der Interaktionseffekt zwischen den potentiellen Kovariaten und den abhängigen Variablen regressionsanalytisch auf Signifikanz überprüft. Die Ergebnisse befinden sich in Tab. A-8 im Anhang (auf S. 215), wobei in der Tabelle noch weitere Interaktionseffekte zusammengefasst sind, die an späterer Stelle benötigt werden. Wie der Tabelle entnommen werden kann, liegen zwei Verstöße gegen die Forderung homogener Regressionssteigungen vor: Die Kovariate PSQ<sub>0</sub> eignet sich nicht zur Varianzaufklärung bezüglich den abhängigen Variablen PSQ<sub>1</sub> und PSQ<sub>2</sub> und wird folglich aus den Modellen ausgeschlossen. Mit Blick auf die Komplexität des Variablenplans wird an dieser Stelle darauf verzichtet, Interaktionen höherer Ordnung zu berücksichtigen (und ggf. zu korrigieren). Der Anteil der Fehlervarianz, der durch den Ausschluss von PSQ<sub>0</sub> unaufgeklärt bleibt, beträgt etwa 25% (vgl. Tab. 35).

Als Zwischenergebnis der durchgeführten Voranalysen zur Modellspezifikation ergeben sich die Kovarianzmodelle gemäß Tab. A-9 (siehe Anhang, S. 215).

## 11.2 Ergebnisse zur Hauptfragestellung 1: Wirksamkeit

Gemäß der Strukturierung der Hauptanalysen und der Festlegung der Modellspezifikationen wurden für jede Zielvariable in der Regel drei Kovarianzanalysen mit den abhängigen Variablen  $X_1$ ,  $X_2$  sowie  $X_{2-1}$  (= Änderung zwischen  $t_1$  und  $t_2$ ) durchgeführt — Ausnahmen bilden KiRC und NZ, da diese Variablen nicht zu allen Zeitpunkten erhoben wurden. Diese Ergebnisse sind in Tabelle 37 dargestellt und werden im Folgenden variablenweise erläutert.

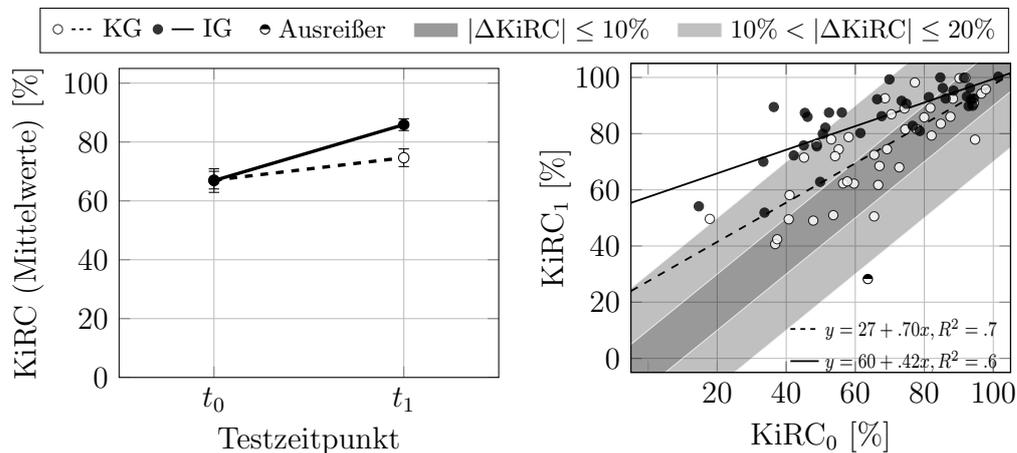
**Tab. 37:** Ergebnisse der Kovarianzanalysen zu den Hauptfragestellungen

Hyp	AV	$df_F$	KV / Faktor	$F$ -Wert	$p$ -Wert	$\eta_p^2$	$\eta^2$	$\omega^2$	$d$		
1.1	KiRC <sub>1</sub>	70	Gruppe	28.42	< .001	.29	.13	.11	.72		
			Abinote	3.52	.065	.05	.02				
			Ma.-Vorleistung	3.42	.068	.05	.02				
			KiRC <sub>0</sub>	92.69	< .001	.57	.47				
1.2	Kon <sub>1</sub>	65	Gruppe	5.29	.025	.08	.04	.03	.34		
			M-Vorleistung	3.07	.084	.05	.03				
			Kon <sub>0</sub>	48.87	< .001	.43	.40				
	Kon <sub>2</sub>	55	Gruppe	.35	.554	.01	.00	-	-		
Kon <sub>0</sub>			25.35	< .001	.32	.24					
	Kon <sub>2-1</sub>	59	Gruppe	1.71	.196	.03	.03	-	-		
1.3	RA <sub>1</sub>	72	Gruppe	6.34	.014	.08	.05	.03	.37		
			RA <sub>0</sub>	58.47	< .001	.43	.42				
	RA <sub>2</sub>	60	Gruppe	.67	.416	.01	.00	-	-		
			RA <sub>0</sub>	51.58	< .001	.46	.46				
	RA <sub>2-1</sub>	60	Gruppe	3.64	.061	.06	.06	(.04)	(.41)		
	SK <sub>1</sub>	69	Gruppe	2.64	.109	.04	.02	-	-		
			CL <sub>0</sub>	.81	.370	.01	.01				
			SK <sub>0</sub>	33.12	< .001	.32	.31				
	SK <sub>2</sub>	58	Gruppe	.01	.931	.00	.00	-	-		
			SK <sub>0</sub>	11.69	.001	.17	.17				
	SK <sub>2-1</sub>	59	Gruppe	2.17	.146	.04	.04	-	-		
	IE <sub>1</sub>	71	Gruppe	.05	.826	.00	.00	-	-		
			IE <sub>0</sub>	35.91	< .001	.34	.34				
IE <sub>2</sub>	58	Gruppe	.05	.828	.00	.00	-	-			
		IE <sub>0</sub>	35.91	< .001	.34	.34					
IE <sub>2-1</sub>	59	Gruppe	.00	.998	.00	.00	-	-			
dA <sub>1</sub>	71	Gruppe	3.60	.062	.05	.03	(.02)	(.25)			
		dA <sub>0</sub>	47.71	< .001	.40	.39					
dA <sub>2</sub>	56	Gruppe	3.03	.087	.05	.03	-	-			
		dA <sub>0</sub>	49.97	< .001	.47	.46					
dA <sub>2-1</sub>	59	Gruppe	2.35	.131	.04	.04	-	-			
Atnm <sub>1</sub>	73	Gruppe	.18	.671	.00	.00	-	-			
		Atnm <sub>0</sub>	6.77	.011	.08	.07					
Atnm <sub>2</sub>	58	Gruppe	.02	.894	.00	.00	-	-			
		Atnm <sub>0</sub>	6.28	.015	.10	.10					
Atnm <sub>2-1</sub>	59	Gruppe	.09	.763	.00	.00	-	-			
1.4	NZ <sub>1</sub>	73	Gruppe	3.82	.054	.05	.05	.04	.39		
			NZ <sub>2</sub>	61	Gruppe	1.41	.240	.02	.02	-	-
			NZ <sub>2-1</sub>	60	Gruppe	1.42	.238	.02	.02	-	-

*Bemerkungen:* AV = abhängige Variable,  $df_F$  = Fehlerfreiheitsgrade, KV = Kovariate (konstrastarme Darstellung). Der Zählerfreiheitsgrad beträgt jeweils 1. Die Effektstärkemaße  $\omega^2$  und  $d$  werden nur für Gruppeneffekte berichtet. Effektstärken werden nicht berichtet (-), wenn  $p \gg 0.05$ ; bei Tendenzen ( $0.05 < p < 0.10$ ) sind die Effektstärken in Klammern angegeben.

### 11.2.1 Repräsentationskompetenz

Notwendig für den Nachweis eines Effekts des Instruktionsmaterials auf die Repräsentationskompetenz (Hyp 1.1) ist ein signifikanter Haupteffekt unter dem Faktor „Gruppe“. Die Analyse beruht auf der optimierten Kern-Itemgruppe des KiRC-Instruments. Abbildung 26 zeigt den zeitlichen Verlauf der Gruppenmittelwerte und das Streudiagramm (Prä/Post-Datenpunkte einzelner Studierender). Ein multivariater Ausreißer wurde aus der Analyse ausgeschlossen.



**Abb. 26:** Unadjustierte Gruppenmittelwerte und Standardfehler bzgl. des KiRC-Tests (Repräsentationskompetenz) zu den beiden Erhebungszeitpunkten  $t_0$  und  $t_1$  (links) sowie Streudiagramm zum Lernzuwachs mit Regressionsgeraden in den Untergruppen (rechts). Die gefärbten Flächen indizieren Bereiche mit Lernzuwachs  $|\Delta\text{KiRC}| \leq 10\%$  bzw.  $10\% < |\Delta\text{KiRC}| \leq 20\%$ . Dabei entsprechen 10 Prozentpunkte 2.2 Testsummenpunkten.

#### Ergebnisse der Kovarianzanalyse (optimierte Kern-Itemgruppe)

Gemäß der Kovarianzanalyse unterscheidet sich die korrigierte Repräsentationskompetenz zum Post1-Zeitpunkt signifikant zwischen den Gruppen ( $F(1, 70) = 28.42, p < 0.001$ ), vgl. Tab. 37. Die Effektstärke nach Cohen ist nach der in Kap. 10.5.2 gegebenen Klassifizierung als mittelgroß bis groß einzustufen ( $d = 0.72$ ). Im Modell wurden die Prä-Werte, die Abiturnote und die Mathematik-Vorleistung berücksichtigt, die allesamt signifikant zur Reduktion der Fehlervarianz beitragen. Die Prä-Leistung  $\text{KiRC}_0$  trägt am stärksten zur Varianzaufklärung ( $\eta^2 = 0.47$ ) bei. Zur Einschätzung, wie stark die Berücksichtigung von Kovariaten das Ergebnis beeinflusst, wurde die Analyse zusätzlich ohne Kovariaten durchgeführt. In diesem Fall fällt der  $F$ -Bruch geringer aus (da weniger Fehlervarianz korrigiert wurde) und der  $p$ -Wert ist entsprechend größer ( $F(1, 73) = 10.69, p = 0.002, d = 0.72$ ). Das Ergebnis bleibt hochsignifikant mit gleich großer Effektstärke.

Aus dem zeitlichen Verlauf der Gruppenmittelwerte sowie dem Streudiagramm (Abb. 26) geht hervor, dass beide Gruppen einen Lernzuwachs verzeichnen — die meisten Datenpunkte liegen oberhalb der Winkelhalbierenden. Während sich die schon zum Prä-Zeitpunkt leistungsstarken Studenten infolge des Deckeneffekts nur gering steigern können, weisen vor allem die Studierenden der Interventionsgruppe im unteren bis mittleren Fähigkeitsbereich deutliche Lernzuwächse auf, teils sogar außerhalb des schraffierten 20%-Bereichs. Der insgesamt höhere Lernzuwachs der Interventionsgruppe gegenüber der Kontrollgruppe wird neben

dem varianzanalytischen Ergebnis durch die Unterschiede der Hake-Indizes quantifiziert:

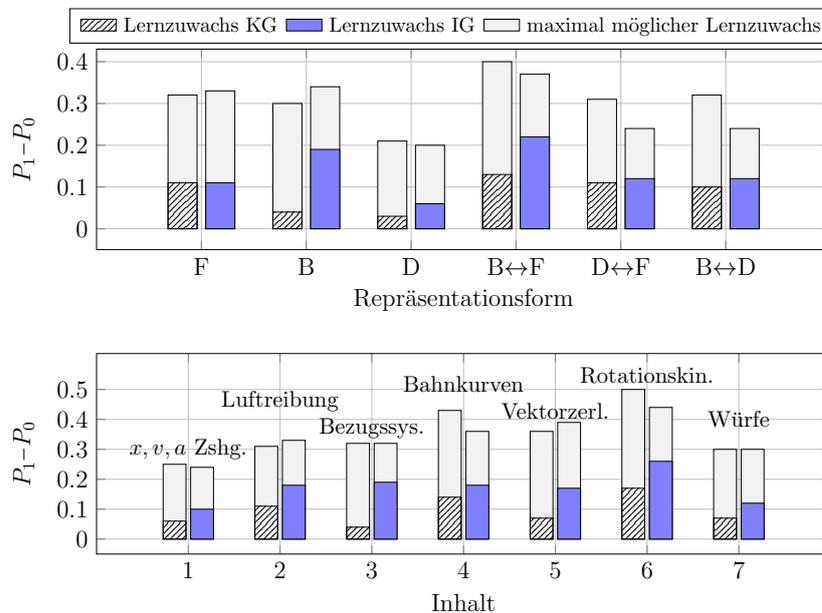
$$g_{H,IG} \stackrel{Gl. (36)}{=} \frac{P(t_1) - P(t_0)}{1 - P(t_0)} \stackrel{Tab. A-10}{=} \frac{0.86 - 0.67}{1 - 0.67} = 0.57$$

$$g_{H,KG} = \frac{0.75 - 0.67}{1 - 0.67} = 0.23$$

### Lernzuwachs auf Itemebene

Um festzustellen, auf welche Items der größere Lernzuwachs der IG zurückzuführen ist, sind die deskriptiven Statistiken aller Items in Tab. A-11 dargestellt (Anhang, S. 217). Zunächst ist festzustellen, dass sich beide Gruppen bei den meisten Items im Laufe der Zeit verbesserten (Spalte  $P_j(t_1) - P_j(t_0)$ ). Die in der Spalte  $\Delta_{IG-KG}$  angegebenen Lernzuwachsdifferenzen geben an, wie stark sich der Lernzuwachs zwischen den Gruppen auf Itemebene unterscheidet. Die geringe Anzahl negativer Werte bedeutet, dass IG bei fast allen Items einen vergleichbaren oder größeren Lernzuwachs als die KG verzeichnete.

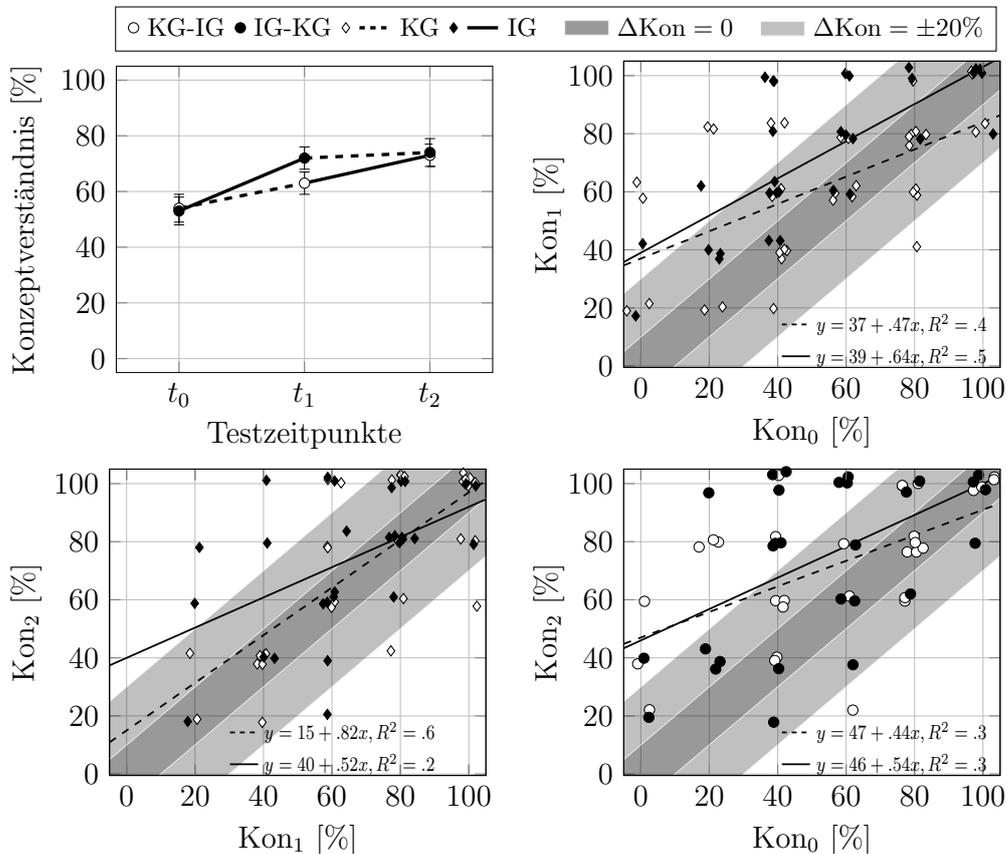
Eine systematische Differenzierung der Items hinsichtlich der dargebotenen Repräsentationsformen und der thematisierten Lerninhalte ist durch die normativ getroffene Zuordnung in Tab. 4 gegeben. Auf dieser Grundlage wurden die mittleren Lernzuwächse in den jeweiligen Itemgruppen in beiden Untersuchungsgruppen bestimmt und in Abb. 27 dargestellt. Wie festgestellt werden kann, zeigt die IG besonders im bildhaften sowie beim Wechsel zwischen bildhaften und algebraisch-formalen Repräsentationsformat einen größeren Lernzuwachs als die KG. Insbesondere hervorzuheben sind in diesem Zusammenhang die Items K3, K5, K11, K22, K23 und K25, bei denen die Differenz mindestens 10% beträgt (vgl. Tab. A-11). Inhaltlich ergeben sich die größten Differenzen bezüglich des Umgangs mit Bezugssystemen und der Komponentenerlegung von Vektoren. Bei allen drei Items, die Bezugssysteme thematisieren (K6, K9 und K22), zeigen sich Lernzuwachs-Differenzen  $\Delta_{IG-KG} \geq 0.10$ . Der größte Unterschied in der Inhaltsgruppe 5 (Vektorzerlegung) beruht auf Item K11 ( $\Delta_{IG-KG} = 0.28$ ), bei dem der Richtungscharakter einer beschleunigten Kreisbewegung beurteilt werden muss. Im Diskussionskapitel werden diese Ergebnisse auf die Art der kognitiven Anforderungen der videobasierten Aufgaben zurückgeführt (Kap. 12.1.1).



**Abb. 27:** Mittlere Lernzuwächse  $P_1 - P_0$  und möglicher Lernzuwachs  $1 - P_0$  für Itemgruppen, die nach Repräsentationsformen und Inhalten gemäß Tab. 4 zusammengefasst sind.

### 11.2.2 Konzeptverständnis

Zur Prüfung der Hypothese 1.2, dass Videoanalyse-Aufgaben das konzeptionelle Verständnis besser fördern als traditionelle Aufgaben, werden die Leistungsverläufe zwischen Interventions- und Kontrollgruppe miteinander verglichen.



**Abb. 28:** Unadjustierte Gruppenmittelwerte und Standardfehler bzgl. des Konzept-Tests zu Bewegungen sowie Streudiagramme zum Lernzuwachs zwischen je zwei Zeitpunkten mit Regressionsgeraden in den Untergruppen. Die gefärbten Flächen indizieren Bereiche ohne Lernzuwachs ( $\Delta\text{Kon} = 0$ ) bzw. mit Änderung  $\Delta\text{Kon} = \pm 20\%$  (Letzteres entspricht einem Testsummenpunkt). Um ein Aufeinanderliegen der Datenpunkte im Streudiagramm zu vermeiden, wurde den Daten ein geringes Rauschen hinzugefügt.

#### Ergebnisse der Kovarianzanalyse

Im Gruppenvergleich adjustierter Post-Werte ( $\text{Kon}_1$ ) zeigt sich ein signifikanter Unterschied bzgl. der Bedingung „Gruppe“ kleinen Effekts ( $F(1, 65) = 5.29$ ,  $p = 0.025$ ,  $d = 0.34$ ). Wie der deskriptive Vergleich der Gruppenmittelwerte zeigt, erfährt die IG in der ersten Interventionsphase einen höheren Lernzuwachs als die KG, vgl. Abb. 28 (oben links). Beide Untersuchungsgruppen verzeichnen einen Lernzuwachs, der sich hinsichtlich der Hake-Indizes unterscheidet ( $g_{H,KG} = 0.19$ ,  $g_{H,IG} = 0.40$ ; berechnet aus Werten von Tab. A-10).

Nach der zweiten Interventionsphase unterscheiden sich die Gruppenmittelwerte  $\text{Kon}_2$  nicht signifikant voneinander ( $F(1, 55) = 0.35$ ,  $p = 0.554$ ). Auch der Lernzuwachs zwischen den beiden Post-Zeitpunkten  $\text{Kon}_{2-1}$  weist keine signifikanten Unterschiede im Gruppenvergleich auf ( $F(1, 59) = 1.71$ ,  $p = 0.196$ ). Zwischen den beiden Post-Testzeitpunkten steigert nur die Interventionsgruppe (KG-IG-Gruppe) ihre Leistung signifikant ( $t_{IG}(28) = 2.04$ ,  $p =$

0.051,  $d_{IG} = 0.40$ ), wohingegen die Änderung des konzeptionellen Verständnisses in der Kontrollgruppe (IG-KG-Gruppe) insignifikant bleibt ( $t_{KG}(31) = 0.57$ ,  $p = 0.572$ ). Die Hake-Indizes betragen für diesen Zeitraum  $g_{H,KG} = 0.07$  bzw.  $g_{H,IG} = 0.27$ . Über den gesamten Zeitraum ( $t_0$  bis  $t_2$ ) betrachtet, erzielen beide Untersuchungsgruppen ähnlich hohe normierte Lernzuwächse ( $g_{H,KG-IG} = 0.41$  bzw.  $g_{H,IG-KG} = 0.44$ ).

In den Streudiagrammen (Abb. 28) können die Veränderungen des Konzeptverständnisses zwischen je zwei Testzeitpunkten auf Individualebene nachvollzogen werden. Da der Test zur Erfassung des Konzeptverständnisses nur fünf Items umfasst, gruppieren sich die Datenpunkte rasterartig um die Gitterpunkte mit Abständen von jeweils 20%. Um ein Übereinanderliegen und damit eine Ununterscheidbarkeit der Datenpunkte zu verhindern, wurde (nur zu Darstellungszwecken) ein geringes Rauschen hinzugefügt. Insgesamt zeigt sich ein deutlicher Zusammenhang zwischen den Leistungen zu je zwei Zeitpunkten, vgl. Tab. 35 sowie Fit-Indizes ( $R^2$ ) in den Streudiagrammen. Dementsprechend erklären die Prä-Werte auch einen hohen Anteil der Variabilität der abhängigen Variable ( $\eta^2 = 0.40$  bzw.  $\eta^2 = 0.24$ , vgl. Tab. 37). Allerdings liegt in allen drei Streudiagrammen eine beachtliche Menge von Datenpunkten auf oder unterhalb der Winkelhalbierenden, was bedeutet, dass die Studierenden ihre Leistung nicht steigern können (bzw. sogar verschlechtern). Ferner gibt es einige Studenten, die einen enormen Sprung von niedriger oder mittlerer zu hoher (höchster) Fähigkeitsausprägung schaffen.

Im nächsten Abschnitt wird untersucht, bei welchen Items die Videoanalyse zu einem höheren Lernerfolg führte. Dabei beschränkt sich die Analyse auf die erste Interventionsphase.

### Item- und Distraktorenanalyse

Aus den deskriptiven Daten in Tabelle 38 geht hervor, dass die Interventionsgruppe vor allem bei den Items Kon3 und Kon4 einen hohen Lernzuwachs erzielte, während sich die Kontrollgruppe bei diesen Items kaum bzw. vergleichsweise weniger stark verbesserte.

**Tab. 38:** Deskriptiver Vergleich der Item-Schwierigkeiten (Konzepttest)

Item	KG			IG		
	$P(t_0)$	$P(t_1)$	$\Delta$	$P(t_0)$	$P(t_1)$	$\Delta$
Kon1	.51	.44	-.08	.42	.56	.15
Kon2	.66	.74	.09	.67	.77	.10
Kon3	.57	.69	.12	.46	.74	.29
Kon4	.47	.49	.01	.45	.64	.19
Kon5	.64	.77	.13	.74	.85	.12

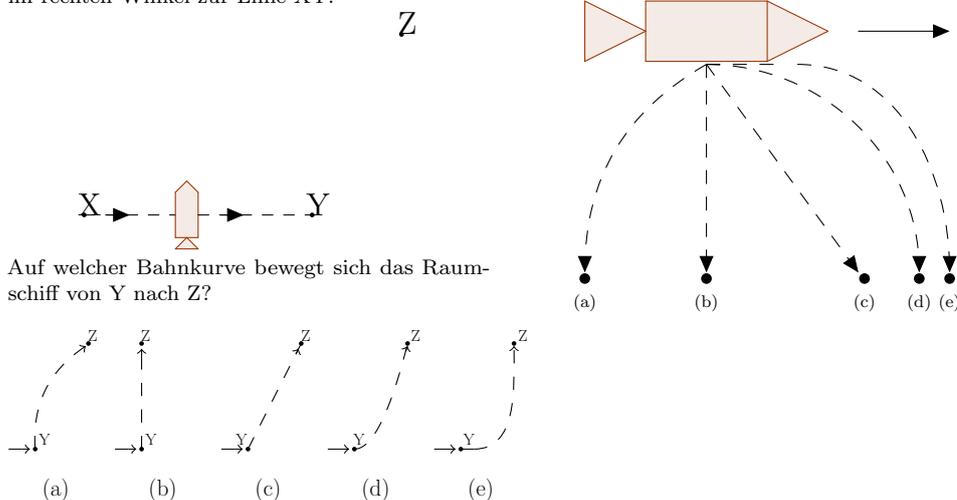
*Bemerkung:*  $\Delta$  = Differenzwerte.

Die beiden Items Kon3 und Kon4 sind im Anhang (S. 202) abgedruckt, aber für einen besseren Lesefluss auch in Abb. 29 dargestellt. Eine Distraktorenanalyse bei diesen beiden Items zeigt, dass die Impetus-Vorstellung beider Gruppen bei der Raumschiff-Aufgabe infolge der Instruktion abnimmt (Tab. 39). In der IG entscheiden sich zum Post-Zeitpunkt zudem viel weniger Studenten für die Alternative (c), in der das Vorhandensein einer konstanten Kraft fälschlicherweise nicht in einer beschleunigten Bewegung resultiert. Die beiden naiven Vorstellungen sind zum Prä-Zeitpunkt nur selten vertreten und verschwinden zum Post-Zeitpunkt, was vermutlich darauf zurückzuführen ist, dass die bildhafte Konfrontation mit den Bewegungsvorgängen bei vorhandenen Fehlvorstellungen einen kognitiven Konflikt auslöst.

Auch in Aufgabe Kon4 verliert die Impetus-Vorstellung in beiden Gruppen an Attraktivität. In der Kontrollgruppe findet eine Umverteilung der Antworthäufigkeiten zwischen falschen Distraktoren statt, wodurch kein Lernzuwachs zustande kommt. Im Gegensatz zur Aufgabe Kon3 reduziert sich die Häufigkeit der Alternative (c) in der Interventionsgruppe nicht. Der Lernzuwachs wird neben dem Überwinden der Impetus-Vorstellung vor allem durch das Ablegen naiver Vorstellungen erklärt.

**Item Kon3** Ein Raumschiff bewegt sich im Welt-  
raum ohne den Einfluss von äußeren Kräften von  
Position X nach Position Y. An Punkt Y wird das  
Raketentriebwerk gezündet und das Raumschiff  
bewegt sich auf einen Punkt Z zu. Das Raketen-  
triebwerk erzeugt eine konstante Beschleunigung  
im rechten Winkel zur Linie XY.

**Item Kon4** Eine Kugel fällt aus dem Frachtraum  
eines Flugzeuges während das Flugzeug mit kon-  
stanter Geschwindigkeit in horizontaler Richtung  
fliegt. Vernachlässigen Sie Luftreibungskräfte! Der  
Vorgang wird von der Erde aus beobachtet.  
Welche Bahnkurve beschreibt die Kugel?



**Abb. 29:** Item Kon3 und Kon4 aus dem Konzepttest

**Tab. 39:** Distraktorenanalyse ausgewählter Konzept-Items

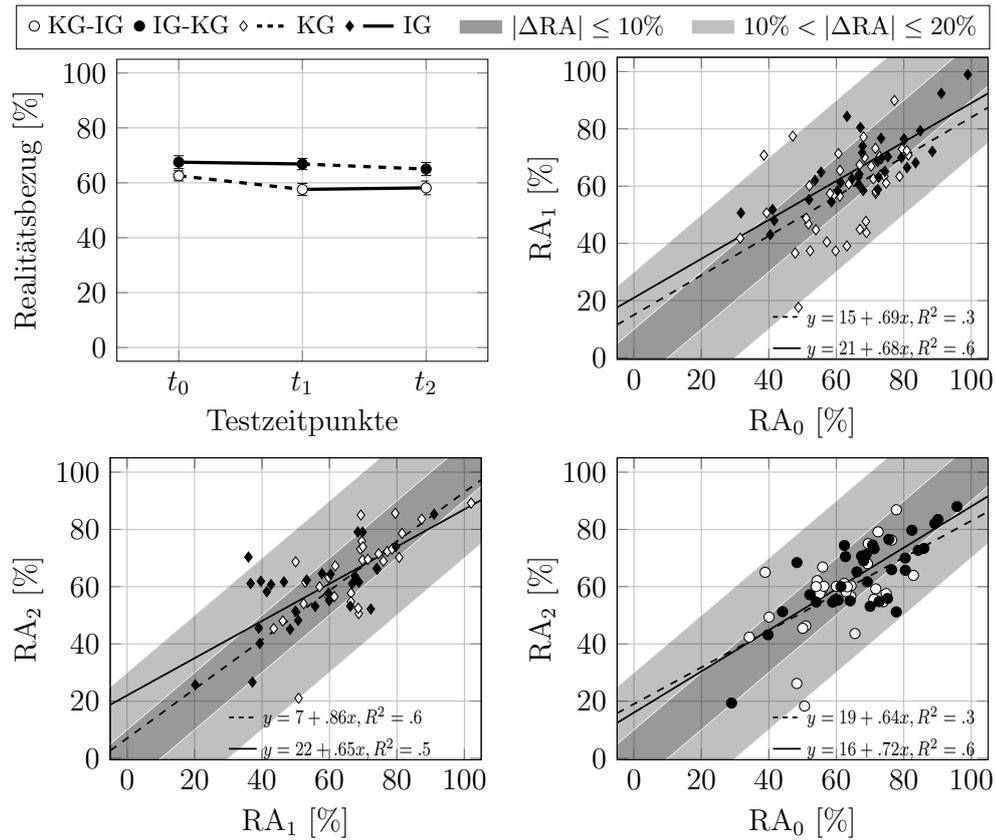
	KG			IG		
	$N(t_0)$	$N(t_1)$	$\Delta$	$N(t_0)$	$N(t_1)$	$\Delta$
<b>Raumschiff-Aufgabe Kon3</b>						
a) naiv	1	0	-1	2	0	-2
b) naiv	1	0	-1	0	0	0
c) Kraft $\rightarrow$ keine Beschl.	6	6	0	8	2	-6
d) korrekt	20	26	+6	15	26	+11
e) Impetus-Vorstellung	8	4	-4	9	6	-3
<b>Flugzeug-Aufgabe Kon4</b>						
a) naiv	2	0	-2	2	2	0
b) naiv	4	8	+4	5	2	-3
c) Kraft $\rightarrow$ keine Beschl.	10	9	-1	7	8	+1
d) korrekt	16	17	+1	15	20	+5
e) Impetus-Vorstellung	3	1	-2	3	0	-3

*Bemerkungen:*  $N$  = Häufigkeit der gewählten Alternativen,  $\Delta$  = Differenzwerte.

### 11.2.3 Motivation

Das Motivationskonstrukt unterteilt sich gemäß den theoretischen Vorarbeiten und den Testanalysen in sechs Subkomponenten, die im Folgenden nacheinander analysiert werden. Wegen der Doppelrolle des Experimentbezugs als abhängige Variable und als Variable des Manipulation Checks wird die Analyse unter Kap. 11.4.1 eingeordnet (methodische Fragestellungen).

Im Anschluss an die Kovarianzanalysen werden die Statistiken der Items, die nicht zur Skalenbildung genutzt wurden, auf Gruppenebene miteinander verglichen.



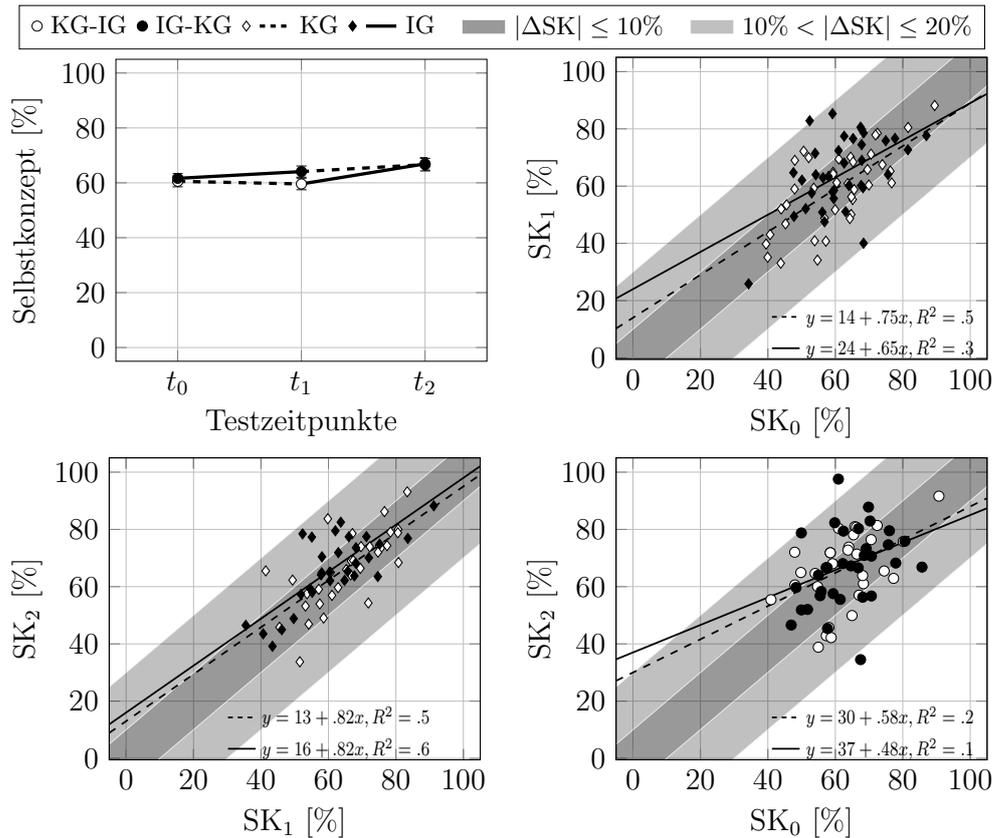
**Abb. 30:** Unadjustierte Gruppenmittelwerte und Standardfehler bzgl. des Realitätsbezugs der Aufgaben sowie Streudiagramme mit Regressionsgeraden innerhalb der Untergruppen.

**Realitätsbezug der Aufgaben**

Der Realitätsbezug der Aufgaben unterscheidet sich zum Post1-Zeitpunkt signifikant und mit kleiner Effektstärke zwischen den Gruppen ( $F(1, 72) = 6.34, p = 0.014, d = 0.37$ ; vgl. Tab. 37), wobei die Kovariate  $RA_0$  einen signifikanten Anteil der Varianz aufklärt ( $\eta^2 = 0.42$ ). Aus den deskriptiven Daten (siehe Abb. 30) geht hervor, dass die IG nach der ersten Interventionsphase den Aufgaben einen höheren Realitätsbezug zuschreibt als die KG. Die Gruppenmittelwerte sind zwar schon zum Prä-Zeitpunkt ( $RA_0$ ) leicht verschieden mit höherer Ausprägung bei der IG, doch diese Differenz wird durch die ANCOVA korrigiert ( $RA_1$  wird in der IG nach unten und in der KG nach oben adjustiert), sodass der Gruppenunterschied gerade nicht durch die anfängliche Differenz erklärt wird. Ohne diese Korrekturmaßnahme würde der Effekt größer ausfallen und (fälschlicherweise) überschätzt werden.

Zum zweiten Post-Zeitpunkt ist der Gruppenunterschied insignifikant ( $F(1, 60) = 0.67, p = 0.416$ ). Das Verschwinden des Gruppenunterschieds wird durch einen schwach signifikanten Interaktionseffekt ( $RA_{2-1}$ ) erklärt ( $F(1, 60) = 3.64, p = 0.061, d = 0.41$ ).

Das erste Streudiagramm in Abb. 30 (oben rechts) verdeutlicht, dass sich die AV bei den meisten Studierenden der IG während der ersten Semesterphase nicht stark verändert, während sie in der Kontrollgruppe vor allem bei den Studierenden abfällt, die den Realitätsbezug der Übungsaufgaben vor der Intervention gering bis mittelgroß einschätzten. Aus dem Post2-Post1-Streudiagramm geht hervor, dass ein Teil der Studierenden mit niedrigem Realitätsbezug zum Post1-Zeitpunkt (die nun zur IG gehören), in der zweiten Interventionsphase einen moderaten Anstieg des Realitätsbezugs wahrnehmen.



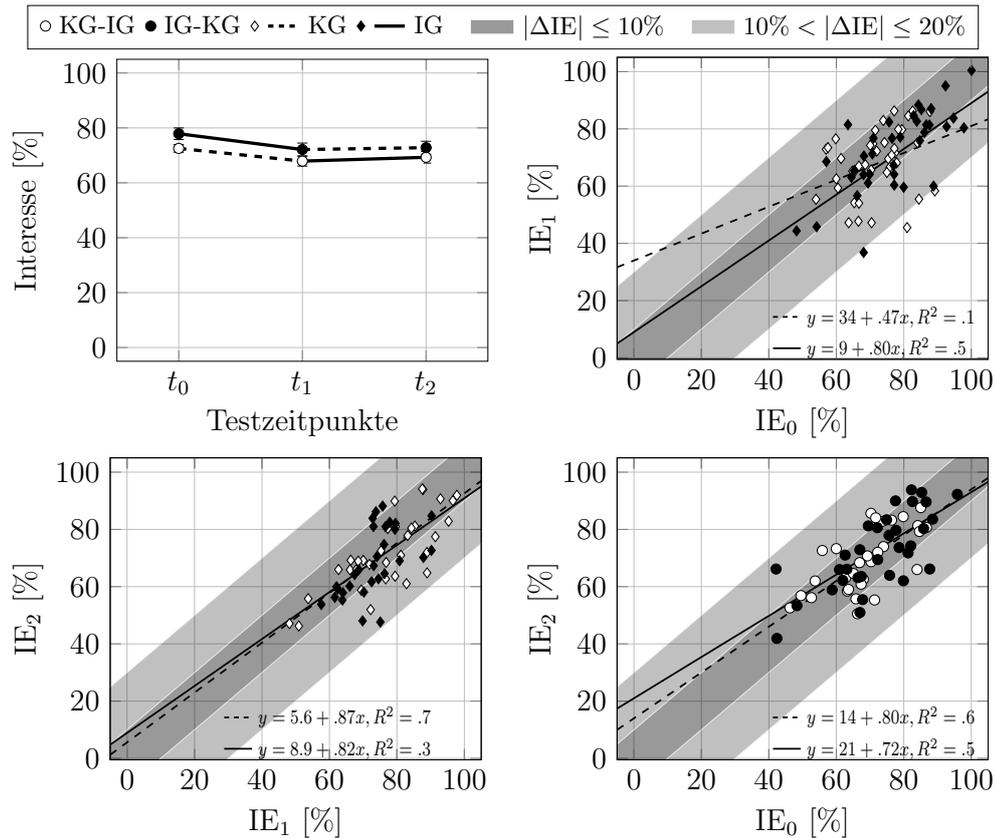
**Abb. 31:** Unadjustierte Gruppenmittelwerte und Standardfehler bzgl. des akademischen Selbstkonzepts sowie Streudiagramme mit Regressionsgeraden innerhalb der Untergruppen.

### Selbstkonzept

Es wurden keine signifikanten Unterschiede in der Ausprägung des Selbstkonzepts festgestellt, weder zum Post1-Zeitpunkt ( $F(1, 69) = 2.64, p = 0.109$ ) noch zum Post2-Zeitpunkt ( $F(1, 58) = 0.01, p = 0.931$ ), vgl. Tab. 37. Demzufolge unterscheidet sich auch der Zeitverlauf zwischen den beiden Post-Zeitpunkten nicht signifikant voneinander ( $F(1, 59) = 2.17, p = 0.146$ ). In den beiden Post-Analysen wurden die Prä-Werte als Kovariate berücksichtigt, die jeweils signifikant zur Varianzaufklärung beitragen. Die Berücksichtigung der Aufgabenschwierigkeit ( $CL_0$ ) im Post1-Vergleich trägt hingegen nicht zur Aufklärung der Fehlervarianz bei.

Die deskriptiven Daten lassen eine zeitliche Stabilität des Selbstkonzepts in der ersten Untersuchungshälfte erkennen (Abb. 31). Der geringe Gruppenunterschied zum ersten Post-Zeitpunkt ist wie oben berichtet insignifikant. Zwischen den beiden Post-Zeitpunkten ist auf Stichprobenebene ein signifikanter Anstieg des Selbstkonzepts mit kleiner Effektstärke zu verzeichnen ( $t(62) = 2.4, p = 0.019, d = 0.24$ ). Die Interventionsgruppe zeigt gemäß den deskriptiven Daten in beiden Phasen einen Trend zur Steigerung des Selbstkonzepts.

Die Streudiagramme in Abb. 31 geben keine Hinweise auf systematische Einflüsse der Intervention auf bestimmte Personengruppen.

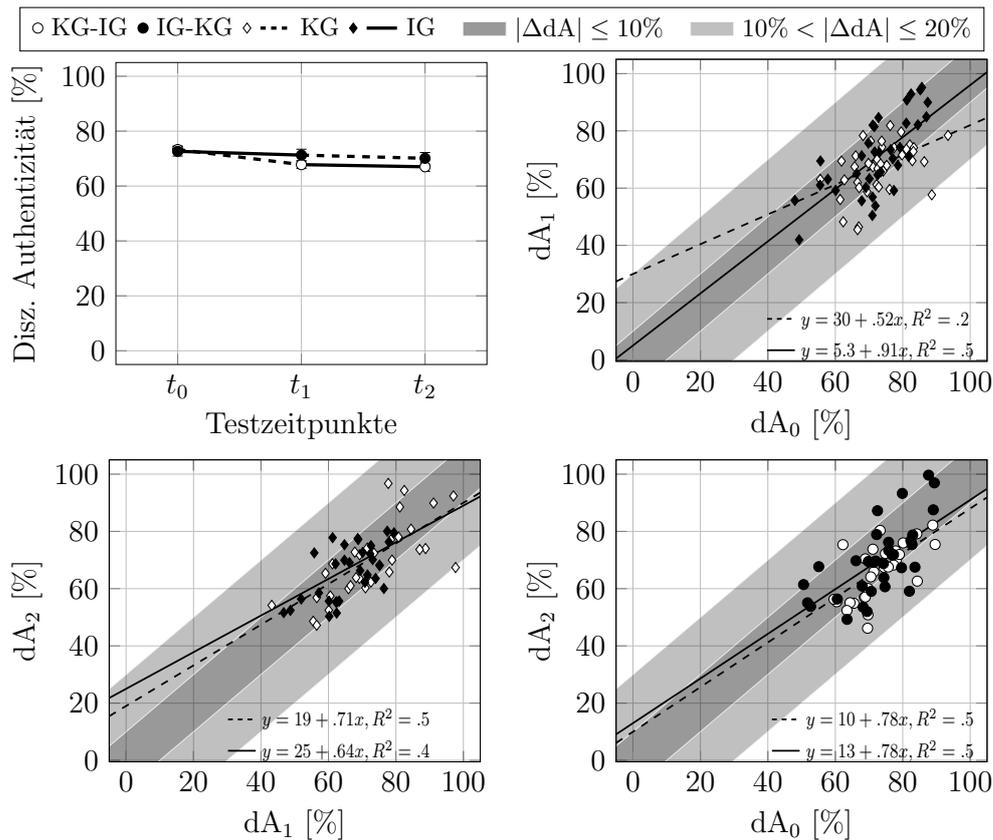


**Abb. 32:** Unadjustierte Gruppenmittelwerte und Standardfehler bzgl. des Interesses / Engagements sowie Streudiagramme mit Regressionsgeraden innerhalb der Untergruppen.

### Interesse / Engagement

Die Kovarianzanalysen mit Interesse / Engagement als abhängige Variable ergaben keine signifikanten Effekte unter der Experimentalbedingung „Gruppe“, weder zum ersten ( $F(1, 71) = 0.05, p = 0.826$ ), noch zum zweiten Post-Testzeitpunkt ( $F(1, 58) = 0.05, p = 0.828$ ), vgl. Tab 37. Da sich die Variable  $IE_0$  wegen Verletzung der Gruppenunabhängigkeit zum Prä-Zeitpunkt als Kovariate nur bedingt eignet (vgl. Tab. 36), wurde zusätzlich eine Varianzanalyse (ohne Kovariaten) durchgeführt. Diese ist auch sowohl für die Post1- als auch für die Post2-Variable insignifikant ( $F(1, 72) = 2.05, p = 0.156$  bzw.  $F(1, 59) = 1.22, p = 0.273$ ). Da keine Unterschiede zwischen den Gruppen zu den beiden Post-Zeitpunkten festgestellt wurden, bleibt auch das Ergebnis der Differenzwerte  $IE_{2-1}$  insignifikant ( $F(1, 59) = 0.00, p = 0.998$ ).

Der in Abb. 32 dargestellte Zeitverlauf zeigt zunächst einen Abfall des Interesses (gleichmaßen in beiden Untersuchungsgruppen), gefolgt von einer Stabilität der Interessensausprägung. Die Abnahme des Interesses ist auf Stichprobenebene statistisch signifikant ( $t(75) = 3.80, p < 0.001, d = 0.44$ ). Das zugehörige Streudiagramm (Abb. 32 oben rechts) zeigt, dass das Interesse vor allem bei Studierenden im unteren und mittleren Interessensbereich abnimmt, unabhängig ihrer Gruppenzugehörigkeit.

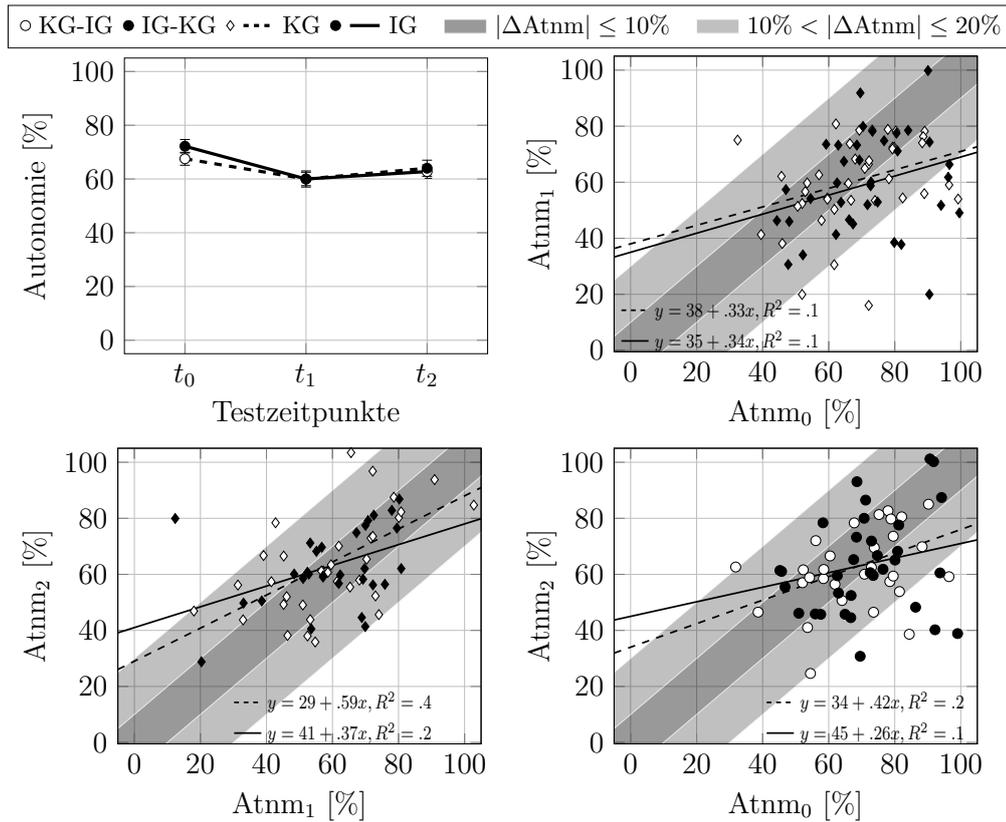


**Abb. 33:** Unadjustierte Gruppenmittelwerte und Standardfehler bzgl. der disziplinären Authentizität sowie Streudiagramme mit Regressionsgeraden innerhalb der Untergruppen.

### Disziplinäre Authentizität

Die wahrgenommene disziplinäre Authentizität des Instruktionsmaterials unterscheidet sich zum ersten Post-Zeitpunkt schwach signifikant zwischen den Untersuchungsgruppen ( $F(1, 71) = 3.60, p = 0.062$ ), wobei die Prä-Werte signifikant zur Varianzaufklärung beitragen ( $\eta^2 = 0.39$ ). Der Effekt ist klein ( $d = 0.25$ ) und nur von geringer praktischer Relevanz. Im zweiten Interventionszeitraum bleibt der wahrgenommene Professionsbezug der Aufgaben in beiden Gruppen weitgehend stabil. Obwohl die Zeitverläufe in diesem Zeitraum in zwischen den Gruppen nahezu parallel erscheinen (siehe Abb. 33), weist die Post2-Variable keinen sign. Gruppenunterschied auf ( $F(1, 56) = 3.03, p = 0.087$ ). Der Grund liegt hauptsächlich in der geringeren Stichprobengröße, die zu einer geringeren Anzahl von Fehlerfreiheitsgraden führt, und dem etwas kleineren  $F$ -Bruch. Die Interaktionsvariable  $dA_{2-1}$  unterscheidet sich nicht signifikant zwischen den Gruppen ( $F(1, 59) = 2.35, p = 0.131$ ).

Nach den deskriptiven Daten zeigt die Interventionsgruppe in beiden Phasen eine zeitliche Stabilität der wahrgenommenen disziplinären Authentizität, die auch statistisch durch  $t$ -Tests belegt werden kann ( $t_{1-0}(29) = 0.29, p = 0.777$  für die erste,  $t_{2-1}(35) = 1.15, p = 0.257$  für die zweite Interventionsphase). Die disziplinäre Authentizität fällt in der Kontrollgruppe in der ersten Phase hingegen tendenziell ab ( $t_{1-0}(37) = 3.96, p < 0.001$ ). Durch die Streudiagramme wird ersichtlich, dass die intraindividuellen Veränderungen zwischen zwei Testzeitpunkten eher gering ausfallen (vgl. Abb. 33). Große Änderungen ( $\Delta dA > 20\%$ ) treten nicht auf. Zwischen den ersten beiden Testzeitpunkten fällt auf, dass die videobasierten Aufgaben besonders positiv auf die Studenten wirkten, die schon vor der Intervention eine positive Erwartungshaltung an die disziplinäre Authentizität der Übungsaufgaben hatten.



**Abb. 34:** Unadjustierte Gruppenmittelwerte und Standardfehler bzgl. der Autonomie (Atnm) sowie Streudiagramme mit Regressionsgeraden innerhalb der Untergruppen.

**Experimentbezug**

Siehe Kap. 11.4.1.

**Autonomie**

Auch die Autonomie-Subskala zeigt keine signifikanten Gruppenunterschiede, weder zum Post1-Zeitpunkt ( $F(1, 73) = 0.18, p = 0.671$ ), noch zum Post2-Zeitpunkt ( $F(1, 58) = 0.02, p = 0.894$ ). In beiden Fällen trägt die Prä-Variable  $\text{Atnm}_0$  signifikant zur Varianzaufklärung bei ( $\eta^2 = 0.07$  bzw.  $\eta^2 = 0.10$ ) — allerdings fällt die Reduktion der Fehlervarianz durch die modellspezifische Prä-Variable viel geringer aus als bei den anderen Motivationskomponenten.

Die deskriptiven Daten in Abb. 34 zeigen zunächst einen Abfall der empfundenen Autonomie, der sich in beiden Gruppen als statistisch signifikant erweist ( $t_{KG}(39) = 2.67, p = 0.011$ ;  $t_{IG}(35) = 3.58, p = 0.001$ ). Anschließend stabilisieren sich die Werte in beiden Gruppen. Die Streudiagramme zeigen eine große Streuung innerhalb beider Gruppen. Zwischen dem Prä- und Post1-Test fällt die empfundene Autonomie während der Aufgabenbearbeitung insbesondere bei den Studierenden ab, die mit einer hohen Erwartungshaltung an die Autonomie der Aufgaben in das Studium starteten. Nur wenige Studierende zeigen einen Anstieg empfundener Autonomie oberhalb des 10%-Bereichs.

### Deskriptiver Vergleich ausgeschlossener Items

Auf Grundlage der klassischen Testtheorie wurden 13 Items aus dem ursprünglichen Fragebogen zur Erfassung der Motivation ausgeschlossen (über alle Subskalen), die folglich nicht in den Skalen berücksichtigt wurden, vgl. Kap. 6.4.3. Um die statistische Analyse der einzelnen Motivationskomponenten zu ergänzen, stellt Tab. A-12 (im Anhang, S. 218) die deskriptiven Daten dieser Items dar, wobei sich auf den Prä- und Post1-Zeitpunkt eingeschränkt wird.

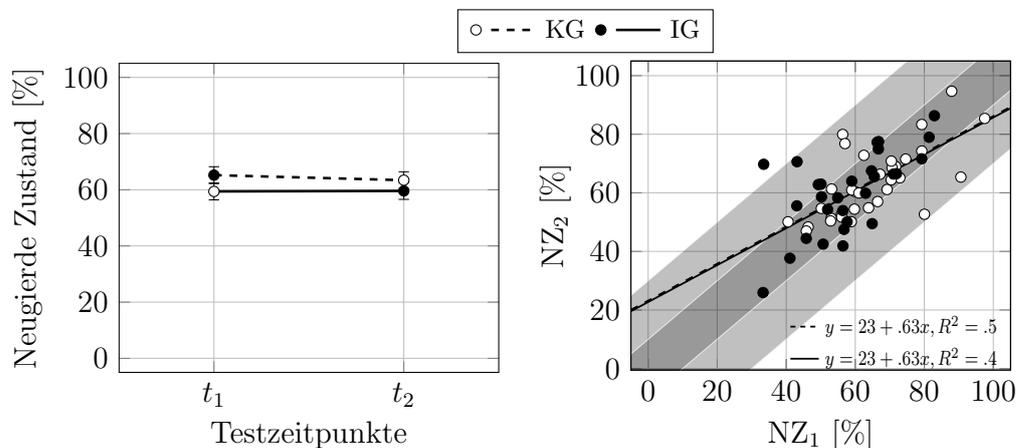
Die Kontrollgruppe verzeichnet bei den beiden Items M02 („Ich schaue zu Hause in Büchern, im Internet oder ähnlichem nach, um mehr zu den Übungsaufgaben zu erfahren“) und M16 („Der Lösungsweg einer Aufgabe ist fest vorgeschrieben“) einen Motivationsanstieg von 3% bzw. 1%. Ansonsten bleiben die Ausprägungen zwischen den Testzeitpunkten stabil oder fallen ab. Der größte Motivationsabfall (14%) zeigt Item M37 („Unabhängig von Prüfungsanforderungen beschäftige ich mich intensiv mit interessanten Aufgaben“).

Im Unterschied zur KG nimmt die Zustimmung bzgl. Item M16 bei der IG um 8% ab. Dahingegen weisen die Items M19 („Ich beschäftige mich auch über Übungsaufgaben hinaus mit Themen, die mit den darin behandelten Inhalten zu tun haben“), M23 („Die Aufgaben geben mir die Möglichkeit, neue Themengebiete selbstständig zu erschließen“) und M45 („Den Dingen, die wir in den Physik-Übungen lernen, messe ich persönlich eine große Bedeutung zu“) stabilere Ausprägungen auf.

Auffällig ist, dass das Item M27 („Ich strenge mich an, alle gestellten Übungsaufgaben zu bearbeiten“) zum Prä-Zeitpunkt von beiden Gruppen fast volle Zustimmung erfährt ( $M_{KG} = 0.93$ ,  $M_{IG} = 0.96$ ), die zum Post-Zeitpunkt jeweils um etwa 10% abfällt.

### 11.2.4 Neugierde

In diesem Abschnitt werden die varianzanalytischen Ergebnisse der NZ-Skalen berichtet (epistemische Neugierde bzgl. des Instruktionsmaterials — aktueller Zustand). Aufgrund der engen konzeptionellen Verwandtschaft schließt sich eine Analyse der NP-Skalenwerte (Neugierde als Personeneigenschaft) an den Ergebnisbericht der NZ-Skalen an.



**Abb. 35:** Unadjustierte Gruppenmittelwerte und Standardfehler der Neugierde (als aktueller Zustand, NZ) nach der ersten bzw. zweiten Interventionsphase (links) sowie Streudiagramm (rechts). Während die Symbolkodierung der Messpunkte in dem linken Diagramm die Gruppenzugehörigkeit unmittelbar vor dem jeweiligen Messzeitpunkt angibt, steht sie in dem rechten Diagramm für die Zugehörigkeit während der zweiten Interventionsphase.

Die Varianzanalyse ergab einen signifikanten Unterschied bezüglich der epistemischen Neugierde zum Post1-Zeitpunkt mit kleiner bis mittelgroßer Effektstärke ( $F(1, 73) = 3.82$ ,  $p = 0.054$ ,  $d = 0.39$ ). Da NZ nicht zum Prä-Zeitpunkt erhoben wurde, kann keine Korrektur auf

eventuell vorhandene Unterschiede vor der Intervention erfolgen. Das zur Analyse zugeordnete Diagramm ist in Abb. 35 (links) dargestellt und verdeutlicht, dass die Interventionsgruppe eine höhere epistemische Neugierde aufweist als die Kontrollgruppe.

Zum zweiten Post-Zeitpunkt ergeben sich keine signifikanten Unterschiede bzgl. der abhängigen Variable zwischen den Gruppen ( $F(1, 61) = 1.42, p = 0.240$ ). Auch der Verlauf zwischen den beiden Post-Zeitpunkten bleibt insignifikant ( $F(1, 60) = 1.42, p = 0.238$ ). Das zugehörige Streudiagramm  $NZ_2$  vs.  $NZ_1$  (Abb. 35, rechts) gibt keine Hinweise darauf, dass die Veränderung der Neugierde vom Anfangszustand abhängt.

Neben dem Einfluss des Instruktionsmaterials auf die situative Neugierde wurde getestet, ob sich die Untersuchungsgruppen hinsichtlich der Neugierde als Personeneigenschaft (NP) voneinander unterscheiden. Es konnten zu keinem Zeitpunkt signifikante Unterschiede zwischen den Gruppen festgestellt werden. Es zeigt sich auf Stichprobenebene ein signifikanter Abfall der Skalenmittelwerte über die Zeit ( $F(2, 60) = 12.29, p < 0.001$ ).

### 11.3 Ergebnisse zur Fragestellung 2: Aufgabeninduzierte Belastung

In Anlehnung an den Ergebnisbericht zur Hauptfragestellung 1 stellt Tabelle 40 die Ergebnisse der Kovarianzanalysen zu den abhängigen Variablen „Aufgabenschwierigkeit“ (abgekürzt durch CL wegen dem Schwerpunkt auf intrinsic Cognitive Load) und „Beanspruchung“ (abgekürzt durch PSQ in Anlehnung an die Quelle der Skala — Perceived Stress Questionnaire) dar. Die Analysen bzgl. der Beanspruchung wurden ohne Kovariate durchgeführt, da  $PSQ_0$  gegen das harte Voraussetzungskriterium homogener Regressionssteigungen verstößt (vgl. Kap. 11.1.5).

Im Folgenden werden die Ergebnisse ausformuliert und mit deskriptiven Daten ergänzt (Kap. 11.3.1 und 11.3.2). Anschließend folgt die Auswertung der benötigten Zeit zur Bearbeitung des Instruktionsmaterials (Kap. 11.3.3). Die zugehörigen Streudiagramme sind – bis auf eine diskussionswürdige Ausnahme — im Anhang dargestellt (Abb. A-4 auf S. 218).

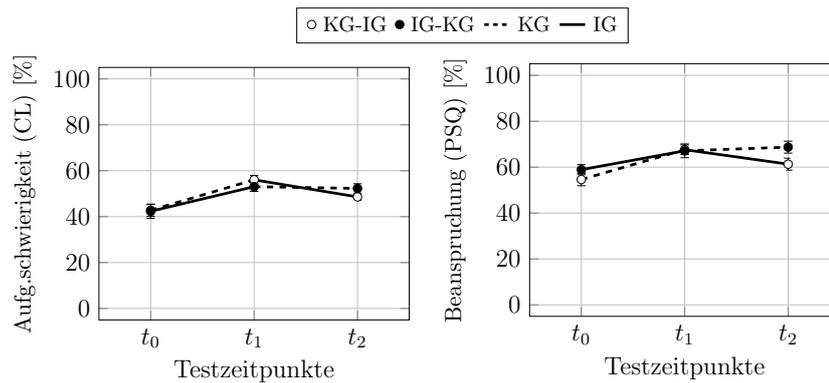
**Tab. 40:** Ergebnisse der Kovarianzanalysen zur Belastung

Hyp	AV	$df_F$	KV / Faktor	$F$ -Wert	$p$ -Wert	$\eta_p^2$	$\eta^2$	$\omega^2$	$d$
2.1	CL <sub>1</sub>	73	Gruppe	1.62	.208	.02	.02	-	-
			CL <sub>0</sub>	41.58	< .001	.36	.36		
	CL <sub>2</sub>	60	Gruppe	2.51	.118	.04	.03	-	-
			CL <sub>0</sub>	21.31	< .001	.26	.25		
	CL <sub>2-1</sub>	61	Gruppe	5.55	.022	.08	.08	.07	.54
2.2	PSQ <sub>1</sub>	73	Gruppe	.02	.902	.00	.00	-	-
	PSQ <sub>2</sub>	62	Gruppe	3.95	.051	.06	.06	.04	.43
	PSQ <sub>2-1</sub>	61	Gruppe	4.18	.045	.06	.06	.05	.44

*Bemerkungen:*  $df_F$  = Fehlerfreiheitsgrade. Der Zählerfreiheitsgrad beträgt bei jeder Analyse 1. Die Effektstärkemaße  $\omega^2$  und  $d$  werden nur für Haupteffekte (Faktor Gruppe) berichtet. Auf die Angabe von Effektstärken wird verzichtet (-), wenn  $p \gg 0.05$ .

#### 11.3.1 Aufgabenschwierigkeit (Cognitive Load)

Die Aufgabenschwierigkeit unterscheidet sich zum Post1-Zeitpunkt nicht signifikant zwischen den Gruppen ( $F(1, 73) = 1.62, p = 0.208$ ). Die Einschätzung der Aufgabenschwierigkeit zum Prä-Zeitpunkt CL<sub>0</sub> (aufgrund der ersten beiden Übungsblätter vor der Intervention)



**Abb. 36:** Unadjustierte Gruppenmittelwerte bzgl. der empfundenen Aufgabenschwierigkeit (links) und der wahrgenommenen Beanspruchung (rechts). Die Streudiagramme befinden sich im Anhang (Abb. A-4).

trägt einen signifikanten Anteil zur Varianzaufklärung bei ( $\eta^2 = 0.36$ ). Auch zum zweiten Post-Zeitpunkt werden keine signifikanten Gruppenunterschiede in der Aufgabenschwierigkeit festgestellt, wenn auf die Prä-Werte  $CL_0$  korrigiert wird ( $F(1, 60) = 2.51$ ,  $p = 0.118$ ). Die Entwicklung der wahrgenommenen Aufgabenschwierigkeit zwischen den beiden Post-Zeitpunkten unterscheidet sich marginal signifikant zwischen den beiden Untersuchungsgruppen ( $F(1, 61) = 5.55$ ,  $p = 0.022$ ). Die Effektstärke liegt im mittleren Bereich ( $d = 0.54$ ).

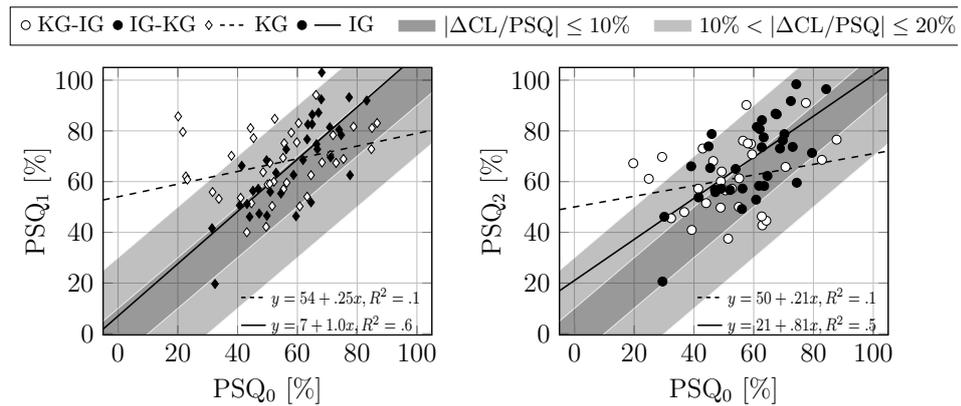
Aus der Darstellung des Zeitverlaufs in Abb. 36 (links) geht hervor, dass die Aufgabenschwierigkeit zunächst in beiden Untersuchungsgruppen stark ansteigt ( $t_{KG}(39) = 6.85$ ,  $p < 0.001$ ,  $d = 0.81$ ;  $t_{IG}(35) = 4.09$ ,  $p < 0.001$ ,  $d = 0.58$ ) und anschließend in der Kontrollgruppe stabil bleibt. Die Gruppe, die von traditionellen Aufgaben zu Videoanalyse-Aufgaben wechselte (KG-IG-Gruppe), verzeichnet einen signifikanten Abfall der Aufgabenschwierigkeit in der zweiten Interventionsphase ( $t_{IG}(29) = 3.62$ ,  $p = 0.001$ ,  $d = 0.48$ ).

### 11.3.2 Beanspruchung

Der Zeitverlauf der wahrgenommenen psychischen Beanspruchung des Physik-Studiums ist ebenfalls in Abb. 36 (rechts) dargestellt. Es wurden keine sign. Gruppenunterschiede zum ersten Post-Zeitpunkt festgestellt ( $F(1, 73) = 0.02$ ,  $p = 0.902$ ). Die Beanspruchung zum Prä-Zeitpunkt  $PSQ_0$  wurde infolge der unterschiedlichen Regressionssteigungen nicht als Kovariate verwendet. In den beiden Streudiagrammen (Abb. 37) wird die Verletzung gegen diese notwendige Bedingung der ANCOVA deutlich: Die Steigungen der Ausgleichsgeraden unterscheiden sich zwischen den Untersuchungsgruppen beachtlich.

In der Kontrollgruppe stieg die empfundene Beanspruchung im ersten Interventionszeitraum an ( $t_{KG}(39) = 4.63$ ,  $p < 0.001$ ,  $d = 0.77$ ). Dabei ist der Anstieg insbesondere bei den Studierenden sehr hoch (teilweise über 30%), die im Prä-Test eine sehr niedrige Beanspruchung angaben (vgl. Abb. 37 links). Auch in der Interventionsgruppe ist ein Anstieg der Beanspruchung im ersten Interventionszeitraum zu verzeichnen ( $t_{IG}(34) = 4.63$ ,  $p < 0.001$ ,  $d = 0.63$ ), jedoch lässt sich keine systematische Wirkung der Intervention in Abhängigkeit der Prä-Disposition erkennen (die Ausgleichsgerade ist winkelhalbierend). Diese gruppenspezifischen Zusammenhänge haben über den gesamten Interventionszeitraum bestand, wie das rechte Streudiagramm in Abb. 37 zeigt.

Es besteht ein signifikanter Gruppenunterschied bezüglich der Post2-Variable ( $F(1, 62) = 3.95$ ,  $p = 0.051$ ,  $d = 0.43$ ) und bezüglich der Differenzvariable  $PSQ_{2-1}$  ( $F(1, 61) = 4.18$ ,  $p = 0.045$ ,  $d = 0.44$ ) mit jeweils mittlerer Effektstärke. Die IG-KG-Gruppe empfindet einen leichten Anstieg der Beanspruchung, der jedoch nicht signifikant ist. Die KG-IG-Gruppe dagegen beurteilt die psychische Beanspruchung durch das Studium am Ende der zweiten



**Abb. 37:** Streudiagramme der Beanspruchung. Die Verletzung gegen die Homogenität der Regressionssteigungen wird durch die stark unterschiedlichen Steigungsparameter der Ausgleichsgeraden innerhalb der Untersuchungsgruppen deutlich.

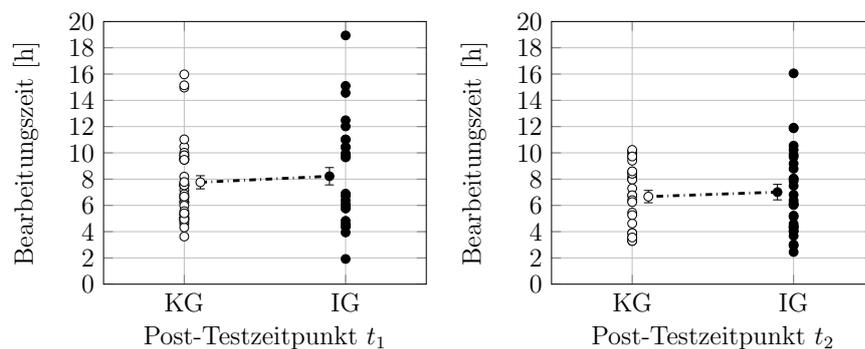
Interventionsphase signifikant geringer im Vergleich zum ersten Post-Zeitpunkt ( $t_{IG}(29) = 2.18, p = 0.037, d = 0.35$ ).

### 11.3.3 Bearbeitungszeit

Die wöchentliche Bearbeitungszeit der Aufgaben wurde zu den Zeitpunkten  $t_1$  und  $t_2$  erhoben. Abbildung 38 stellt die mittlere Bearbeitungszeit der Gruppen zu den beiden Testzeitpunkten dar. Die Datenpunkte vermitteln einen Eindruck über die Verteilung der Bearbeitungszeit innerhalb der Gruppen sowie über extremale Angaben einzelner Studenten. Die Auswertung der Daten erfolgte ohne Korrekturen auf Kontrollvariablen.

Die wöchentliche Bearbeitungszeit der Aufgaben (*Time-on-Task*) beträgt in der Kontrollgruppe nach der ersten Interventionsphase durchschnittlich 7.8 h ( $SD = 3.1$  h), in der Interventionsgruppe 8.2 h ( $SD = 3.7$  h). Die beiden Werte unterscheiden sich nicht signifikant voneinander ( $t(67) = 0.56, p = 0.580$ ).

Zum zweiten Post-Zeitpunkt beträgt die Bearbeitungszeit der Aufgaben 6.7 h ( $SD = 3.2$  h) in der KG bzw. 7.01 h ( $SD = 2.5$  h) in der IG. Auch dieser Unterschied ist insignifikant ( $t(55) = 0.45, p = 0.660$ ). Die Bearbeitungszeit nimmt insgesamt zwischen den beiden Testzeitpunkten signifikant mit kleiner Effektstärke ab ( $t(52) = 2.42, p = 0.019, d = 0.24$ ).



**Abb. 38:** Mittlere Zeitangabe zur Bearbeitung eines Übungszettels mit Standardfehlern nach den beiden Instruktionsphasen. Neben den Mittelwerten ist die Verteilung der Bearbeitungszeit dargestellt, wobei ein Datenpunkt einem Studierenden entspricht.

## 11.4 Ergebnisse zu den methodischen Fragestellungen

Im nachfolgenden Abschnitt werden die Ergebnisse der durchgeführten Analysen bezüglich den methodischen Fragestellungen vorgestellt: Nachdem die Resultate bzgl. des Experimentbezugs (Manipulation Check) beschrieben werden (Kap. 11.4.1), folgt eine detaillierte Untersuchung der in Kap. 11.2.1 diskutierten instruktionsbedingten Steigerung der Repräsentationskompetenz. In der vorangegangenen Analyse wurde die optimierte Kern-Itemgruppe betrachtet, innerhalb der alle Items mit dem gleichen Gewicht zu einer Gesamttestsumme beitragen. Diese Analyse wird in Kap. 11.4.2 um folgende Aspekte erweitert: Erstens wird die Gleichgewichtung der Testsummen aufgehoben, d.h. es wird untersucht, welchen Einfluss eine relative Gewichtung beider Testhälften auf die Effektstärke der Wirksamkeitsuntersuchung nimmt (Testmischungsanalyse). In diesem Zusammenhang ist es von praktischem Interesse, den Einfluss dieser Gewichtung auf die Reliabilität des Instruments ebenfalls zu analysieren. Zweitens werden die Effektstärken für die in Kap. 7 isolierte Rasch-Itemgruppe sowie für die Kern-Itemgruppe (ohne psychometrische Optimierung) berechnet. Drittens wird der Einfluss der Ratekorrektur auf das Ergebnis der Vergleichsanalyse untersucht (Kap. 11.4.3).

### 11.4.1 Manipulation Check

Wie in Kap. 11.1.5 festgestellt wurde, unterscheidet sich der wahrgenommene Experimentbezug der Aufgaben bereits zum Prä-Zeitpunkt signifikant zwischen den Gruppen. Um den Unterschied zu den Post-Zeitpunkten zu analysieren, wurde deshalb zunächst auf „einfache“ Varianzanalysen zurückgegriffen. Anschließend wurde  $\text{Exp}_0$  in das Modell eingeschlossen, um die Auswirkung auf das varianzanalytische Ergebnis zu untersuchen und die Diskussion zu ergänzen. Beide Analyseergebnisse sind in Tab. 41 zusammengefasst.

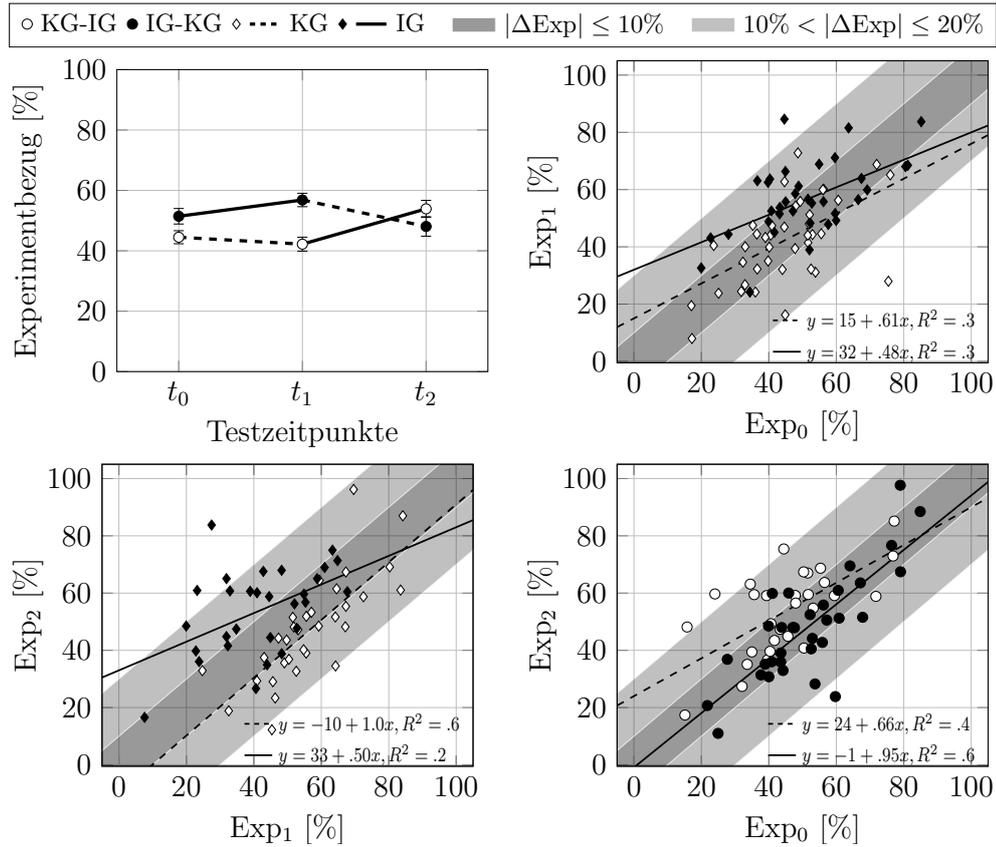
Im unkorrigierten Modell zeigen sich signifikante Gruppenunterschiede zum Post-Zeitpunkt ( $F(1, 73) = 20.57, p < 0.001$ ) und im Zeitverlauf  $\text{Exp}_{2-1}$  ( $F(1, 59) = 33.81, p < 0.001$ ). In beiden Fällen liegen große Effektstärken vor. Der in Abb. 39 (links) dargestellte Zeitverlauf verdeutlicht, dass die Interventionsgruppe zum Post1-Zeitpunkt einen höheren Experimentbezug der Aufgaben wahrnimmt als die Kontrollgruppe. Weiterhin fällt in der zweiten Instruktionsphase der Experimentbezug innerhalb der KG ab, während er in der IG ansteigt.

Im korrigierten Modell zeigen sich bezüglich den Variablen  $\text{Exp}_1$  und  $\text{Exp}_{2-1}$  ebenfalls signifikante Gruppenunterschiede. Die Effektstärke ( $d = 0.75$ ) ist zum ersten Post-Zeitpunkt substantiell kleiner als im unkorrigierten Modell ( $d = 1.02$ ), da die Gruppenunterschiede durch

**Tab. 41:** Ergebnisse der Kovarianzanalysen zum Manipulation Check

AV	$df_F$	KV / Faktor	$F$ -Wert	$p$ -Wert	$\eta_p^2$	$\eta^2$	$\omega^2$	$d$
$\text{Exp}_1$	73	Gruppe	20.57	< .001	.22	.22	.21	1.02
$\text{Exp}_2$	60	Gruppe	1.84	.180	.03	.03	-	-
$\text{Exp}_{2-1}$	59	Gruppe	33.81	< .001	.40	.40	.38	1.57
Unter Einschluss von $\text{Exp}_0$ :								
$\text{Exp}_1$	72	Gruppe	16.62	< .001	.19	.14	.12	.75
		$\text{Exp}_0$	33.24	< .001	.32	.27		
$\text{Exp}_2$	59	Gruppe	13.36	< .001	.19	.09	.08	.58
		$\text{Exp}_0$	70.07	< .001	.54	.49		
$\text{Exp}_{2-1}$	58	Gruppe	43.67	< .001	.43	.41	.40	1.63

*Bemerkungen:* AV = abhängige Variable,  $df_F$  = Fehlerfreiheitsgrade, KV = Kovariate. Der Zählerfreiheitsgrad beträgt bei jeder Analyse 1. Die Effektstärkemaße  $\omega^2$  und  $d$  werden nur für Haupteffekte (Faktor Gruppe) berichtet. Auf die Angabe von Effektstärken wird verzichtet (-), wenn  $p \gg 0.05$ .



**Abb. 39:** Unadjustierte Gruppenmittelwerte bzgl. des Experimentbezugs sowie Streudiagramme mit Regressionsgeraden innerhalb der Untergruppen.

die Korrektur auf die Prä-Werte verringert wird. Die Skalenmittelwerte der IG-Studenten werden im korrigierten Modell nach unten, die der KG-Studenten nach oben adjustiert. Aus dem gleichen Grund wird in diesem Modell auch der Faktor „Gruppe“ bzgl. der abhängigen Variable  $Exp_2$  signifikant ( $F(1, 59) = 13.36, p < 0.001, d = 0.58$ ). Die einbezogene Kovariante trägt bei den Post-Variablen jeweils signifikant zur Varianzaufklärung bei ( $\eta^2 = 0.27$  bzw.  $\eta^2 = 0.49$ ).

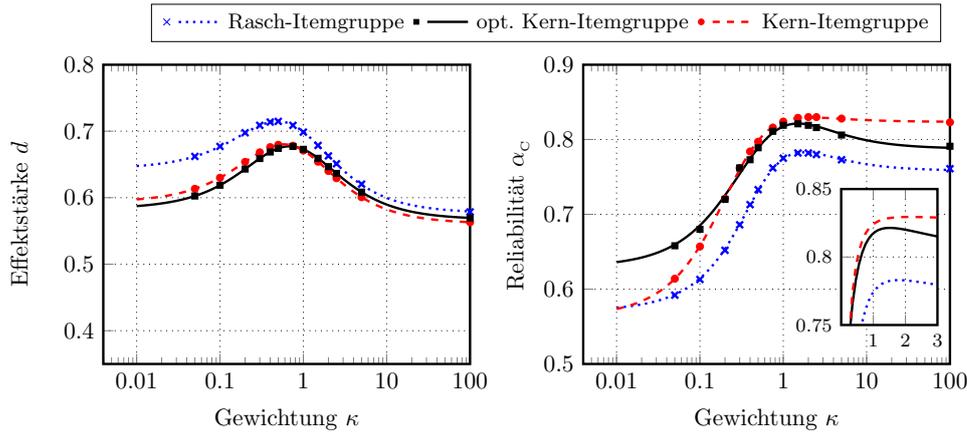
### 11.4.2 Effektstärken- und Reliabilitätsanalyse von KiRC mit variabler Testsummengewichtung

Gemäß den Ausführungen im Methodenteil (Kap. 10.5.5) wurden neue KiRC-Testsummen berechnet, indem die Annahme einer Gleichgewichtung beider Testhälften verworfen wurde. Stattdessen ergibt sich die neue Testsumme  $T$  gemäß Gl. (39) aus einer relativen Gewichtung beider Testsummen:

$$T = T_1 + \kappa T_2.$$

Die daraus resultierenden Effektstärken sind in Abb. 40 für die optimierte Kern-Itemgruppe (durchgezogene Linie) in Abhängigkeit des Gewichtungsparmeter  $\kappa$  dargestellt. Der Effektstärkenberechnung liegt Gl. (24) zugrunde, d.h. es wurden die mittleren Testsummen beider Gruppen zum Post-Zeitpunkt verglichen und zur Standardabweichung der Kontrollgruppe (zum Post-Zeitpunkt) ins Verhältnis gesetzt:

$$d = \frac{T(\kappa)_{IG(post)} - T(\kappa)_{KG(post)}}{SD(\kappa)_{KG(post)}}.$$



**Abb. 40:** Testmischungsanalyse: Effektstärke  $d$  der Vergleichsuntersuchung (links) und Reliabilität  $\alpha_c$  des KiRC-Instruments (rechts) in Abhängigkeit der Gewichtung  $\kappa$  der zweiten KiRC-Testhälfte für verschiedene Itemgruppen. Die Datenpunkte wurden für diskrete Werte von  $\kappa$  unmittelbar aus den empirischen Daten nach einer Transformation der Testsummen gemäß  $T = T_1 + \kappa T_2$  (mit  $T_{1,2} =$  Testsumme der MC3- / MTF-Items) ermittelt. Die Fitkurven ergeben sich durch Parameteranpassung der Funktionen aus Gl. (46) bzw. (60).

Die analytischen  $d(\kappa)$ -Funktionen resultieren aus einer Kurvenanpassung von Gleichung (46) an die händisch berechneten Effektstärken, wobei die Varianzanteile  $\phi$ ,  $\theta_1$  und  $\theta_2$  als freie Parameter behandelt wurden. In der Abbildung wurde eine logarithmische Abszisse gewählt, um bei Darstellung des rechten Grenzwerts  $\kappa \rightarrow \infty$  eine ausreichende Auflösung des Bereichs  $0 \leq \kappa \leq 1$  zu erreichen. Neben der optimierten Kern-Itemgruppe enthält das Diagramm die Effektstärken, die auf Grundlage anderer Itemgruppen berechnet wurden. Die (nicht-optimierte) Kern-Itemgruppe beinhaltet alle KiRC-Items, die in beiden Studien eingesetzt wurden, ungeachtet ihrer psychometrischen Güte. Die Rasch-Itemgruppe ist eine rasch-skalierte Teilmenge der optimierten Kern-Itemgruppe (vgl. Kap. 7).

Wie in Abb. 40 abgelesen werden kann, variiert die Effektstärke zwischen dem mittleren und oberen mittleren Bereich. Die Kurvenverläufe der Kern- und optimierten Kern-Itemgruppe unterscheiden sich kaum voneinander. Der Grund liegt darin, dass die optimierte Kern-Itemgruppe aus der Kern-Itemgruppe hauptsächlich durch den Ausschluss von zu leichten Items (Deckeneffekte) hervorging. Solche varianzarmen Items verhalten sich ähnlich zu additiven Konstanten und haben keinen Einfluss auf die Effektstärke. Weiterhin wird deutlich, dass die MC3-Itemgruppe ( $\kappa = 0$ ) zu höheren Effektstärken führt als die MTF-Itemgruppe ( $\kappa \rightarrow \infty$ ) und dass sich ein klares Maximum bei  $\kappa \approx 0.7$  ausbildet.

Die Rasch-Itemgruppe erzielt für jede Gewichtung  $\kappa$  etwas größere Effektstärken als die anderen beiden Itemgruppen, wobei der Unterschied vor allem auf die MC3-Items zurückzuführen ist. Rasch- und optimierte Kern-Itemgruppe unterscheiden sich hinsichtlich der MC3-Items K9 und K10, die gegen die lokale stochastische Unabhängigkeit verstoßen. Obwohl die IG bei Item K10 den gleichen und bei Item K9 einen deutlich höheren Lernzuwachs als die KG aufweist (vgl. Tab. A-11, Spalte  $\Delta_{IG-KG}$ ) erzeugt die Rasch-Itemgruppe nach Ausschluss dieser Items größere Effektstärken. Die geringere Effektvarianz der Rasch-Itemgruppe wird durch eine kleinere Fehlervarianz überkompensiert. Das Maximum der Rasch-Itemgruppe ist etwas nach links verschoben ( $\kappa \approx 0.5$ ), da der Unterschied zwischen MC3- und MTF-Itemgruppe größer ist als bei den anderen Itemgruppen.

Im rechten Teil der Abbildung ist die Reliabilität  $\alpha_c$  (nach Cronbach) gegen  $\kappa$  für die drei Itemgruppen aufgetragen. Zur Darstellung der kontinuierlichen Kurven wurde Gl. (60) mit den freien Parametern  $c_i$  ( $i = 1, \dots, 5$ ) an die Daten angepasst. Das Diagramm verdeutlicht, wie sich das Zusammenspiel zwischen psychometrischer Qualität der Items innerhalb einer

Skala und Skalenlänge auf die Reliabilität auswirkt:

- Die MC3-Itemgruppe erreicht stets geringere Reliabilitäten als die MTF-Itemgruppe, was auf die geringere Itemanzahl zurückgeführt werden kann, vgl. Tab. 42.
- Die itemschwächste Rasch-Itemgruppe besitzt die niedrigste Reliabilität.
- Trotz einer geringeren Itemanzahl besitzt der MC3-Stamm der optimierten Kern-Itemgruppe eine höhere Reliabilität als der MC3-Stamm der Kern-Itemgruppe, da das ausgeschlossene MC3-Item psychometrisch unpassend ist.
- Bei der MTF-Itemgruppe wurden neun psychometrisch unpassende Items aus der Kern-Itemgruppe ausgeschlossen. Die Reliabilität der Kern-Itemgruppe ist nach Ausschluss dieser Items trotzdem geringfügig höher als die der optimierten Kern-Itemgruppe, d.h. die viel größere Itemanzahl überkompensiert die psychometrische Skalenoptimierung.
- Obwohl die Kern-Itemgruppe zehn Items mehr besitzt als die optimierte Gruppe, unterscheiden sich die Reliabilitäten im Bereich um  $\kappa = 1$  nur marginal voneinander.

Auch bei der Lage der Maxima wird das o.g. Zusammenspiel zwischen psychometrischer Optimierung und Testlänge deutlich: Die eingebettete Grafik mit Ausschnittsvergrößerung zeigt, dass die Maxima aller Itemgruppen in der Nähe von  $\kappa = 1$  liegen. Die effektive Anzahl der Items ist für  $\kappa = 1$  am größten, aber die MTF-Itemgruppe besitzt eine höhere Reliabilität als die MC3-Itemgruppe, wodurch sich das Maxima zu  $\kappa$ -Werten größer 1 verschiebt.

**Tab. 42:** Zusammensetzung der KiRC-Itemgruppen

Format <sup>a</sup>	Kern-Itemgruppe	opt. Kern-Itemgruppe	Rasch-Itemgruppe
MC3	9	8	6
MTF	23	14	12

*Bemerkung:* Zusammensetzung der Itemgruppen siehe Tab. 14 und Kap. 7.2.2.

<sup>a</sup>Gezählt werden einzelne TF-Items eines übergeordneten MTF-Items (vgl. Kap. 5.1.3).

### 11.4.3 Einfluss einer Ratekorrektur auf die Ergebnisse der Vergleichsuntersuchung

Die methodische Forschungsfrage 3.4 wurde bereits theoretisch beantwortet (Kap. 10.5.5): Eine Ratekorrektur ist ein Spezialfall einer  $\kappa$ -Gewichtung mit  $\kappa' = \beta/\alpha$ , vgl. Gl. (32) und Gl. (39), und erzeugt für beliebige Parameter  $\alpha, \beta$  folglich den gleichen Effektstärkebereich wie  $\kappa$ -gewichtete Testsummen. In dem hier vorliegenden Fall beträgt  $\beta/\alpha = 4/3$ .

Abbildung 41 illustriert am Beispiel der optimierten Kern-Itemgruppe, dass der  $d$ -Wertebereich zunächst ratekorrigierter und anschließend  $\kappa$ -gewichteter Testsummen identisch ist zum  $d$ -Wertebereich lediglich  $\kappa$ -gewichteter Testsummen. Die multiplikative Streckung durch die Ratekorrektur entspricht in der gewählten logarithmischen Darstellung einer Verschiebung der nicht-korrigierten Kurve (entlang der Abszisse). Daraus wird sofort ersichtlich, dass die Effektstärkebereiche beider Kurven übereinstimmen. Insbesondere stimmt die Effektstärke ratekorrigierter Werte bei  $\kappa = 3/4$  mit der Effektstärke überein, die durch eine gleichgewichtete, unkorrigierte Testsumme ( $\kappa = 1$ ) generiert wird (denn dann ist  $\kappa \cdot \beta/\alpha = 1$ ). Dies wird durch die vergrößerte Darstellung im eingebetteten Diagramm illustriert.

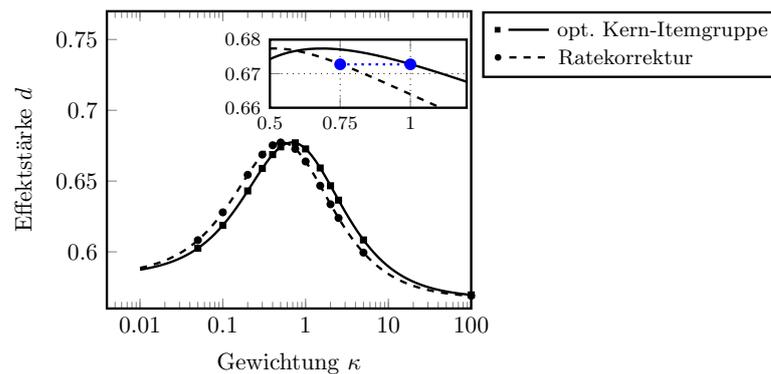


Abb. 41: Mischungsanalyse ratekorrigierter Testsummen der optimierten KiRC-Itemgruppe.

## 11.5 Ergebnisse zur Robustheitsfrage

In diesem Abschnitt werden die Ergebnisse zu den Forschungsfragen 4.1 bis 4.5 dargestellt. Dabei wird untersucht, welche Rolle verschiedene Einflussfaktoren auf die abhängigen Variablen der Hauptforschungsfrage (FF 4.1 und 4.2) bzw. auf den Lernzuwachs im Speziellen (FF 4.3–4.5) nehmen. In allen Fällen beschränkt sich die Analyse auf die erste Interventionsphase.

### 11.5.1 Einfluss des Geschlechts auf die Zielvariablen

Die in Kap. 11.1.5 definierten und in Kap. 11.2 berechneten Kovarianzmodelle wurden um den zusätzlichen Faktor „Geschlecht“ erweitert. Die Ergebnisberichte der Kovarianzanalyse beinhalten dann neben den Ergebnissen des ersten Haupteffekts (Faktor „Gruppe“) und der Varianzaufklärung der Kovariaten zusätzlich die durch den zweiten Hauptfaktor „Geschlecht“ sowie die durch die Interaktion „Gruppe  $\times$  Geschlecht“ aufgeklärte Varianz. Die Statistiken des zweiten Haupteffekts geben Aufschluss darüber, ob sich die abhängige Variable geschlechtsspezifisch unterscheidet (unter Berücksichtigung der Kovariaten im Modell), die des Interaktionseffekts darüber, ob sich die Intervention auf das Geschlecht unterschiedlich auswirkt. Durch die Erweiterung der Modelle um den Faktor „Geschlecht“ ändert sich die totale aufgeklärte Varianz und damit auch die Fehlervarianz. Folglich ändern sich auch die Ergebnisse des Haupteffekts „Gruppe“ sowie der Kovariaten (jeweils geringfügig). Durch eine Überprüfung konnte ausgeschlossen werden, dass sich die Ergebnisse der Haupteffekte zwischen den ein- und zweifaktoriellen Modellen wesentlich unterscheiden. Deshalb beschränken sich die Ergebnisberichte auf die Varianzaufklärung durch den zweiten Haupteffekt „Geschlecht“ sowie den Interaktionseffekt, siehe Tab. 43.

Die zweifaktoriellen Kovarianzanalysen ergeben mit Ausnahme der IE-Skala keine signifikanten Haupt- oder Interaktionseffekte hinsichtlich der motivationalen abhängigen Variablen und der Neugierde. Bei den beiden kognitiven Variablen (Repräsentationskompetenz und Konzeptverständnis) zeichnen sich Tendenzen bzgl. des Haupteffekts mit kleiner Effektstärke ab. Darüber hinaus zeigen sich schwach signifikante Interaktionseffekte bzgl. der Repräsentationskompetenz und des Interesses / Engagements mit ebenfalls kleinen Effektstärken. Zur inhaltlichen Deutung der Ergebnisse sind die deskriptiven Daten in Tab. 44 aufgeführt. Frauen zeigen niedrigere Ausprägungen der Repräsentationskompetenz und des Konzeptverständnisses als Männer. Zwar schneiden sowohl Männer als auch Frauen in der Interventionsgruppe besser ab als in der Kontrollgruppe, aber im Gruppenvergleich verzeichnen die Frauen höhere Lernzuwächse. Dieser Interaktionseffekt wird wegen der geringeren Messgenauigkeit für die Konzept-Variable nicht signifikant, kann aber tendenziell bestätigt werden. Das Item Kon2, welches sich als nicht Rasch-konform herausstellte (Verletzung gegen die Gruppenhomogenität, Teilungskriterium Geschlecht; vgl. Kap. 7.3), wurde einzeln analysiert. Bei diesem Item zeigte sich auch innerhalb der Kern-Stichprobe ein großer

Unterschied in den Lösungswahrscheinlichkeiten zu Beginn der Studie zwischen Männern ( $P(t_0) = 0.76$ ) und Frauen ( $P(t_0) = 0.37$ ), der auch nach der Intervention Bestand hat ( $P(t_1) = 0.87$  bzw.  $P(t_1) = 0.42$ ). Die Ergebnisse auf Stichprobenebene zeigen, dass bei diesem Item Männer einen viel größeren Lernzuwachs ( $g_H = 0.45$ ) erreichen als Frauen ( $g_H = 0.08$ ). Eine differenzierte Betrachtung der Ergebnisse auf Gruppenebene erbrachte, dass videobasierte Aufgaben bei diesem Item zu keiner besonderen Förderung von weiblichen Studenten führten.

Bezüglich des Interesses werden insgesamt keine Geschlechtsunterschiede festgestellt; der Haupteffekt ist insignifikant. Die deskriptiven Daten bzgl. des Interaktionseffekts zeigen, dass weibliche Studierende in der Interventionsgruppe nach dem ersten Interventionszeitraum ein höheres Interesse zeigen als ihre Kommilitoninnen der Kontrollgruppe, während sich die Interessensausprägung männlicher Studenten zwischen den Gruppen nicht unterscheidet. Schließlich sind in Tab. A-13 noch die Ausprägungen von Vorleistungen und relevanter Prä-Variablen geschlechtsspezifisch aufgeführt, die spätere Diskussion (Kap. 12.1) ergänzen.

**Tab. 43:** Ergebnisse zweifaktorieller Kovarianzanalysen

AV (Post1-Zeitpunkt)	Geschlecht			Geschl. × Gruppe		
	<i>F</i>	<i>p</i>	<i>d</i>	<i>F</i>	<i>p</i>	<i>d</i>
Repräsentationskompetenz	3.53	.065	(.14)	3.62	.061	(.26)
Konzepttest	3.32	.073	(.21)	2.44	.123	-
Realitätsbezug	.45	.504	-	1.36	.248	-
Interesse / Engagement	.02	.886	-	3.83	.054	.30
Selbstkonzept	1.96	.166	-	.06	.801	-
Disziplinäre Authentizität	1.01	.318	-	.01	.942	-
Experimentbezug	.00	.980	-	.23	.625	-
Autonomie	1.09	.299	-	.52	.473	-
Neugierde (aktueller Zustand)	.00	.962	-	.38	.541	-

*Bemerkungen:* Kovarianzmodelle siehe Tab. 37. Die Varianzaufklärung durch Kovariaten ist nicht aufgeführt ist. Die Ergebnisse unter dem Faktor „Geschlecht“ beziehen sich auf den Vergleich der AV zwischen Männer und Frauen, die Ergebnisse unter dem Faktor „Geschlecht × Gruppe“ auf Interaktionseffekte. In Klammern angegebene Effektstärken stehen in Zusammenhang mit einer Tendenz zur statistischen Signifikanz des zugehörigen Tests ( $0.05 < p < 0.10$ ).

**Tab. 44:** Deskriptive Daten zu Gender-Effekten

AV		Gesamt			Gruppenvergleich				
		<i>N</i>	<i>M*</i>	$\sigma_N$		<i>N</i>	<i>M*</i>	$\sigma_N$	
KiRC	♂	55	.82	.01	KG	♂	29	.77	.02
	♀	20	.77	.02		♀	11	.68	.03
					IG	♂	26	.86	.02
						♀	9	.86	.03
Konzepttest	♂	51	.70	.02	KG	♂	28	.67	.03
	♀	18	.61	.04		♀	10	.50	.05
					IG	♂	23	.73	.04
						♀	8	.71	.07
Interesse / Eng.	♂	56	.70	.01	KG	♂	29	.70	.11
	♀	20	.70	.02		♀	11	.64	.12
					IG	♂	27	.70	.14
						♀	9	.80	.12

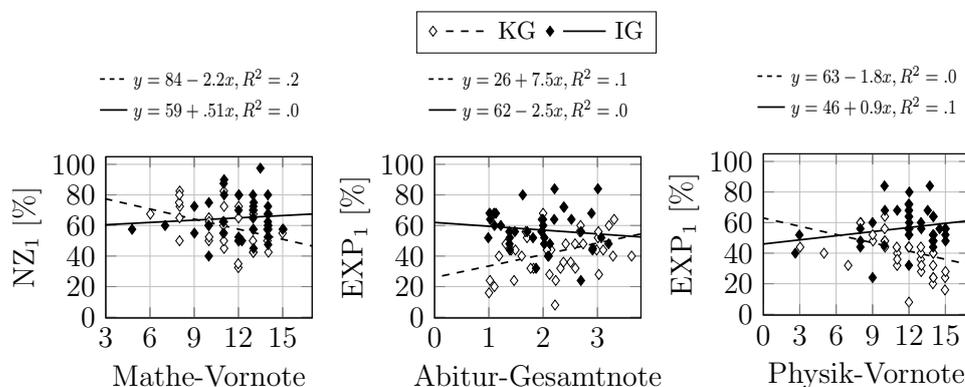
*Bemerkungen:*  $M^*$  = auf Kovariaten adjustierte Mittelwerte,  $\sigma_N$  = Standardfehler des Mittelwerts.

### 11.5.2 Einfluss der Vorleistungen auf die Zielvariablen

Der Einfluss der Vorleistungen auf die Zielvariablen wurde teilweise im Rahmen der Voranalyse implizit überprüft. So geht aus Tabelle 34 hervor, dass die Abiturnote mit der Repräsentationskompetenz korreliert ( $r = -0.33$ ) und die Mathematik-Vorleistung sowohl mit der Repräsentationskompetenz ( $r = 0.38$ ) als auch mit dem Konzeptverständnis ( $r = 0.25$ ) zusammenhängt. Ebenso wurde implizit überprüft, ob sich die Zusammenhänge zwischen den Vorleistungen und den Zielvariablen zwischen den Gruppen unterscheiden (Interaktionseffekt „Gruppe  $\times$  Kovariate“). Ein Verstoß gegen die Homogenität der Regressionssteigungen wäre nämlich gleichbedeutend damit, dass ein unterschiedlich starker Zusammenhang zwischen der betrachteten Kovariate und der betrachteten abhängigen Variable zwischen IG und KG besteht. Aus der Voranalyse geht hervor, dass solche Modellverletzungen bei den o.g. Variablenpaaren nicht vorliegen.

Es bleibt übrig zu untersuchen, ob Interaktionseffekte bei den KV-AV-Paaren vorliegen, die nicht in der Voranalyse betrachtet wurden. Die Ergebnisse sind in Tab. A-8 im Anhang aufgeführt. Bei fast allen abhängigen Variablen zeigt sich die Intervention robust gegenüber den getesteten Einflussfaktoren. Dies bedeutet, dass die Intervention zu keinem unterschiedlichen Zusammenhang zwischen den abhängigen Variablen und den Vorleistungen innerhalb der Untersuchungsgruppen führte. Ausnahmen bilden die Neugierde und der Experimentbezug. Die Neugierde als aktueller Zustand (NZ) fällt in der KG ab, je besser die Mathematik-Vorleistung der Studierenden ist. In der IG besteht hingegen kein Zusammenhang zwischen NZ und Mathe-Vorleistung. Dieser Interaktionseffekt „Gruppe  $\times$  Mathe-Vorleistung“ ist signifikant ( $F(1, 71) = 4.20$ ,  $p = 0.044$ ,  $d = 0.40$ ).

Der Experimentbezug steigt in der IG mit wachsender Vorleistung (bessere Abiturnote, bessere Physik-Vorleistung) an, während er in der KG abfällt. Die beiden Interaktionseffekte „Gruppe  $\times$  Abiturnote“ bzw. „Gruppe  $\times$  Physik-Vorleistung“ sind von kleiner bis mittlerer Größe ( $F(1, 71) = 4.59$ ,  $p = 0.036$ ,  $d = 0.39$  bzw.  $F(1, 70) = 5.91$ ,  $p = 0.018$ ,  $d = 0.51$ ). Die drei Streudiagramm in Abb. 42 verdeutlichen die beschriebenen Zusammenhänge, die im Diskussionskapitel (Kap. 12.1) interpretiert werden.



**Abb. 42:** Interaktionseffekte zwischen abhängigen Variablen und Vorleistungen: Nach der ersten Interventionsphase zeigen sich zwischen den Gruppen unterschiedliche Zusammenhänge zwischen NZ und Mathematik-Vorleistung (links), zwischen Exp und Abiturnote (mittig), sowie zwischen Exp und Physik-Vorleistung (rechts). Zu berücksichtigen ist die umgekehrte Polung der Abiturnote im Vergleich zu den Vorleistungsindikatoren.

### 11.5.3 Einflussfaktoren auf den Lernzuwachs

Ähnlich zum Vorgehen im vorangegangenen Abschnitt wird hier untersucht, ob der Lernzuwachs gegenüber verschiedenen Einflussparametern robust ist. Unter Lernzuwachs wird in

diesem Abschnitt die ungewichtete Differenz der Prä- und Post-Mittelwerte der Leistungstests verstanden (Konzepttest bzw. KiRC-Test mit optimierter Kern-Itemgruppe). Wiederum werden Interaktionseffekte zwischen den Einflussfaktoren und dem Faktor „Gruppe“ bzgl. den abhängigen Variablen (Lernzuwächse) untersucht. Tabelle 45 fasst die Ergebnisse für beide Leistungstests zusammen. Der CAQ-Index, der wegen der Testkürze nicht für den Konzepttest berechnet werden kann (siehe Kap. 8.3.1), dient als Maß für die Adäquatheit der Selbsteinschätzung und wurde in Kap. 8.1 eingeführt (Gl. (22)).

Der repräsentationale Lernzuwachs zeigt sich robust gegenüber allen getesteten Kovariaten, d.h. weder das Selbstkonzept, noch die empfundene Aufgabenschwierigkeit, Beanspruchung oder die Bearbeitungszeit der Aufgaben stehen im Zusammenhang mit dem Lernzuwachs. Insbesondere existiert auch kein Zusammenhang zwischen der Adäquatheit der Selbsteinschätzung und dem Leistungszuwachs. Allerdings wurde ein signifikanter Zusammenhang zwischen dem konzeptionellen Lernzuwachs und der Bearbeitungszeit der Aufgaben festgestellt, der praktisch bedeutsam ist ( $F(1, 59) = 6.27$ ,  $p = 0.015$ ,  $d = 0.58$ ). Ferner zeigt sich sowohl für den Repräsentations- als auch für den Konzepttest, dass sich der Zusammenhang zwischen den Variablen innerhalb der Untersuchungsgruppen nicht signifikant voneinander unterscheidet.

**Tab. 45:** Ergebnisse zur Robustheit des Lernzuwachses

KV	„Gesättigtes Modell“ <sup>a</sup>		Interaktionseffekt <sup>b</sup>	
	<i>F</i>	<i>p</i>	<i>F</i>	<i>p</i>
<u>AV = Lernzuwachs KiRC</u>				
SK <sub>0</sub>	1.63	.206	.60	.440
CL <sub>0</sub>	1.70	.197	.90	.346
PSQ <sub>0</sub>	1.41	.240	.08	.775
ToT <sub>1</sub>	2.21	.142	.60	.443
CAQ <sub>0</sub>	.01	.927	.11	.740
<u>AV = Lernzuwachs Konzepttest</u>				
SK <sub>0</sub>	2.45	.123	.20	.656
CL <sub>0</sub>	.21	.650	.04	.842
PSQ <sub>0</sub>	.68	.413	.08	.777
ToT <sub>1</sub>	6.27	.015*	.00	.966

<sup>a</sup>Test auf einen Zusammenhang zwischen AV (= abhängige Variable) und KV (= Kovariate). <sup>b</sup>Testet, ob es zwischen den Gruppen einen unterschiedlich starken Zusammenhang zwischen KV und AV gibt.

\*Die Effektstärke beträgt  $d = 0.58$ .

## 11.6 Analyse der Klausurergebnisse

In diesem Abschnitt werden zunächst die Zusammenhänge zwischen dem Klausurergebnis und den kognitiven abhängigen Variablen der Intervention sowie den Vorleistungen analysiert, um anschließend Gruppenunterschiede zu untersuchen. Da zum Zeitpunkt der Klausur (am Semesterende) jeder Student mit Videoanalyse-Aufgaben gearbeitet hat, werden zwar keine gruppenspezifischen Leistungsunterschiede in dem Gesamtergebnis erwartet, aber es wurde versucht, einzelne Klausuraufgaben den Interventionsphasen zuzuordnen.

Deshalb bestand ein erster Schritt darin, die in den beiden Semestern (WS 13/14 und 14/15) gestellten Klausuraufgaben thematisch zu strukturieren. Beide Klausuren beinhalteten fünf Aufgaben über experimentalphysikalische Themen, von denen sich die ersten vier Aufgaben semesterübergreifend sehr ähnelten: Die erste Aufgabe (A1) behandelte die Kinematik (Flugbewegung eines Skispringers; schiefer Wurf), A2 thematisierte ein- bzw. zweidimensionale Stoßvorgänge, A3 erforderte Kenntnisse über die Dynamik starrer Körper (Rotation um raumfeste Achse, Trägheitsmomente) und A4 behandelte Reibungsvorgänge

(wirkende Kräfte, Bewegungsgleichung). Die letzte Aufgabe wurde im WS 13/14 zum Thema Schwingungen und Wellen gestellt, behandelte im WS 14/15 jedoch die Hydromechanik (Auftrieb, Bernoulli-Gleichung). Deshalb wurde davon abgesehen, die Daten von A5 zur Berechnung von Korrelationen auf Stichprobenebene zu aggregieren. Die Reihenfolge der Aufgaben entsprach in etwa der inhaltlichen Abfolge im zeitlichen Verlauf der Vorlesung, wobei die Aufgaben A2–A5 themenübergreifende Unteraufgaben besaßen. Aufgabe 1 hingegen beschränkte sich im Wesentlichen auf die Inhalte und Konzepte der Kinematik und kann deshalb eindeutig der ersten Interventionsphase zugeordnet werden. Neben den fünf Aufgaben zur EP1 umfasste die Klausur fünf theoretische Aufgaben zu den mathematischen Grundlagen der Physik („MGdP“). Im weiteren Ergebnisbericht wird deshalb zwischen dem Klausur-Gesamtergebnis (EP1- und MGdP-Aufgaben) und den Teilergebnissen differenziert. Es ist zu berücksichtigen, dass die Gruppenmittel nicht unmittelbar aus den deskriptiven Daten der Teilergebnisse abgeleitet werden können, da sich die Anzahl bearbeiteter MGdP-Aufgaben zwischen den Studierenden aufgrund unterschiedlicher Prüfungsanforderungen der verschiedenen Studiengänge unterscheidet.

In einem zweiten Schritt wurden die Korrelationen zwischen den Klausuraufgaben zu den abhängigen Variablen und Kovariaten bestimmt, siehe Tab. 46. In der Tat zeigt sich, dass die erste Klausuraufgabe A1 (mit Inhalt Kinematik) am stärksten mit dem KiRC-Test ( $r = 0.42$ ,  $p < 0.01$ ) und als einzige der fünf Klausuraufgaben mit dem Konzepttest zu Bewegungen korreliert ( $r = 0.44$ ,  $p < 0.01$ ). Dabei wurden jeweils die Testergebnisse verwendet, die zeitlich am nächsten an der Klausur lagen (KiRC<sub>1</sub> und Kon<sub>2</sub>). Auf Ebene der Gesamtergebnisse zeigt sich, dass die Klausur hochsignifikant mit der Abiturnote<sup>4</sup>, der Vorleistung in Mathematik und dem KiRC-Test korreliert. Dies gilt sowohl für das Klausurergebnis unter Einschluss aller Aufgaben als auch für die beiden Teilergebnisse (EP1 und MGdP). Darüber hinaus weist das EP1-, nicht aber MGdP- und Gesamtergebnis eine signifikante Korrelation zu dem Konzeptverständnis auf.

**Tab. 46:** Zusammenhang zwischen Klausur(aufgaben) und kognitiven Variablen

	Vorleistungen				
	Abiturnote	Mathe	Physik	KiRC <sub>1</sub>	Kon <sub>2</sub>
Klausur (gesamt) <sup>a</sup>	-.51**	.50**	.20	.50**	.15
Klausur (nur EP1)	-.45**	.43**	.15	.40**	.31*
Klausur (nur MGdP)	-.56**	.58**	.19	.37*	.05
A1: Kinematik	-.18	.11	.00	.42**	.44**
A2: Stöße	-.30*	.25	.24	.27*	.12
A3: Dynamik	-.42**	.49*	.14	.22	.18
A4: Reibung	-.38**	.28*	.08	.29*	.24

*Bemerkungen:* Es sind die zur Klausur zeitnächsten Datensätze der kognitiven Variablen aufgeführt. Die Klausur besteht aus fünf Aufgaben (A1–A5) zur EP1 (Experimentalphysik 1) sowie fünf Aufgaben zu den mathematischen Grundlagen der Physik (MGdP). Aufgabe A5 ist semesterübergreifend thematisch verschieden und damit nicht vergleichbar. Die Stichprobengröße variiert zwischen 53 und 57.

\* $p < 0.05$ , \*\* $p < 0.01$ .

Schließlich zeigt Tab. 47 die Ergebnisse auf Aufgaben- und Klausurebene im Gruppenvergleich. Bei Aufgabe A1 zeichnet sich die Tendenz ab, dass die Gruppe, die in der ersten Semesterhälfte mit Videoanalyse arbeitete (IG-KG-Gruppe), eine höhere Punktzahl erreicht als die andere Gruppe ( $t(55) = 1.95$ ,  $p = 0.057$ ,  $d = 0.45$ ). Sonstige Effekte auf Klausuraufgabenebene liegen nicht vor. Es fällt jedoch auf, dass die Differenz der Mittelwerte abnimmt, je weiter der Inhalt im zeitlichen Verlauf der Lehrveranstaltung voranschreitet. Zudem wurde ein schwach signifikanter Gruppenunterschied bzgl. des Gesamtergebnisses festgestellt ( $t(55) = 1.90$ ,  $p = 0.063$ ,  $d = 0.43$ ). Auch hier erreicht die IG-KG-Gruppe ein höheres Ergebnis als die KG-IG-Gruppe.

<sup>4</sup>An dieser Stelle sei erneut auf die umgekehrte Polung der Abiturnote hingewiesen.

**Tab. 47:** Deskriptive Statistiken der Klausuraufgaben und Gruppenvergleich

Klausuraufgabe	KG-IG		IG-KG		$\Delta$	$t$ -Wert	$p$	$d$
	$M$	$SD$	$M$	$SD$				
A1: Kinematik	.80	.21	.90	.16	.10	1.95	.057	.45
A2: Stöße	.59	.24	.67	.24	.09	1.36	.181	-
A3: Dynamik	.69	.29	.69	.23	.00	.21	.983	-
A4: Reibung	.43	.31	.41	.22	-.02	.32	.748	-
Gesamt (EP1)	.61	.20	.63	.13	.02	.53	.601	-
Gesamt (EP1 & MGdP)	.55	.20	.64	.15	.09	1.90	.063	.43

## 11.7 Korrelationsanalysen

Als Ergänzung des Ergebnisberichts sind in Tabelle 48 die Zusammenhänge zwischen den abhängigen Variablen zu den beiden Post-Zeitpunkten berichtet. Die Berechnung erfolgte auf Stichprobenebene ( $N = 76$ ). Es wurden zu beiden Testzeitpunkten substantielle Korrelationen ( $0.36 \leq r \leq 0.60$ ) zwischen der Neugierde (NZ) und den Motivationskomponenten festgestellt (mit Ausnahme von SK). Außerdem korrelieren die Motivationskomponenten untereinander — hervorzuheben sind die hohen Korrelationskoeffizienten zwischen den Authentizitäts-Skalen (Realitätsbezug, disziplinäre Authentizität und Experimentbezug). Die beiden Leistungsvariablen Repräsentationskompetenz und Konzeptverständnis korrelieren sehr hoch miteinander ( $r = 0.62$ ), weisen aber keine Zusammenhänge zu den anderen Variablen auf.

**Tab. 48:** Korrelationsmatrix zwischen den AV zu  $t_1$  (obere Hälfte) und  $t_2$  (untere Hälfte)

	CL	PSQ	NZ	RA	IE	SK	dA	Atnm	Exp	ToT	KiRC	Kon
CL	■	.24	.07	-.03	-.07	-.37	.06	-.03	-.13	.29	-.04	-.03
PSQ	.34	■	-.02	.16	-.16	-.07	.21	.23	-.01	.42	.23	.17
NZ	.20	.24	■	.55	.54	.06	.53	.38	.48	-.18	-.21	-.13
RA	.23	.39	.57	■	.51	.20	.63	.32	.51	-.14	.08	.01
IE	-.03	.21	.60	.51	■	.39	.56	.35	.30	-.16	.05	-.12
SK	-.18	.08	.36	.24	.50	■	.12	.06	.16	-.22	.17	.04
dA	.13	.42	.51	.70	.65	.40	■	.33	.34	-.08	.05	-.04
Atnm	.17	-.10	.37	.44	.19	.36	.35	■	.31	.34	-.20	-.02
Exp	.14	.14	.45	.34	.03	.10	.18	.24	■	-.06	.02	.09
ToT	.31	.42	.17	.10	.16	-.01	.09	-.11	.01	■	.09	-.07
KiRC	-	-	-	-	-	-	-	-	-	-	■	.62
Kon	.23	.18	.10	.02	.11	.05	.10	.22	-.02	.06	-	■

*Bemerkungen:* Oberhalb der Diagonalen sind die (Pearsonschen) Korrelationskoeffizienten der abhängigen Variablen (= AV) zum Post-Testzeitpunkt  $t_1$ , unterhalb zum Zeitpunkt  $t_2$  angegeben. Korrelationen mit  $r \geq 0.25$  sind auf dem 5%-Niveau signifikant. Korrelationen mit Betrag  $< 0.25$  sind aus Übersichtsgründen kontrastarm dargestellt. Die Korrelationen zwischen den affektiven Variablen sind — ebenfalls aus Übersichtsgründen — eingerahmt. Der äußere Rahmen beinhaltet die Korrelationen aller affektiven Variablen untereinander, der innere im Speziellen die der Motivationskomponenten



# Kapitel 12

## Diskussion

In diesem Kapitel werden zunächst die Ergebnisse zu den Hauptfragestellungen diskutiert (Kap. 12.1 und 12.2). Die Diskussion wird für beide Interventionszeiträume getrennt geführt, wobei der Interpretationsschwerpunkt dabei nach den in Kap. 10.2 genannten Gründen auf dem ersten Interventionszeitraum liegt. Die in Kap. 9 formulierten Hypothesen zur Lernwirksamkeit und zur Belastung werden der Reihe nach aufgegriffen und die methodischen Fragestellungen sowie die Robustheitsfragen den zugehörigen Variablen zugeordnet. Nach Diskussion der beiden Interventionsphasen folgt die Interpretation der Klausurergebnisse unter Beachtung der eingeschränkten Aussagekraft wegen des Studienverlaufsplans (Kap. 12.3). Tabelle 49 bietet einen Überblick über die Ergebnisse.

**Tab. 49:** Hypothesen und Ergebnisse im Überblick

<i>Videobasierte Aufgaben fördern ...</i>		Phase 1		Phase 2	
		hyp-konform?	<i>d</i>	hyp-konform?	<i>d</i>
1.1	Repräsentationskompetenz	ja	0.72	nicht überprüft	
1.2	Konzeptverständnis	ja	0.34	nein	-
-	(Klausur-Ergebnis) <sup>a</sup>	(ja)	(0.45)	-	-
1.3	Motivation				
	Realitätsbezug	ja	0.37	ja	0.41
	Interesse / Engagement	nein	-	nein	-
	Selbstkonzept	nein	-	nein	-
	Disziplinäre Authentizität	ja	0.25	nein	-
	Exp. (auch Hyp 3.1) <sup>b,c</sup>	ja	0.75	ja	1.63
	Autonomie	nein	-	nein	-
1.4	Neugierde	ja	0.39	nein	-
<i>... besser als inhaltsähnliche traditionelle Aufgaben</i>					
<i>Die Untersuchungsgruppen unterscheiden sich nicht bzgl. der wahrgenommenen ...</i>					
2.1	Aufgabenschwierigkeit	ja	-	nein <sup>d</sup>	0.55
2.2	Beanspruchung	ja	-	nein <sup>d</sup>	0.44
2.3	Aufg.-Bearbeitungszeit	ja	-	ja	-

*Bemerkung:* Effektstärken *d* werden bei signifikanten Gruppenunterschieden berichtet.

<sup>a</sup>Das Ergebnis bezieht sich auf eine Teilaufgabe (Kinematik), die inhaltlich der ersten Interventionsphase zugeordnet werden kann. Da bzgl. der Klausur-Leistung keine explizite Hypothese formuliert wurde, sind die Ergebnisse in Klammern angegeben. <sup>b</sup>Die Variable Experimentbezug (= Exp) nimmt eine Doppelrolle als abhängige Variable und als Variable des Manipulation Checks ein, vgl. Kap. 10.4. <sup>c</sup>Angegeben sind die korrigierten Wert gemäß Tab. 41, die bedeutend niedriger als die unkorrigierten Werte ausfallen. <sup>d</sup>Videobasierte Aufgaben führten in Phase 2 zu einer geringeren Belastung.

Es schließt sich eine Interpretation der Vergleichsanalyse zwischen Studienabbrechern und Weiterstudierenden an, die losgelöst von den Hauptfragestellungen durchgeführt wurde (Kap. 12.4). Das Kapitel endet mit Bemerkungen zur Generalisierbarkeit der gefundenen Ergebnisse und zu Einschränkungen, denen die Studie unterworfen ist.

## 12.1 Erster Interventionszeitraum

Zunächst kann festgehalten werden, dass der „Manipulation Check“ erfolgreich war, d.h. die unterschiedlich konstruierten Instruktionsmaterialien wurden von den Lernenden wie intendiert wahrgenommen: Die IG beurteilte den Experimentbezug der Übungsaufgaben signifikant höher als die KG (großer Effekt,  $d = 1.02$ ; vgl. Kap. 11.4.1). Neben der methodischen Absicherung der Existenz substantiell verschiedener Faktorstufen der unabhängigen Variable (Gruppe) bedeutet dieses Ergebnis inhaltlich, dass eine wichtige Teilkomponente disziplinären Arbeitens durch videobasierte Aufgaben besser abgebildet wird als durch traditionelle Aufgaben. Dazu gehören die Entwicklung weitergehender Fragen zu einem physikalischen Problem, das Formulieren eigener Hypothesen, das Durchführen experimenteller Tätigkeiten und die Erhebung experimenteller Daten. Zwar kann aus dem Ergebnis nicht geschlossen werden, in welchem Maße videobasierte Aufgaben diese experimentellen Fähigkeiten und Kompetenzen förderten, wohl aber, dass sie eine Auseinandersetzung mit experimentellen Prozessen anregten.

### 12.1.1 Repräsentationskompetenz

Zur Überprüfung der Hypothese 1.1 wurde das KiRC-Instrument zur Erfassung repräsentationaler Fähigkeiten unmittelbar vor und nach der Intervention eingesetzt. Der folgenden Diskussion liegen die Testsummen der optimierten KiRC-Itemgruppe zugrunde; die Ergebnisse bzgl. den anderen Itemgruppen und der Testmischungsanalyse werden weiter unten diskutiert.

#### Haupteffekt

Die Ergebnisse in Kap. 11.2.1 zeigen hypothesenkonform, dass die IG nach der Intervention eine signifikant größere Repräsentationskompetenz aufweist als die KG (mittelgroßer bis großer Effekt,  $d = 0.72$ ). Zwar steigerte auch die KG ihre Leistung im Laufe der Intervention und erreichte einen positiven normierten Lernzuwachs ( $g_{H,KG} = 0.23$ ), dieser fiel jedoch bedeutend niedriger aus als in der IG ( $g_{H,IG} = 0.57$ ). Die erzielten Hake-Indizes stehen in guter Übereinstimmung mit den Werten, die in der Literatur für traditionelle ( $g \approx 0.2$ ) bzw. innovative Lehrmethoden ( $g = 0.3 - 0.6$ ) berichtet werden (Coletta et al., 2007). Eine differenzierte Betrachtung der Testergebnisse nach Repräsentationsform und Inhalt der verwendeten KiRC-Items ergab, dass die IG in jeder Kategorie bessere oder zumindest gleichwertige Lernzuwächse erzielte als die KG. Besonders deutliche Gruppenunterschiede bestehen bei Aufgabeninhalten, die (i) den Umgang mit Bezugssystemen erfordern und bei denen (ii) Vektorkomponenten eine Rolle spielen sowie bei Aufgaben, die (iii) bildhafte Repräsentationsform beinhalten und bei denen (iv) bildhafte und formale Inhalte verknüpft werden. Diese Ergebnisse spiegeln die Lernwirkung wesentlicher Elemente der videobasierten Instruktion wider und lassen sich mit Bezug auf den lerntheoretischen Hintergrund multipler Repräsentationen interpretieren:

- (i) Die explizite Festlegung von Bezugssystemen ist eine notwendige Voraussetzung zur mathematischen Beschreibung der Bewegung eines Objekts, die zu Beginn jeder Videoanalyse durch *Skalierung* des Videos zu leisten ist (vgl. Kap. 2.1.1). Dies erfordert eine kognitive Auseinandersetzung mit der experimentellen Situation in Bezug auf die Aufgabenlösung, beispielsweise, um die Bewegung in nur einer Dimension zu beschreiben. Die Lernwirkung lässt sich auf diese kognitive Aktivierung bei der Explizitheit der Wahl

des Bezugssystems zurückführen und steht in Einklang mit Ergebnissen von Koleza und Pappas (2008).<sup>1</sup>

- (ii) Die Erklärung der hohen Lernleistung bzgl. Vektorzerlegungen folgt einem ähnlichen Argument: Die Vektoren im Realraum werden durch die Videoanalyse immer in Bezug zu einem Koordinatensystem betrachtet — alle Bewegungsinformationen liegen als  $(x, y)$ -Koordinatenpaare vor. Die Selektion der relevanten Bewegungskomponente sowie das Verrechnen beider Komponenten zu einer physikalischen Größe (z.B. Bahngeschwindigkeit, kinetische Energie, Schwerpunkt,...) erfordern jeweils die explizite Auseinandersetzung mit beiden Teilkomponenten. Beispielsweise kann bei traditionellen Aufgaben der vektorielle Charakter eindimensionaler Bewegungen formal vernachlässigt werden, wohingegen die weggedachte Komponente bei der Videoanalyse stets präsent ist. Eine Auseinandersetzung mit beiden Komponenten ist demnach zwingend erforderlich, um die Messergebnisse richtig einordnen zu können. Die visuelle Einblendung von Vektoren unterstützt zudem die Erfassung des Richtungscharakters kinematischer Größen und wurde beispielsweise in den Aufgaben 2.2 und 5.1 explizit gefordert (vgl. Instruktionsmaterial im Anhang, ab S. 220). Wie die Ergebnisse zeigen, stellte sich dies besonders bei Items zur Rotationskinematik als hilfreich heraus.
- (iii) Der hohe Lernzuwachs bei bildhaften Aufgabenformaten begründet sich theoriekonform in der Darstellungsart der Inhalte. Da sich Studierende der IG vermehrt mit realen Abbildern von Experimenten auseinandersetzten waren sie im Stande, bildhafte Informationen zu physikalischen Prozessen besser zu interpretieren als die Kontrollgruppe.
- (iv) Darüber hinaus waren Studierende der IG erfolgreicher, bildhafte Inhalte mit formalen Inhalten zu verknüpfen. Gemäß der CATLM unterstützt die Videoanalyse durch die zeitsynchrone Präsentation visueller Informationen unterschiedlichen Abstraktionsgrades (reales Bild, Diagramm, Vektor, Formel) die Verarbeitungstiefe und Vernetzungsdichte der präsentierten Inhalte (vgl. Kap. 3.1.3).

Auch wenn das Ergebnis für eine gesamtheitliche Förderung repräsentationaler Kompetenz spricht und theoretisch gut begründet werden kann, führte die Intervention nur zu einem geringfügigen Gruppenunterschied bzgl. des Verständnisses von Diagrammen. Dieses Ergebnis überrascht zunächst, denn einerseits legen die Ergebnisse von Beichner (1996) eine Lernwirkung gerade in diesem Bereich nahe und andererseits wurde der Umgang mit Diagrammen an mehreren Stellen explizit durch das Instruktionsmaterial angeleitet. Die hohen Lösungswahrscheinlichkeiten der Diagramm-Items zum Prä-Zeitpunkt zeigen, dass bereits vor der Intervention ein Deckeneffekt vorliegt ( $P_{KG} = 0.79$ ,  $P_{IG} = 0.80$ ). Um die Wirkung videobasierter Aufgaben auf das Verständnis und den Umgang mit graphischen Repräsentationsformen zu messen, sind also Anpassungen am KiRC-Instrument notwendig.

### Beeinflussung des Ergebnisses durch Wahl der Itemgruppe und Testmischung

Im Rahmen der Forschungsfragen 3.2–3.4 wurde untersucht, welchen Einfluss die Wahl der Itemgruppe bzw. die Mischung der Teilsummen beider KiRC-Testhälften zu einer Gesamttestsumme ( $T = T_1 + \kappa T_2$ , vgl. Kap. 10.5.5) auf das Ergebnis der Vergleichsstudie nimmt. Im Gegensatz zu Kovarianzanalysen berücksichtigt die Behandlung gemischter Testsummen keine Kovariaten, wodurch tendenziell geringere Effektstärken als oben berichtet ( $d = 0.72$ ; optimierte Kern-Itemgruppe) zu erwarten sind.

Die Vergleichsanalyse ergab für alle betrachteten Itemgruppen mindestens mittelgroße Effekte zu Gunsten der Interventionsgruppe ( $0.50 < d \leq 0.72$ ). Der Effektstärken-Unterschied zwischen den Itemgruppen (Rasch-, optimierte Kern- und Kern-Itemgruppe) hängt zwar lokal von der Gewichtung  $\kappa$  ab, ist aber für alle  $\kappa$ -Werte nur geringfügig ausgeprägt ( $< 0.1$ ).

<sup>1</sup>In dem Artikel „The right frame of reference makes it simple [...]“ wird auf die Rolle des Koordinatensystems bei der Videoanalyse sehr detailliert eingegangen (Klein et al., 2015c).

Aus der Analyse ging weiter hervor, dass die Reliabilitäten der eingesetzten Itemgruppen sowohl von der Anzahl der Items als auch von deren psychometrischen Qualität abhängt. So ließe sich mit einem ersten Blick auf Abb. 40 vermuten, dass die aufwendige Vorarbeit der Testvalidierung ungerechtfertigt sei — schließlich führt die Rasch-Itemgruppe zu den geringsten Reliabilitäten und die beiden Kern-Itemgruppen unterscheiden sich im relevanten Gewichtungsbereich ( $\kappa = 0.5 - 1.0$ ) nur unwesentlich voneinander. Es muss allerdings beachtet werden, dass die optimierte Kern-Itemgruppe mit 22 Items wesentlich ökonomischer ist als die Kern-Itemgruppe (32 Items) und dass die Rasch-Itemgruppe (18 Items) weitaus strengere psychometrische Bedingungen erfüllt als die beiden klassischen Itemgruppen.

Das entwickelte analytische Verfahren zur Behandlung gemischter Testsummen führt damit zu einer verstärkten methodischen Absicherung der Ablehnung der Nullhypothese 1.1 (die in dieser Arbeit stets als Alternativhypothese formuliert ist). Außerdem wurde sowohl analytisch gezeigt als auch empirisch bestätigt, dass in der Betrachtung gemischter Testsummen der Spezialfall einer Ratekorrektur enthalten ist. Dies bedeutet, dass die Ergebnisse gegenüber Durchführung einer Ratekorrektur beständig sind und praktisch bedeutsam bleiben.

### **Robustheit des Ergebnisses gegenüber verschiedenen Einflussfaktoren**

Im Rahmen der Forschungsfragen 4.1–4.5 wurde untersucht, ob die gefundenen Unterschiedseffekte bzgl. Repräsentationskompetenz von Persönlichkeitseigenschaften oder Vorleistungen abhängen und ob sich die Zusammenhänge zwischen Repräsentationskompetenz und verschiedenen Faktoren gruppenspezifisch unterscheiden.

Zwar wurden signifikante Zusammenhänge zwischen Repräsentationskompetenz und Vorleistungsindikatoren (Abiturnote:  $r = -0.33$ , Mathematik-Vorleistung:  $r = 0.38$ , Prä-Leistung:  $r = 0.72$ ) nachgewiesen, aber die Stärke der Zusammenhänge zwischen diesen Variablenpaaren unterschied sich nicht zwischen den Untersuchungsgruppen, d.h. die Gruppenzugehörigkeit hatte keinen signifikanten Einfluss auf die genannten Korrelationen. Auch der Lernzuwachs zeigte sich robust gegenüber allen getesteten Einflussparametern. Insbesondere zeigten die Belastungsvariablen und das Selbstkonzept keinen gruppenspezifischen Zusammenhang zum Leistungszuwachs. Entgegen möglichen gegenteiligen Befürchtungen wirkte sich eine Überforderung oder ein geringes Selbstkonzept also nicht stärker auf den Leistungszuwachs aus, wenn videobasierte Aufgaben statt traditionelle Aufgaben bearbeitet wurden. Auch der Einfluss der Aufgaben-Bearbeitungszeit auf die Lernwirkung unterschied sich gemäß den Ergebnissen nicht zwischen den Gruppen.

Bei Untersuchung des Einflusses des Geschlechts auf das Ergebnis der Wirksamkeitsuntersuchung wurde zunächst festgestellt, dass Männer nach der Intervention auf Stichprobenebene bessere Leistungen als Frauen aufwiesen. In diesem Zusammenhang ist es nennenswert, dass Frauen zwar geringfügig bessere Vorleistungen als Männer besaßen (Tab. A-13), dies aber keinen Niederschlag in besseren KiRC-Ergebnissen fand (weder zum Prä- noch zum Postzeitpunkt). Die Problematik fehlschlagender Transfers schulischer zu akademischen Leistungen bei Frauen wurde in der fachdidaktischen Literatur zwar schon dokumentiert (Adams et al., 2006), aber noch nicht hinreichend diskutiert. Die vorliegenden Daten replizieren dieses Ergebnis, lassen aber keinen Rückschluss auf die Ursache des genannten Sachverhalts zu.

Nach der Intervention bestand in der Interventionsgruppe kein Leistungsunterschied zwischen Männer und Frauen, d.h. der „Gender Gap“ konnte durch videobasierte Aufgaben geschlossen werden, während er in der Kontrollgruppe weiterhin vorhanden blieb. Dieser Interaktionseffekt ist an der Schwelle zur Signifikanz, sollte aber angesichts der geringen Effektstärke ( $d = 0.26$ ) sowie der geringen Stichprobengröße (insgesamt 20 Frauen) nicht überbewertet werden. Dennoch geben die Daten Anlass zur Vermutung, dass videobasierte Aufgaben wegen der höheren Authentizität und dem unmittelbaren Anwendungsbezug für Frauen ansprechender sind als traditionelle Aufgaben, die ein abstrakteres Vorstellungsvermögen verlangen. Auch wenn die langjährige Forschung zu geschlechtsspezifischen Unterschieden bzgl. der allgemeinen Intelligenz (g-Faktor) kontroverse Ergebnisse hervorbrachte, gibt es Hinweise darauf, dass Frauen (im Allgemeinen) bei abstrakten Aufgabenstellungen größere

Schwierigkeiten als Männer haben (Lynn, 1994), was o.g. These unterstützt. Es ist allerdings nicht auszuschließen, dass ein Teil des dokumentierten Gender-Effekts auf einen höheren Interessensanstieg bei Frauen zurückzuführen ist (siehe Diskussion zur Motivation, Kap. 12.1.3).

## 12.1.2 Konzeptverständnis (Kinematik)

### Haupteffekt

Die in Kap. 11.2.2 berichteten Ergebnisse zeigen, dass videobasierte Aufgaben im betrachteten Interventionszeitraum zu einem signifikant größeren Konzeptverständnis führten als traditionelle Aufgaben (kleiner Effekt,  $d = 0.34$ ). Die Nullhypothese ist deshalb zu Gunsten der Alternativhypothese 1.2 abzulehnen. Die normierten Lernzuwächse beider Gruppen sind positiv und fallen mit  $g_{H,KG} = 0.19$  und  $g_{H,IG} = 0.40$  in den Ergebnisbereich traditioneller bzw. innovativer Lehrmethoden (Coletta et al., 2007). Zu Beginn der Intervention liegen die Mittelwerte beider Gruppen noch knapp unter der 60%-Marke ( $P_{KG} = 0.54$ ,  $SD = 0.30$ ;  $P_{IG} = 0.53$ ,  $SD = 0.29$ ), die von den Konstrukteuren des FCI als „Eintrittsschwelle“ zur Newtonschen Prinzipienwelt verstanden wird (Hestenes & Halloun, 1995). Sowohl traditionelle Aufgaben als auch videobasierte Aufgaben führen zur Überschreitung dieser Marke ( $P_{KG} = 0.63$ ,  $SD = 0.23$ ;  $P_{IG} = 0.72$ ,  $SD = 0.26$ ).<sup>2</sup> Wie im Streudiagramm (Abb. 28) und an den angegebenen Standardabweichungen ersichtlich wird, besteht eine große Variabilität der Messwerte, die auf die Kürze der Skala (5 Items) und auf deren geringe Reliabilität zurückzuführen ist ( $\alpha_c(t_0) = 0.51$  in Kern-Stichprobe, vgl. Tab. 33).

Inhaltlich erforderten die Aufgaben dieses Kurztests zu Trajektorien die korrekte Anwendung des Superpositionsprinzips bei überlagerten Bewegungen in verschiedenen Alltagssituationen. Wie oben bereits diskutiert wurde, führten videobasierte Aufgaben zu einer besseren Interpretationsfähigkeit bildhaft repräsentierter physikalischer Szenarien. Dies wirkt sich ebenfalls positiv auf die Beantwortung der Konzept-Items aus, denn wie bei der Testkonstruktion dargelegt wurde, liegen alle Items dieses Kurztests im bildhaften Format vor (vgl. Kap. 5.2). Das bessere Ergebnis im Konzepttest ist aber nicht ausschließlich auf die gesteigerte Repräsentationskompetenz zurückzuführen: Statistisch ist der Zusammenhang zwischen den Lernzuwächsen (Konzeptverständnis und Repräsentationskompetenz) insignifikant ( $r = 0.06$ ) und inhaltlich verlangen die Konzept-Items außerdem eine Transferleistung der physikalischen Konzepte auf Alltagssituationen sowie formale Kenntnis des zweiten Newtonschen Gesetzes. Im Rahmen videobasierter Aufgaben wurde durch das (vorgegebene oder selbst erstellte) Realexperiment stets ein unmittelbarer Anwendungsbezug hergestellt (vgl. Kontextorientierung, Kap. 3.1.2), wodurch die Übertragbarkeit physikalischer Prinzipien auf reale Situationen eingeübt wurde. Die Förderung des konzeptionellen Verständnisses kann nach obiger Argumentation auf eine erhöhte Verarbeitungstiefe und auf den unmittelbaren Anwendungsbezug physikalischer Prinzipien beim Lösen videobasierter Aufgaben zurückgeführt werden.

Ferner wurde durch eine Distraktorenanalyse ausgewählter Items festgestellt, dass in der Interventionsgruppe eine Fehlerkorrektur betreffend des zweiten Newtonschen Axioms stattfand: Im Post-Test wurde eine konstante Kraft häufiger in Verbindung mit einer beschleunigten Bewegung gebracht als im Prä-Test. Auch wenn die Kinematik nur zu einem geringen Anteil expliziter Bestandteil der vierwöchigen Interventionsphase war, spielten die Zusammenhänge kinematischer Größen bei jeder videobasierten Aufgabe eine implizite Rolle. Insbesondere wurden durch Beschleunigungsmessungen explizite Bezüge zu den wirkenden Kräften hergestellt, wie etwa bei den Aufgaben 2.1 (b), 3.2 (b) und 5.1 (b); vgl. Instruktionsmaterial (ab S. 220).

<sup>2</sup>Dabei sei angemerkt, dass der hier verwendete Kurztest nur eine Teilmenge der FCI-Items abbildet, auf die sich Hestenes & Halloun beziehen.

### Robustheit des Ergebnisses gegenüber verschiedenen Einflussfaktoren

Es konnten keine gruppendifferenten Einflüsse der betrachteten Faktoren auf das konzeptionelle Verständnis bzw. den Lernzuwachs nachgewiesen werden. Auch hing der Lernzuwachs nicht vom Selbstkonzept, von der Aufgabenschwierigkeit oder von der Beanspruchung im Studium, wohl aber von der Bearbeitungszeit der Aufgaben ab. Je mehr Zeit zur Bearbeitung der Aufgaben aufgewendet wurde, desto höher war der erzielte Lernerfolg ( $r = 0.33$ ,  $p < 0.001$ ). Diese Relation gilt unabhängig von der Gruppenzugehörigkeit und ist plausibel (vgl. Berliner, 1990).

Ähnlich zur Repräsentationskompetenz zeichnete sich ferner die Tendenz eines Gender-Effekts ab. Auf Stichprobenebene wiesen Frauen vor und nach der Intervention zwar geringere Leistungen als Männer auf, in der Interventionsgruppe konnte dieser Unterschied allerdings beseitigt werden. Dies verstärkt die oben geäußerte Vermutung, dass videobasierte Aufgaben besonders bei Frauen eine größere Lernwirkung haben als traditionelle Aufgaben. Es sei darauf hingewiesen, dass sich das Item Kon2 bzgl. der Gender-Frage als problematisch herausstellte. Erstens verstieß dieses Item gegen die Rasch-Kriterien (Stichprobenhomogenität nach Teilung des Datensatzes gemäß Geschlecht; siehe Kap. 7.3), zweitens zeigten Frauen bei diesem Item auf Stichprobenebene einen deutlich niedrigeren Lernzuwachs als Männer (siehe Kap. 11.5.1) und drittens führte die Videoanalyse bei diesem Item nicht den gewünschten Lernerfolg herbei. Um die Beobachtungen auf einzelne Elemente der Videoanalyse zurückzuführen, sind allerdings weitere Untersuchungen bzgl. der Gender-Problematik notwendig, z.B. qualitative Interviews mit Lernenden weiblichen Geschlechts nach dem Bearbeiten der unterschiedlichen Aufgabentypen.

### 12.1.3 Motivation

Die in Kap. 9 formulierte Hypothese 1.3 bezüglich des Gesamtkonstrukts der Motivation kann nicht bestätigt werden, denn dazu müssten alle Teilkomponenten einen Gruppenunterschied zu Gunsten der IG aufweisen. Wohl aber konnte die Wirksamkeit videobasierter Aufgaben theoriekonform für die verschiedenen Facetten der Authentizität (Realitätsbezug, Professionsbezug, Experimentbezug) nachgewiesen werden.

Im Folgenden werden die in Kap. 11.2.3 berichteten Ergebnisse für die verschiedenen Teilkomponenten der Motivation einzeln diskutiert:

#### Realitätsbezug

Die Interventionsgruppe assoziierte einen höheren Realitätsbezug mit den Übungsaufgaben als die Kontrollgruppe (kleiner Effekt  $d = 0.37$ ). Da die Inhalte und die physikalischen Kontexte in korrespondierenden traditionellen und videobasierten Aufgabenpaaren identisch waren<sup>3</sup>, kann dieses Ergebnis (i) auf die höhere Authentizität des bildbasierten Mediums und (ii) auf das Durchführen von Realexperimenten zurückgeführt werden:

- (i) Es ist theoriekonform, dass die Visualisierung physikalischer Prozesse zu einem größeren Realitätsbezug führt als eine rein gedankliche Auseinandersetzung mit verbal beschriebenen Abläufen, vgl. Kap. 3.1.1 und 3.1.2.
- (ii) Die Interventionsgruppe führte selbst Experimente durch, wodurch der Bezug der Aufgaben zu realen Phänomenen offensichtlich wurde.

Aufgrund der Mischung mobiler und non-mobiler Videoanalyse-Aufgaben im Studiendesign kann nicht eindeutig entschieden werden, ob der Effekt auf einen spezifischen Aufgabentyp zurückgeführt werden kann. Der diagnostizierte Gruppenunterschied ist zu groß, als dass er auf einen Neuigkeitseffekt der Aufgaben zurückgeführt werden kann.<sup>4</sup>

<sup>3</sup>z.B. die Aufgaben zum Bungee-Springer (A 3.2) oder zum rollenden Zylinder (A 4.1) (ab S. 220)

<sup>4</sup>Für Neuigkeitseffekte (Hawthorne-Effekte) gilt typischerweise  $d < 0.3$  (Adair, Sharpe & Huynh, 1989).

### Akademisches Selbstkonzept

Bezüglich des akademischen Selbstkonzepts wurden keine signifikanten Gruppenunterschiede festgestellt. In Kap. 3.1.4 wurde die Vermutung begründet, dass die durch das Theorie-Experiment-Wechselspiel induzierten Reflexionsmechanismen zu einer Steigerung des Selbstkonzepts führen. Zwar weisen die deskriptiven Daten und die Analyseergebnisse eine positive Tendenz zur Ablehnung der Nullhypothese auf, die Wirkung blieb jedoch hinter den Erwartungen zurück. Aufgrund der Datenlage kann nicht ausgeschlossen werden, dass die intendierten Kontrollmechanismen von den Lernenden nicht wahrgenommen wurden und deshalb der Effekt ausblieb. Qualitative Befragungen könnten diese Fragestellung klären.

### Interesse / Engagement

Auch bezüglich des Interesses / Engagements konnten keine Gruppenunterschiede festgestellt werden. Das Interesse war schon zu Semesterbeginn im Durchschnitt sehr hoch ausgeprägt (Stichprobenwert  $M = 0.75$ ,  $SD = 0.11$ ), was auf eine Interessenselektion zurückgeführt werden kann: Personen, die sich für ein Physik-Studium entscheiden, sind in der Regel hoch intrinsisch motiviert, zeigen eine hohe Identifikation mit dem gewählten Fach und besuchten schon in der Schule naturwissenschaftliche Leistungskurse (vgl. Albrecht, 2011, S. 19 und Kap. 10.1).

Trotz des nachweislich höheren Realitätsbezugs der videobasierten Aufgaben wurde in der Interventionsgruppe keine Interessenstabilisierung erreicht. Es ist allerdings nicht auszuschließen, dass der anfängliche signifikante Unterschied bzgl. des Interesses ( $M_{IG} = 0.78$ ,  $SD = 0.12$ ,  $M_{KG} = 0.72$ ,  $SD = 0.08$ ) darauf zurückzuführen ist, dass die Interventionsgruppe über die neue Lernmethode informiert war (Einführungsphase, vgl. Kap. 10.2.2). Es ist also möglich, dass ein Teil der interventionsbedingten Wirkung vorweggenommen wurde. Auch wenn die Hypothese zur Interessensteigerung nicht bestätigt werden kann, ist ein interessenförderliches oder (aufgrund des Deckeneffekts) zumindest interessenstabilisierendes Potential videobasierter Aufgaben deshalb nicht gänzlich auszuschließen.

Es wurde festgestellt, dass das Interesse in beiden Gruppen signifikant abnahm ( $d = 0.44$ ). Gründe dafür können in der steigenden Belastung durch das Studium in der ersten Interventionsphase (vgl. Kap. 12.1.5) sowie im hohen fachlichen Anspruch des Lerninhalts liegen. Während zum Prä-Zeitpunkt fast alle Studierende überzeugt waren, alle gestellten Übungsaufgaben zu lösen (Item M27;  $M = 0.94$ ,  $SD = 0.13$ ), fiel die Zustimmung in beiden Gruppen zum ersten Post-Zeitpunkt um etwa 10% ab.

Außerdem wurde ein kleiner Interaktionseffekt bzgl. Geschlecht und Gruppe festgestellt ( $d = 0.30$ ): Frauen der Interventionsgruppe wiesen ein höheres Interesse auf als Frauen der Kontrollgruppe (und zudem als Männer beider Gruppen). Wie oben bereits geäußert, sind die gefundenen Gender-Effekte wegen der kleinen Stichprobe (9 Frauen in der IG) nicht überzubewerten und bedürfen Anschlussuntersuchungen, um Rückschlüsse auf die Ursache der Effekte zu ziehen.

### Disziplinäre Authentizität

Bezüglich der disziplinären Authentizität wurde ein kleiner Effekt zu Gunsten der Interventionsgruppe an der Signifikanzschwelle festgestellt ( $d = 0.25$ ), der unabhängig von Persönlichkeitsmerkmalen Bestand hat. Die Einschätzung darüber, inwiefern das eigene Arbeiten während der Studieneingangsphase an das wissenschaftliche Arbeiten heranreicht, kann zum Prä-Zeitpunkt mangels Erfahrung hauptsächlich als Erwartungshaltung interpretiert werden. Der nachgewiesene Gruppenunterschied beruht auf der Stabilisierung dieser Erwartungshaltung durch die Bearbeitung videobasierter Aufgaben bei gleichzeitiger Minderung in der Kontrollgruppe. Die Unterschiedshypothese fußte auf dem durch die videobasierten Aufgaben intendierten Wechsel zwischen theoretischen und experimentellen Tätigkeiten sowie der Möglichkeit, Experimente zur Erkenntnisgewinnung / zum Lernen zu nutzen. Der kleine Effekt bzgl. des Professionsbezugs wirft mit Blick auf den großen Gruppenunterschied

bzgl. des wahrgenommenen Experimentbezugs ( $d = 1.02$ , siehe unten) die Frage auf, inwiefern Studierende die genannten fachmethodischen Arbeitsweisen als solche auffassten. Das Durchführen von Experimenten und der Umgang mit eigenen Daten sind Tätigkeiten, die offensichtlich als solche erkannt wurden, aber die für Studierende keine zwingende Voraussetzung physikalischer Arbeitsweisen bildeten. Ansonsten hätte der Effekt bzgl. der disziplinären Authentizität deutlicher ausfallen müssen. Das Ergebnis der Wirksamkeitsuntersuchung beinhaltet eine Varianzquelle, die nicht berücksichtigt wurde, nämlich die intraindividuelle Vorstellung über naturwissenschaftliche Arbeitsweisen. So bestand in der Kontrollgruppe nach der Intervention eine 68%-ige Zustimmung, dass das Bearbeiten ausschließlich traditioneller Aufgaben physikalische Arbeitsmethoden hinreichend abbildet, was aus fachlicher Perspektive fragwürdig erscheint.

### Experimentbezug

Wie bereits im Einleitungstext zu Kap. 12.1 diskutiert wurde, liegt ein großer Effekt bzgl. des Experimentbezugs vor ( $d = 1.02$ ), der eine gezielte Anreicherung der Aufgaben mit experimentellen Arbeitsweisen bedeutet. Es wurden zwei Interaktionseffekte nachgewiesen: Der wahrgenommene Experimentbezug ist in der KG umso geringer, je besser die Physik-Vorleistung (bzw. je besser die Abiturnote) der Studierenden ist. In der IG ist der Zusammenhang genau umgekehrt, aber von geringerer Bedeutung (vgl. Steigungen in Abb. 42). Der Gruppenunterschied bzgl. des wahrgenommenen Experimentbezugs basiert also vor allem auf den vorleistungsstarken Studierenden, die das Instruktionsmaterial genauer differenzieren können. Da ein höherer Experimentbezug videobasierter Aufgaben per Konstruktion gegeben war, lässt sich das Ergebnis in anderen Worten so formulieren, dass die Adäquatheit von Einschätzungen mit wachsender Fähigkeit steigt. Diese Interpretation zeigt eine Parallele zum *Dunning-Kruger-Effekt* („Die metakognitiven Fähigkeiten von Studierenden steigen mit wachsender Leistung“, bedeutungsgemäß nach Kruger & Dunning, 1999), was die Plausibilität des Befunds verstärkt.

### Autonomie

Die empfundene Autonomie nahm in beiden Gruppen ohne Zwischengruppen-Effekte signifikant ab. Zwischen empfundener Autonomie und Experimentbezug der Aufgaben bestand auf Stichprobenebene zwar theoriekonform ein positiver Zusammenhang ( $r = 0.31$ ,  $p < 0.01$ ), dennoch gelang es nicht, den Studierenden eine größere Entscheidungsfreiheit bzgl. der Aufgabenbearbeitung durch videobasierte Aufgaben bewusst zu machen. Vermutlich ist die Ursache in den engen zeitlichen Rahmenbedingungen der Aufgabenbearbeitung begründet, die kaum Spielraum lässt, experimentellen Variationen nachzugehen. Diese Vermutung wird durch eine substantielle negative Korrelation zwischen Autonomie und Beanspruchung ( $r = -0.23$ ,  $p < 0.05$ ) und zwischen Autonomie und Aufgaben-Bearbeitungszeit gestützt ( $r = -0.33$ ,  $p < 0.01$ ). Je belasteter die Studierenden waren, desto weniger Entscheidungsfreiheit nahmen sie wahr. Der große Belastungsanstieg während der ersten Interventionswochen (vgl. Kap. 12.1.5) erklärt die Abnahme der Autonomie auf Stichprobenebene.

Mobile Videoanalyse-Aufgaben bieten zwar prinzipiell die Möglichkeit, Aufgabenstellungen zu öffnen und verschiedene experimentelle Lösungsansätze zuzulassen. Aber eben wegen den strikten zeitlichen Rahmenbedingungen und zur Gewährleistung der Vergleichbarkeit zu korrespondierenden T-Aufgaben wurde davon abgesehen, dieses Potential bei Gestaltung des Instruktionsmaterials vollständig auszuschöpfen.

### 12.1.4 Neugierde

Gemäß den in Kap. 11.2.4 berichteten Ergebnissen konnte eine Förderung der Zustandskomponente epistemischer Neugierde durch videobasierte Aufgaben gegenüber traditionellen Aufgaben nachgewiesen werden. Das Ergebnis beruht zwar auf einem reinen Gruppenvergleich

ohne Messwiederholung (da eine Erhebung der spezifischen Neugierde-Komponente vor der Intervention unsachgemäß wäre), aber es konnte ausgeschlossen werden, dass sich die Gruppen hinsichtlich der zeitstabilen allgemeinen Neugierde-Komponente (Persönlichkeitseigenschaft) unterscheiden. Der Effekt ist praktisch relevant ( $d = 0.39$ ) und besteht unabhängig vom Geschlecht der Lernenden. Allerdings wurde ein unterschiedlicher Zusammenhang zwischen geweckter Neugierde und Mathematik-Vorleistung zwischen den Gruppen festgestellt. Die ausschließlich theoriebasierten Aufgaben führten in der Kontrollgruppe zu einer umso geringeren Neugierde, je höher die Mathematik-Vorleistung ausfällt. Die geringe Neugierde leistungsstarker Studenten ist vermutlich ein Effekt falscher Anforderungen der Aufgaben, die zu keiner weiteren Vertiefung oder Auseinandersetzung mit den Inhalten führt. Die Feststellung, dass eben jene leistungsstarke Studenten der Kontrollgruppe den Experimentbezug der Aufgaben als sehr gering beurteilen (siehe Diskussion in Kap. 12.1.3 bzw. Abb. 42), unterstreicht die Bedeutsamkeit videobasierter Aufgaben als Möglichkeit zur Vertiefung der Lerninhalte. Dies wird auch dadurch deutlich, dass Neugierde (NZ) und Experimentbezug hoch miteinander korrelieren ( $r = 0.48$ ,  $p < 0.001$ ). In der Interventionsgruppe ist die geweckte Neugierde hingegen unabhängig von der Mathematik-Vorleistung, d.h. videobasierte Aufgaben fördern die weitere Auseinandersetzung mit den Inhalten der Übungsaufgaben weitgehend voraussetzungsfrei. Matthäus-Effekte, die eine höhere Förderung leistungsstarker im Vergleich zu leistungsschwachen Studierenden bedeuten, traten demnach nicht auf.

Die positive Wirkung der Intervention auf die Neugierde unterstreicht im direkten Vergleich mit den ausbleibenden Effekten bzgl. des Interesses die konzeptionelle Verschiedenheit der beiden Variablen NZ und IE, auch wenn sie miteinander hoch korrelieren ( $r = 0.54$ ,  $p < 0.001$ ).

### 12.1.5 Belastungsindikatoren

#### Beanspruchung durch das Studium im Allgemeinen

Die Beanspruchung durch das Studium im Allgemeinen (PSQ-Skala) stieg von einem anfänglich moderaten Niveau signifikant und mit mittelgroßer bis großer Effektstärke an (Stichprobenwert  $d = 0.70$ ), wobei keine Gruppenunterschiede festgestellt wurden. Der Anstieg der Beanspruchung zeigte sich unabhängig von den Vorleistungen und ist auf den Anspruch des Studiums zurückzuführen, der eine hohe Forderung an zeitliche und kognitive Ressourcen der Studierenden stellt. So gaben 77% der Studierenden nach der Intervention an, zu viele Dinge für ihr Studium erledigen zu müssen und 74% fühlten sich durch Abgabetermine der Aufgaben unter Druck gesetzt. Hinzu kommt die Neuorientierung als Universitätsstudent in den ersten Semesterwochen betreffend Änderungen der Lebensbedingungen (z.B. Wohnortwechsel), sozialer Integration in den universitären Lebensraum sowie erhöhter Verantwortung (z.B. Zeit-Management, Studienfinanzierung).

Das Ausbleiben von Gruppenunterschieden bedeutet insbesondere, dass der durch das Medium Videoanalyse generierte Aufwand zur Bearbeitung der Aufgaben keine Auswirkungen auf die Beanspruchung hatte. Vermutlich schlägt sich dieser Aufwand deshalb nicht in den Daten nieder, weil er gemessen an der Gesamtbelastung gering ausfällt. Schon ältere Studien zur Videoanalyse haben gezeigt, dass die Bedienung von Videoanalyse-Programmen auch technisch unerfahrenen Lernenden ohne größeren Aufwand gelingt (vgl. Kap. 3.2). Durch die einheitliche Einführung in die Videoanalyse während der Einführungsphase wurde sichergestellt, dass die Voraussetzungen zum Bearbeiten videobasierter Aufgaben gegeben waren, sodass die durch das Medium induzierte Beanspruchung möglichst gering gehalten wurde.

Es wurde ein sehr großer Anstieg der Beanspruchung bei anfänglich gering belasteten Studierenden der Kontrollgruppe festgestellt, der vermutlich auf eine Unterschätzung der Anforderungen zurückzuführen ist. Gerade der mathematische Anspruch der experimentalphysikalischen Übungsaufgaben, der bei videobasierten Aufgaben reduziert ist, wird von Studienanfängern erfahrungsgemäß oftmals unterschätzt (Kasten, 2011).

### Aufgabenschwierigkeit

Ähnlich wie die Beanspruchung stieg auch die wahrgenommene Aufgabenschwierigkeit auf Stichprobenebene signifikant und mit mittelgroßer bis großer Effektstärke im ersten Interventionszeitraum ohne Gruppenunterschiede an ( $d = 0.70$ ). Die Beurteilung der Aufgabenschwierigkeit zum Prä-Zeitpunkt beruhte auf Einschätzung des ersten Übungsblattes (mit hauptsächlich Wiederholungsaufgaben) und wurde vermutlich durch den Eindruck von Physikaufgaben aus der Schulzeit beeinflusst. Die vorlesungsbegleitenden Übungsaufgaben sind mit typischen Schulaufgaben weder formal, noch inhaltlich vergleichbar, denn sie erfordern eine komplexere, zeitintensivere Auseinandersetzung mit den dargestellten Problemen. Der hohe fachliche Anspruch führte — wie oben beschrieben — einerseits zum Anstieg der Beanspruchung, und andererseits auch zum Anstieg der empfundenen Aufgabenschwierigkeit. In der Tat bestand ein signifikanter Zusammenhang zwischen der Zunahme von Aufgabenschwierigkeit und Beanspruchung ( $r = 0.26$ ,  $p < 0.05$ ), obwohl sich beide Variablen als psychometrisch klar voneinander trennbare Faktoren darstellten (vgl. Kap. 6.4.1).

Das Ausbleiben von Gruppenunterschieden ist auf das Bemühen zurückzuführen, die Instruktionmaterialien unter Prämisse des fairen Vergleichs zu konstruieren. So thematisierten videobasierte und traditionelle Aufgaben die gleichen physikalischen Prinzipien in einem identischen Anwendungskontext (sprich, die physikalische Situation war identisch). Es wurde bei Konstruktion des Materials darauf geachtet, dass die Schwierigkeit aus den physikalischen Inhalten (intrinsic Cognitive Load), und nicht aus deren Darstellung erwächst (extraneous Cognitive Load).

### Bearbeitungszeit der Aufgaben

Die wöchentliche Bearbeitungszeit der Aufgaben wurde von den Studierenden nach der Intervention retrospektiv durch Angabe eines Zahlenwertes geschätzt. Es wurde zwischen beiden Gruppen kein signifikanter Unterschied festgestellt. Entgegen anders anzunehmender Befürchtungen führten videobasierte Aufgaben demnach zu keinem größeren Zeitaufwand.

Die Bearbeitungszeit der Aufgaben hängt erwartungskonform mit der empfundenen Aufgabenschwierigkeit ( $r = 0.29$ ,  $p < 0.05$ ) und mit dem Lernzuwachs ( $r = 0.33$ ,  $p < 0.01$ ) zusammen, was die Validität der Selbsteinschätzung unterstreicht. Die Bearbeitungszeit der Übungsaufgaben variierte zwischen einzelnen Studenten stark (2–19 h pro Übungszettel) und fiel mit dem Stichprobenmittelwert von 8.80 h ( $SD = 3.34$  h) hoch aus; die Einführungsveranstaltung generiert noch einmal fast das Doppelte an Arbeitszeit für Übungsaufgaben wie die Präsenzzeit (4 SWS Vorlesung und 2 SWS Übung pro Woche). Da in dieser Zeit die eigentliche Informationsverarbeitung stattfindet, kann zu Recht gefolgert werden, dass die Übungen eine entscheidende Ressource zum Verständnis der präsentierten physikalischen Inhalte darstellen. Wie oben berichtet, ist die aufgewendete Übungszeit ein wichtiger Prädiktor für den Lernerfolg.

## 12.2 Zweiter Interventionszeitraum

Infolge der systematischen Unterscheidung der Gruppen zu Beginn der zweiten Interventionsphase ist eine spekulationsfreie Diskussion der Ergebnisse der zweiten Phase nicht möglich. Sie ist auch insofern von geringer praktischer Bedeutung, da ein derartiger Aufgabenwechsel den Rahmenbedingungen der Untersuchung geschuldet ist und außerhalb von Studien keinen Bestand hätte. Deshalb beschränkt sich die Diskussion der zweiten Interventionsphase auf ausgewählte Ergebnisse.

Auch in der zweiten Interventionsphase (zwischen den Testzeitpunkten  $t_1$  und  $t_2$ ) war der Manipulation Check erfolgreich: Die Gruppe, die zuvor videobasierte Aufgaben bearbeitete, assoziiert mit den traditionellen Aufgaben einen signifikant geringeren Experimentbezug. Dagegen stieg bei der Gruppe, die zunächst mit T-Aufgaben arbeitete, der Experimentbezug in der zweiten Interventionsphase an. Dieser gruppendifferente Zeitverlauf besitzt

eine große Effektstärke ( $d = 1.57$ ). Dieses Ergebnis zeigt zum einen nachdrücklich, dass mit videobasierten Aufgaben ein höherer Experimentbezug — unabhängig von den konkreten physikalischen Inhalten der Aufgaben — erreicht wurde. Zum anderen zeigen die Zeitverläufe der beiden Gruppen, dass der Aufgabenwechsel statistisch nachweisbar durch Anstieg bzw. Abfall des Experimentbezugs wahrgenommen wird. Die Ursache dafür, dass der Effekt größer im Vergleich zur ersten Interventionsphase ausfällt, liegt darin, dass der Experimentbezug nach einigen Interventionswochen besser differenziert werden kann als zu Beginn der Intervention. So prägten die Aufgaben in den ersten Interventionswochen einen Eindruck über den Experimentbezug, der im Kontrast zu dem Eindruck der Aufgaben im zweiten Zeitraum stand. Die Studierenden erfuhren einen Aufgabenwechsel und antizipierten die Änderung des Experimentbezugs entsprechend. Dagegen konnte der Experimentbezug in der ersten Phase nur in Bezug zu einer Erwartungshaltung oder Eindrücken von Schulaufgaben gesetzt werden.

### 12.2.1 Leistung (Konzeptverständnis)

Auch im zweiten Interventionszeitraum gelang es der Interventionsgruppe, das Konzeptverständnis zu Trajektorien signifikant zu verbessern (normierter Lernzuwachs  $g_H = 0.27$ ), wenn auch mit einem geringeren Lernzuwachs als der Interventionsgruppe im ersten Zeitraum. Der geringere Lernzuwachs kann hauptsächlich auf zwei Gründe zurückgeführt werden: Erstens spielte die Kinematik (und insbesondere das Superpositionsprinzip) in dieser Phase keine explizite Rolle in den Übungsaufgaben mehr. Da sie aber bei jeder Videoanalyse implizit angewendet wurde (Vektordekomposition, Bezug zwischen kinematischen Größen, vgl. Diskussion in Kap. 12.1.1 und 12.1.2), konnte sich das Konzeptverständnis der Interventionsgruppe verbessern. Die Kontrollgruppe war in der zweiten Phase dagegen kaum zu nachträglichen Steigerungen fähig (normierter Lernzuwachs  $g_H = 0.07$ ). Zweitens startete die Interventionsgruppe von einem höheren Ausgangszustand, sodass der maximal mögliche Lernerfolg unter Berücksichtigung von Deckeneffekten niedriger lag. Am Ende der Studie erreichten beide Gruppen vergleichbare Leistungen, unabhängig davon, in welcher Reihenfolge sie videobasierte Aufgaben bearbeiteten. Die Gruppe, die mit videobasierten Aufgaben begonnen hat, konnte schon nach vier Wochen von dem hohen Konzeptverständnis profitieren, sodass anzunehmen ist, dass sie kognitive Kapazitäten für den Ausbau anderer Kompetenzen nutzen konnte.

### 12.2.2 Motivation und Neugierde

Wie in der ersten Interventionsphase zeigten sich bzgl. des Interesses und der Autonomie keine Gruppenunterschiede. Im Gegensatz zur ersten Phase fielen weder Interesse noch Autonomie weiter ab, was durch die Stabilisierung der Belastungsindikatoren erklärt werden kann (siehe Kap. 12.2.3). Die Messung des Realitätsbezugs ergab einen Gruppenunterschied mit einem Effekt in ähnlicher Größenordnung wie in der ersten Phase ( $d = 0.41$ ) und auch bezüglich des akademischen Selbstkonzepts konnte das Ergebnis der ersten Phase repliziert werden: Es zeigte sich ein Trend zur Steigerung des Selbstkonzepts zu Gunsten der Interventionsgruppe, der allerdings nicht signifikant wurde. Die disziplinäre Authentizität blieb in der zweiten Phase weitgehend stabil und weist im zeitlichen Verlauf keine Gruppenunterschiede auf. Wie in Kap. 12.1.3 bereits diskutiert wurde, schlägt sich der große Effekt bzgl. des Experimentbezugs nicht in dieser Variable nieder, was deren Unterschiedlichkeit unterstreicht.

Die Ausprägung der epistemischen Neugierde blieb während der zweiten Interventionsphase ebenfalls stabil. Mögliche Gründe dafür, warum das Ergebnis aus der ersten Phase nicht repliziert werden kann, sind vielfältig. Es wäre möglich, dass die unterschiedlichen Aufgaben in beiden Phasen ein unterschiedliches Potential zur Steigerung der Neugierde besaßen oder dass die erste Interventionsphase zu einer Stabilisierung der Neugierde führte, die unabhängig vom Aufgabentyp während der zweiten Phase Bestand hatte. Das Ergebnis kann — wie zu

Beginn dieses Kapitels erwähnt — infolge des Studiendesigns nicht spekulationsfrei diskutiert werden.

### 12.2.3 Belastungsindikatoren

Während es in der ersten Interventionsphase zu einem starken Anstieg der Belastungsindikatoren „Beanspruchung“ und „Aufgabenschwierigkeit“ in beiden Gruppen ohne Gruppenunterschiede kam, stabilisierten sich beiden Variablen in der zweiten Phase auf Stichprobenebene. Es zeichneten sich während der zweiten Interventionsphase jedoch Gruppenunterschiede im Zeitverlauf ab: Die Gruppe, die von traditionellen Aufgaben zu videobasierten Aufgaben wechselte (KG-IG-Gruppe), empfand eine geringere Beanspruchung im Allgemeinen ( $d = 0.44$ ) und beurteilte die Übungsaufgaben einfacher als die IG-KG-Gruppe ( $d = 0.55$ ). Die Bearbeitungszeit der Aufgaben verringerte sich zwar im Vergleich zur ersten Phase um durchschnittlich 1 h pro Woche, aber es traten keine Gruppenunterschiede auf. Es ist aufgrund des Studiendesigns nicht zu entscheiden, ob die festgestellten Gruppenunterschiede bzgl. Aufgabenschwierigkeit und Beanspruchung daraus resultieren, dass die videobasierten Aufgaben und die korrespondierenden traditionellen Aufgaben in der zweiten Phase weniger vergleichbar waren als in der ersten Phase oder ob das Ergebnis eine intraindividuelle Antizipation auf den Aufgabenwechsel darstellt. Beispielsweise könnten Studierende es als Erleichterung empfunden haben, von rein traditionellen Aufgaben zu anschaulicheren experimentellen Aufgabenstellungen zu wechseln. Umgekehrt könnte der Verlust von Anschauung und experimentellen Anteilen beim entgegengesetzten Wechsel zu einer größeren Belastung geführt haben. Auch wenn diese Argumentation zu einer neuen interessanten Hypothese führt, die in Folgestudien geprüft werden kann, soll an dieser Stelle die Komplementarität beider Aufgabentypen betont werden (vgl. Kap. 4.3). In einem praxisnahen Anwendungsszenario kommen stets beide Aufgabentypen vor (T-Aufgaben und V/mV-Aufgaben), sodass sich die Frage nach einer Antizipation nur theoretisch stellt. Nichtsdestotrotz kann zusammenfassend festgehalten werden, dass videobasierte Aufgaben nicht zu einer erhöhten Belastung führen und nicht schwieriger wahrgenommen werden als korrespondierende T-Aufgaben. Falls das Ergebnis kein Artefakt aus dem Studiendesign sein sollte, besitzen videobasierte Aufgaben sogar das Potential, bei gleichen physikalischen Inhalten eine geringere Belastung zu induzieren. Dies widerspräche zwar dem Bemühen zur Konstruktion belastungsähnlicher Aufgabenpaare, ist aber gemäß der CATLM durchaus theoriekonform (vgl. Kap. 3.1.3).

## 12.3 Klausurergebnis

Die Klausur am Semesterende umfasste fünf Aufgaben zur Experimentalphysik 1 (Mechanik), von denen eine Aufgabe zur Kinematik des Massenpunktes inhaltlich eindeutig der ersten Interventionsphase zugeordnet werden konnte. In dieser Aufgabe erreichte die Interventionsgruppe des ersten Zeitraums eine höhere Punktzahl als die Kontrollgruppe ( $d = 0.45$ ). Wie gezeigt wurde, führte das Bearbeiten von videobasierten Aufgaben zu höherem konzeptionellen Verständnis über zusammengesetzte Bewegungen und zu einer höheren Repräsentationskompetenz im Themenbereich Kinematik. Da die Gesamtstichprobe zum Ende des Semesters vergleichbare Leistungen im Konzepttest aufwies, ist die höhere Leistungsfähigkeit bei der Klausuraufgabe primär auf den Effekt bzgl. der Repräsentationskompetenz zurückzuführen. Dafür spricht zum einen, dass die erreichte Punktzahl bei dieser Aufgabe hoch mit dem KiRC-Test korreliert ( $r = 0.42$ ), und dass Problemlösefähigkeit eher mit Repräsentationskompetenz in Verbindung gebracht wird als mit konzeptionellem Verständnis (Savinainen et al., 2013; Crouch & Mazur, 2001). Obwohl dieses Ergebnis für die Wirksamkeit videobasierter Aufgaben spricht, ist die Klausur kein guter Indikator zur Quantifizierung des Interventionserfolgs, da zwischen dem Interventionszeitraum und dem Klausurzeitpunkt üblicherweise eine intensive Vorbereitung ohne externe Kontrollmechanismen stattfindet. Die Folgerung, dass Studierende durch videobasierte Aufgaben zu

bestimmten Themen besser in zugehörigen Prüfungsaufgaben abschneiden, ist deshalb nicht zwingend.

Neben dem Gruppenunterschied bei dieser Klausuraufgabe wurde ein mittelgroßer Effekt bzgl. des Klausur-Gesamtergebnisses (Experimentalphysik-Aufgaben und MGdP-Aufgaben) festgestellt ( $d = 0.43$ ). Es ist anzuzweifeln, dass dieser Gruppenunterschied tatsächlich auf die Intervention mit videobasierten Aufgaben zurückzuführen ist. Falls dem so wäre, hätte die Steigerung der Repräsentationskompetenz während der ersten Semesterhälfte einen starken Transfereffekt auf die Bearbeitung theoretischer Aufgabenstellungen bewirkt. Die geringen Unterschiede bzgl. den algebraisch-formal orientierten Aufgaben im KiRC-Instrument sprechen gegen diese Vermutung. Es sind weitere Untersuchungen notwendig, um die Beeinflussung des Verständnisses theoretischer Physik durch experimentell-orientierte Lehransätze zu quantifizieren.

Ein weiteres interessantes Ergebnis betrifft die Zusammenhänge zwischen den Klausur-Noten und den Vorleistungen bzw. den leistungsbezogenen abhängigen Variablen. Die Abiturnote und die Mathematik-Vorleistung stellen sich als starke Prädiktoren sowohl für die EP1-Aufgaben als auch für die MGdP-Aufgaben heraus ( $|r| > 0.43$ ), während die Physik-Vornote in keinem Zusammenhang mit den Klausur-Ergebnissen steht. Gerade im ersten Fachsemester spielen Mathematik-Kenntnisse also eine bedeutende Rolle, wenn es um den Studienerfolg geht. Diese Feststellung wird auch dadurch unterstützt, dass sich frühe Studienaussteiger und Weiterstudierende bzgl. den Mathematik- und den Abiturnoten — nicht bzgl. den Physiknoten — deutlich unterscheiden (vgl. nächsten Abschnitt), und deckt sich sowohl mit den Erfahrungen von Studierenden (Kasten, 2011) als auch mit den Ergebnissen einer Meta-Analyse zur Prädiktion des Studienerfolgs: „Die Abiturnote spiegelt Fähigkeiten und Fertigkeiten wider, die [...] für gute Studiennoten relevant sind, wie z.B. kognitive Leistungsfähigkeit, Lernbereitschaft, Leistungsmotivation, Fleiß und sprachliche Ausdrucksfähigkeit. [...] Die beste Vorhersagemöglichkeit ergibt sich sowohl mit der Durchschnittsnote als auch mit der Mathematiknote für die Studienfächer Mathematik, Natur- und Ingenieurwissenschaften“ (Trapmann et al., 2007, S. 24).

Außerdem ist es theoriekonform, dass das konzeptionelle Verständnis mit dem EP1-Teilergebnis korreliert, nicht aber mit der MGdP.

## 12.4 Ursache für Studienausstieg

Im Rahmen der Voranalyse (Kap. 11.1.1) wurde untersucht, ob sich die Hauptfach-Studierenden, die nur zum ersten Testzeitpunkt  $t_0$  (zwei Wochen nach Semesterbeginn) anwesend waren, systematisch von den Studierenden unterscheiden, die auch zum ersten Post-Testzeitpunkt  $t_1$  (sechs Wochen nach Semesterbeginn) an den Übungen teilnahmen. Die Aussteiger-Gruppe ( $N = 35$ ) wurde als Dropout deklariert, die Weiterstudierenden bildeten die Kern-Stichprobe der Untersuchung (verbundene Stichprobe,  $N = 76$ ). Der weitere Schwund von 10 Studierenden zwischen den Testzeitpunkten  $t_1$  und  $t_2$  wurde nicht untersucht.

Die Ergebnisse zeigten signifikante Unterschiede dieser beiden Studierendengruppen in Bezug auf die Abitur-Gesamtnote ( $d = 0.61$ ), die Mathematik-Vorleistung ( $d = 0.77$ ), das Interesse ( $d = 0.66$ ), das Selbstkonzept ( $d = 0.66$ ), die empfundene Schwierigkeit der ersten beiden Aufgabenblätter ( $d = 0.44$ ) sowie bzgl. der Repräsentationskompetenz ( $d = 0.51$ ). Bei dem Gruppenvergleich wurde der systematische Fehler in Kauf genommen, auch Aussteiger aus der Studie (aufgrund von Fehlzeiten bei der Datenerhebung) als Studienabbrecher zu deklarieren. Außerdem wurde allen Studierenden, die zum ersten Testzeitpunkt  $t_0$  in der Übung anwesend waren, eine ernsthafte Studienabsicht unterstellt (keine sog. Parkstudierende, vgl. Düchs & Ingold, 2015). Aufgrund der berechneten Effektstärken, die im mittleren bis oberen Bereich liegen, sind die Ergebnisse trotz den potentiellen Unzulänglichkeiten in der Gruppenbildung von hoher praktischer Relevanz und lassen eine Interpretation und Einordnung in die Studienabbruch-Forschung zu.

Die schulischen Vorleistungen hinsichtlich der Abitur-Gesamtnote und der Mathematik-

Note werden als valide Prädiktoren eines erfolgreichen Studiums im Allgemeinen (Trapmann et al., 2007) und des Physik-Studiums im Speziellen (Heublein & Wolter, 2011) angesehen. Außerdem ist bekannt, dass sowohl das Interesse als auch das Selbstkonzept eine wichtige Rolle im beruflichen Werdegang und für die Studienfachwahl spielen (Bandura et al., 2001; Filipp, 2006; Krapp, 1989). Zusammengefasst spiegeln die in dieser Studie ermittelten Daten zwischen frühen Studienabbrechern/-Fachwechslern und Weiterstudierenden deshalb eine Selektion nach affektiven und kognitiven Variablen wider, die in der Literatur zur Prognose von Studienerfolg dokumentiert ist. Die Schwundquote von 32% zwischen den ersten beiden Testzeitpunkten und von 13% zwischen den letzten beiden Testzeitpunkten bedeutet erstens, dass sich die Studienabbrecher schon frühzeitig zu einem Fachwechsel/Studienabbruch entscheiden. Dies steht in Einklang mit dem Befund aus Albrecht (2011). Zweitens ist die Schwund-Gesamtquote von ca. 40% vergleichbar mit den Angaben aus den jährlichen Ermittlungen der Konferenz der Fachbereiche Physik (Düchs & Ingold, 2015).

## 12.5 Einschränkungen und Generalisierbarkeit

Um die Frage nach der Übertragbarkeit oder Generalisierbarkeit der Ergebnisse beantworten zu können, müssen die Rahmenbedingungen und Methoden der Studie, die verwendeten Testinstrumente sowie das eingesetzte Instruktionsmaterial diskutiert werden:

- *Auswahl des Aufgabenmaterials.* Die Auswahl der videobasierten Aufgaben aus einem initialen Aufgabenpool erfolgte in Abstimmung mit den Vorlesungsinhalten und demnach praxisorientiert. Da die eingesetzten Leistungstests kaum interventionsspezifische Elemente beinhalten (beispielsweise Items, die sich auf bestimmte Aufgaben der Intervention beziehen), sondern vielmehr übergeordnete Konstrukte thematisieren, ist anzunehmen, dass der erzielte Gruppenunterschied nicht bedeutend von der Auswahl der videobasierten Aufgaben abhängig ist. Falls Aufgaben in zukünftigem Lehreinsatz variiert werden, ist für den Bestand der Lernwirkung sicherzustellen, dass die Aufgaben auch tatsächlich die Förderung repräsentationaler Fähigkeiten anregen und einen Experimentbezug besitzen.
- *Interventionsstärke und Dosis.* Es kann keine Aussage darüber getroffen werden, inwiefern die eingesetzte Dosis videobasierter Aufgaben das Ergebnis der Untersuchung beeinflusste. In der vorliegenden Studie bearbeitete die Interventionsgruppe in einem Zeitraum von vier Wochen acht T-, sechs V- und zwei mV-Aufgaben. Es war nicht Ziel der Studie herauszufinden, wie sich eine Variation des Verhältnisses der Aufgabentypen zueinander oder des Gesamtumfangs der Aufgaben auf das Ergebnis auswirkt (*Dosis-Wirkung-Beziehung*). Als Konsequenz gibt es keine empirischen Anhaltspunkte dafür, ob die verschiedenen videobasierten Aufgabenformaten (V bzw. mV) unterschiedliche Wirkungen erzielten. Die Kombination beider Aufgabenformate geschah im Sinne einer guten Praxis, um einerseits die komplementären Vorteile beider Formate zu nutzen und andererseits das Risiko einer Überforderung der Lernenden durch einen zu hohen Anteil videobasierter Aufgaben gering zu halten. Aus inhaltlicher Sicht werden keine wesentlich unterschiedlichen Lernwirkungen erwartet, wenn auf ein Format gänzlich verzichtet wird, da beide Aufgabentypen sehr ähnliche instruktionale Ziele verfolgen und nur zwei Variationsformen des vorgestellten Nutzungskonzepts des Mediums Videoanalyse darstellen.
- *Lerneffekt-generierende Elemente.* Die gefundenen Lerneffekte begründen sich auf unterschiedlichen Elementen videobasierter Aufgaben: Das Nutzen multipler Repräsentationsformen, die Bezüge zu disziplinärem Arbeiten (eigenständiges Experimentieren), das Theorie-Experiment-Wechselspiel, die Anschaulichkeit des Mediums, usw. Ein Nachweis, welche Elemente zu welcher Wirkung führten, kann in der vorliegenden Studie nicht erbracht werden. In Folgestudien können solche Elemente Konstruktionskriterien

sein, die je nach Aufgabe mit verschiedener Gewichtung bedient und systematisch variiert werden können.

- *Inhalt der Leistungsinstrumente.* Eine Einschränkung der Aussagekraft bezüglich der Lernwirksamkeit besteht insofern, als dass die Instrumente auf den Bereich der linearen Kinematik und Rotationskinematik eingeschränkt wurden. Dies war erforderlich, um weitestgehend inhaltsunabhängige Aussagen über repräsentationale Fähigkeiten zu treffen, lässt aber einen großen Anteil physikalischer Inhalte, die Gegenstand der Intervention waren, außer Acht.

Für die Generalisierbarkeit der Ergebnisse spricht, dass die Begleituntersuchung an ein *realistisches Einsatzszenario* der Aufgaben im universitären Regelbetrieb gekoppelt war (Feldstudie). Die Studienteilnehmer waren alle anwesenden Regelstudenten, sodass die Zielgruppe adäquat abgebildet wurde. Die Gruppeneinteilung geschah zufällig — es wurde lediglich auf vergleichbare Stichprobengrößen in KG und IG geachtet. Unterschiedliche Vorleistungen der Studierenden wurden durch die Analysemethode berücksichtigt und die Betreuung durch den Übungsleiter wurde kontrolliert.

Ein weiteres Argument für die Validität der Ergebnisse ist die gute psychometrische Qualität der Testinstrumente, auf deren Konstruktion ein Schwerpunkt dieser Arbeit lag. Die Extraktion gut interpretierbarer Skalen geschah im Vorfeld der Vergleichsanalyse auf Basis einer Eich-Stichprobe ( $N = 165$ ). Klassische Testanalysen der verwendeten Items zur Neugierde, Belastung und Motivation lieferten zu allen Messzeitpunkten zufriedenstellende Ergebnisse über die psychometrische Qualität der Skalen ( $\alpha_c = 0.70 - 0.92$ ). Zudem erfüllt das KiRC-Instrument mit einer Item-Teilgruppe die Voraussetzungen des Rasch-Modells. Die Ergebnisse bzgl. der Repräsentationskompetenz basieren demnach auf einer Skala mit höchster statistischer Qualität. Der sehr kurze Konzepttest (5 Items,  $\alpha_c = 0.38 - 0.57$ ) und die Klausur sind im Vergleich weniger zuverlässige Leistungsindikatoren.

Schließlich wurde das positive Ergebnis des primären Ziels der Studie — die Steigerung der Repräsentationskompetenz — methodisch abgesichert. Dazu wurde ein Verfahren zur analytischen Behandlung gemischter Testsummen, wie sie im KiRC-Instrument auftreten, entwickelt und angewendet. Auch die Berücksichtigung von Ratekorrekturen und verschiedenen Itemgruppen führten zu der gleichen Schlussfolgerung, nämlich dass die Intervention repräsentationale Fähigkeiten mit einem praxisrelevanten Unterschied zur Kontrollgruppe steigern konnte.



# Fazit und Ausblick

Die Videoanalyse von Bewegungen besitzt eine lange Tradition in der Physikdidaktik und es wurden bisher zahlreiche Anwendungsbeispiele konzipiert. In dieser Arbeit wurde ein instruktionales Nutzungskonzept dieses Physikmediums in Form von Aufgabenstellungen für experimentalphysikalische Lehrveranstaltungen realisiert. Dabei wurde die Videoanalyse genutzt, um vorgegebene Realexperimente zu analysieren (V-Aufgaben) oder Freihandexperimente selbst durchzuführen, mit Tablet-PCs aufzunehmen und auszuwerten (mV-Aufgaben).

## Studienbezogene Ergebnisse und Folgerungen

Als Anreicherung zu traditionellen Aufgaben im instruktionalen Rahmen vorlesungsbegleitender Übungen zur Experimentalphysik 1 (Mechanik) wurde durch ein KG-IG-Rotationsdesign mit Randomisierung der Untersuchungsgruppen nachgewiesen, dass videobasierte Aufgaben im Vergleich zu rein traditionellen Aufgaben eine höhere Lernwirksamkeit bzgl. verschiedenen kognitiven und affektiven Variablen haben (vgl. Überblick über alle Ergebnisse in Tab. 49 auf S. 179). Insbesondere konnte das theoretisch begründete Potential zur Steigerung repräsentationaler Fähigkeiten (Interpretation und Umgang mit Diagrammen, Formeln, Vektoren, Bildern) und des konzeptionellen Verständnisses der Kinematik nachgewiesen werden. Das Ausbleiben von Gruppenunterschieden bzgl. den erhobenen Belastungsindikatoren ist ganz wesentlich, um auszuschließen, dass diese Lernwirksamkeit auf eine erhöhte Bearbeitungszeit der Aufgaben oder auf eine niedrigere Beanspruchung durch das Studium zurückzuführen ist. Außerdem entkräftet das Ergebnis der Belastungsindikatoren die mögliche Befürchtung, dass videobasierte Aufgaben zu negativen Nebenwirkungen wie erhöhte Bearbeitungszeit und höhere kognitive Anstrengung im Vergleich zu traditionellen Aufgaben führen.

Die insgesamt positiven Ergebnisse der Studie sprechen dafür, videobasierte Aufgaben weiter in den Übungen einzusetzen, um die Effektivität des Lernens zu steigern. Ohne den Einschränkungen der Vergleichsstudie unterworfen zu sein, kann das Potential der Aufgaben weiter ausgeschöpft werden. Die in der vorliegenden Studie erzielten Effekte kamen unter konservativen Bedingungen zustande, sodass größere Lerneffekte erwartet werden, wenn die Aufgabenstellungen gezielt optimiert werden. Beispielsweise kann (1) ein stärkerer Fokus auf einzelne Repräsentationen und deren Dynamisierung gelegt, (2) noch attraktivere Kontexte eingebunden oder (3) der Anteil mobiler Videoanalysen zur Förderung selbstständigen Experimentierens erhöht werden. Die in der Studie vorhandene Notwendigkeit, korrespondierende traditionelle Aufgaben zu stellen, entfällt zukünftig. Wie in Kap. 10.2.2 diskutiert wurde, könnte außerdem (4) die gemeinsame Einführungsphase zu Beginn der Studie für beide Gruppen (KG und IG) einen Teil der Lerneffekte vorweggenommen haben, sodass die Untersuchung konsequenterweise weniger sensitiv gegenüber unterschiedlichen Lernzuwachsen war. Die erzielten positiven Ergebnisse sind damit als konservative Abschätzung zu betrachten, die ohne Einführungsphase vermutlich größer ausfallen würden. Um das Potential von mobilen Endgeräten (Tablet-PC und Smartphone) weiter auszunutzen, könnten neben mobilen Videoanalysen auch Experimente mit dem Beschleunigungssensor in Aufgabenstellungen aufgegriffen werden. Dadurch wird das Spektrum experimenteller Möglichkeiten im Rahmen von klassischen Übungen erweitert.

Die Einbindung videobasierter Aufgaben in bestehende Lernszenarien erfolgt ohne

Veränderungen der Kursstruktur. Dadurch und durch die Verfügbarkeit der Aufgaben und Experimentalvideos wird eine hohe Implementationswahrscheinlichkeit an anderen Physik-Fachbereichen erreicht. In der Tat wird der Einsatz von Varianten videobasierter Aufgaben bereits an zwei weiteren Hochschulen geplant (ETH Zürich unter Leitung von Dr. Clemens Wagner und Goethe-Universität Frankfurt am Main unter Leitung von Prof. Dr. Thomas Wilhelm). Auch an der Technischen Universität Kaiserslautern wird der Einsatz videobasierter Aufgaben in den Übungen fortgeführt. Ebenso ist es denkbar, das aufgabenbasierte Nutzungskonzept der Videoanalyse in der Schule mit adaptierten Problemstellungen einzusetzen. So ist eine Wirksamkeitsuntersuchung videobasierter Aufgaben in der Schule mit Schwerpunkt auf der mobilen Videoanalyse Gegenstand einer weiteren Studie.

### Übergeordnete Ergebnisse und Aussagen auf Stichprobenebene

Losgelöst von den Fragestellungen zur Wirksamkeit videobasierter Aufgaben betraf ein weiterer Forschungspunkt die methodische Diagnose und inhaltliche Interpretation von Selbsteinschätzungen beim Lösen repräsentationsbezogener Aufgaben. Dafür wurden in dieser Arbeit die Methoden zur Auswertung fragebezogener Antwortsicherheiten systematisch zusammengefasst und der Stand der physikdidaktischen Forschung um diese Thematik dargestellt. Die Ergebnisse zeigten zum einen, dass die Fähigkeit zur adäquaten Beurteilung der eigenen Leistung von der Leistung selbst abhängt (Dunning-Kruger-Effekt; schwächere Studierende überschätzen sich). Zum anderen trat auch bei Studierenden mit ansonsten zutreffender Selbsteinschätzung bei einigen Items eine systematische Überschätzung der eigenen Fähigkeit auf. Daraus lassen sich zwei praktische Implikationen ableiten: Erstens sollten Ansätze zur Verbesserung metakognitiver Fähigkeiten entwickelt werden, um insbesondere leistungsschwachen Studierenden zu helfen, ihre Wahrnehmung der eigenen Leistung zu erkennen, um damit Defizite gezielt überwinden zu können. Zwar rückte die Diagnose metakognitiver Fähigkeiten in jüngster Zeit zunehmend ins Blickfeld physikdidaktischer Forschung, Fördermaßnahmen dafür wurden hingegen noch nicht gezielt entwickelt. Die Feedbackmöglichkeit der Videoanalyse (Theorie-Experiment-Bezug) bietet eine Ansatzmöglichkeit, um Aufgabenstellungen dahingehend zu variieren, metakognitive Prozesse anzuregen. Zweitens lassen sich durch die dargestellten Analysemethoden auf Itemebene Lernschwierigkeiten und Fehlkonzepte quantitativ diagnostizieren, woraus gezielte Instruktionsmaßnahmen abgeleitet werden können.

Die Untersuchung der Hauptfragestellungen erbrachte positive Effekte videobasierter Aufgaben bzgl. Repräsentationskompetenz, Konzeptverständnis, Neugierde, disziplinärer Authentizität und Experimentbezug. Darüber hinaus wurden in dieser Arbeit wichtige Erkenntnisse über die Entwicklung kognitiver und affektiver Variablen im ersten Semester generiert, die z.T. Befunde aus bestehenden Studien replizieren. Diese werden nachfolgend stichpunktartig mit Querverweisen in der Marginalspalte zu den Ergebnisberichten / Diskussionen aufgeführt, in denen die zugehörigen Resultate erläutert und weitere Literaturbezüge hergestellt wurden. Die Ergebnisse beruhen auf dem Kern-Datensatz von 76 Studierenden, bei denen eine ernsthafte Studienabsicht angenommen werden darf (regelmäßiges Erscheinen in den Übungen und Bearbeiten der Aufgaben):

1. Die Beanspruchung durch das Studium steigt in den ersten Semesterwochen stark an ( $d = 0.70$ ). Studierende fühlen sich unabhängig ihrer Vorleistung unter Zeitdruck gesetzt, vor allem durch das Einhalten von Abgabeterminen der Übungsaufgaben.
2. Studierende verbringen zur Bearbeitung der Übungsaufgaben durchschnittlich 8 Zeitstunden auf — das entspricht mehr als dem 1.5-fachen der Präsenzzeit für die Experimentalphysik 1-Lehrveranstaltungen (Vorlesung und Übung; 6 SWS). Aufgabenbearbeitungszeit und empfundene Belastung hängen zusammen ( $r \approx 0.4$ ).
3. Die Schwundquote im ersten Semester (Fachwechsler, Studienaussteiger) beträgt 40% und ist vergleichbar mit den Angaben anderer Fachbereiche (berichtet durch die Konferenz der Fachbereiche Physik, vgl. Düchs & Ingold, 2015).

4. Die meisten Studienanfänger besuchten schulische Leistungskurse in Physik (65%) bzw. Mathematik (79%) und zeigen ein sehr hohes Fachinteresse (vgl. auch Ramm, 2008 sowie Albrecht, 2011, S. 19). Das hohe Fachinteresse nimmt im Laufe des ersten Semesters ab.
5. Interesse und akademisches Selbstkonzept zum Studienbeginn sind starke Prädiktoren für den Verbleib im Studium.
6. Die Abiturnote und die schulische Mathematik-Vorleistung sind starke Prädiktoren für den Verbleib im Studium und für die Klausurnote (vgl. Trapmann et al., 2007).
7. Die schulische Physik-Vorleistung hängt weder mit dem Verbleib im Studium noch mit der Klausur-Leistung zusammen.
8. Vorleistungsstarke Studenten nehmen einen (fehlenden) Experimentbezug von Aufgaben stärker wahr als vorleistungsschwache. Aufgaben ohne Experimentbezug wecken bei vorleistungsstarken Studierenden weniger Neugierde als bei vorleistungsschwachen.
9. Geweckte Neugierde und Experimentbezug der Aufgaben hängen zusammen ( $r = 0.48$ ).
10. Studierende besitzen in der Regel eine gute metakognitive Kalibrierung, d.h. sie verfügen über eine zutreffende Selbsteinschätzung ihrer eigenen Leistung ( $CAQ = 0.52$ ,  $SD = 0.08$ ). Leistung und Adäquatheit der Selbsteinschätzung hängen zusammen ( $r = 0.30$ ). Leistungsschwache Studierende neigen zur Überschätzung ihrer Fähigkeit (Dunning-Kruger-Effekt, vgl. Lindsey & Nagel, 2015; Rebello et al., 2012; Sharma & Bewes, 2011).

### Instrumente und Methodik

Die im Rahmen der Arbeit (weiter)entwickelten Testinstrumente erfüllen die wesentlichen psychometrischen Qualitätskriterien der klassischen Testtheorie und stellen neben den Aufgabenstellungen und Videos ein Produkt dieser Arbeit dar. Sie ermöglichen auch bei Folgeuntersuchung eine zuverlässige Messung wichtiger Variablen, die im Physikstudium eine Rolle spielen. Insbesondere das neu konzipierte KiRC-Instrument schließt eine Lücke in dem umfangreichen Bestand von forschungsbasierten Testinstrumenten, die bisher keine explizite Quantifizierung repräsentationaler Fähigkeiten erlauben. Die Analyse dieses Tests mithilfe der Item-Response-Theorie zeigte die Existenz eines Rasch-skalierbaren Itemkerns. Diese Rasch-Itemgruppe weist eine akzeptable Reliabilität auf ( $\alpha_c = 0.78$ ) und ist in Folgeuntersuchungen zeitökonomisch einsetzbar (Testdauer), um repräsentationale Fähigkeiten im Bereich der Kinematik zu messen. Dabei eignet sich das Instrument besonders zum Nachweis von Lerneffekten repräsentational-angereicherter Innovationsansätze. Weitere Entwicklungsschritte des Instruments umfassen (i) die Bestimmung und Validierung der Itemparameter an weiteren Stichproben, (ii) die Konstruktion schwierigerer Items für eine bessere Passung zu Studierendenchproben und (iii) die Adaption der Items für den Einsatz in der Schule. Wegen der Bedeutung der Repräsentationskompetenz für das Lernen und Verstehen von Physik bietet sich der letztgenannte Punkt insbesondere an, um Defizite repräsentationaler Fähigkeiten frühzeitig zu diagnostizieren. In diesem Zusammenhang konnte durch die Analyse von Antwortsicherheiten gezeigt werden, dass sich vorliegende Fehlkonzepte in charakteristischen Zusammenhängen zwischen „gegebener Antwort“ und „Zuversicht in die Lösung“ manifestieren. Während Fehlvorstellungen bisher primär durch zeitaufwendige qualitative Studien (Interviews) diagnostiziert wurden, bietet die Erhebung von Antwortsicherheiten eine ökonomische Alternative, um Hinweise auf vorhandene Fehlkonzepte zu erhalten.

Neben dem Ansatz, Repräsentationskompetenz produktorientiert unter Verwendung eines Fragebogens zu erfassen, verfolgt die Didaktik-Arbeitsgruppe seit Ende 2015 in Kooperation mit dem Deutschen Zentrum für Künstliche Intelligenz das Ziel, repräsentationale Fähigkeiten prozessorientiert zu messen. Mit Eye-Tracking wird dabei untersucht, welche Unterschiede in

der Art des Umgangs mit Repräsentationen zwischen Novizen und Experten existieren, wenn verschiedene Repräsentationen zum Problemlösen angeboten werden (Mozafari et al., 2016). Das KiRC-Instrument liefert die materielle Basis für diese Untersuchung.

In dieser Arbeit wurde ferner eine Methode entwickelt, um Testsummen eines aus zwei Komponenten zusammengesetzten Tests analytisch zu behandeln. Diese Testmischungsanalyse liefert den Zusammenhang zwischen einer beliebig gewichteten Komposition zweier Testkomponenten und den resultierenden Effektstärken von Vergleichsanalysen (Prä/Post oder KG/IG) und führt diese analytisch auf die Effektstärken der beiden Teilkomponenten zurück. Die Methode erweitert damit die Angabe einer einzelnen Effektstärke aufgrund einer normativ festgelegten Gewichtung (häufig Gleichgewichtung) durch eine Formel, die alle möglichen Gewichtungen abdeckt. Durch Anwendung des Verfahrens auf Reliabilitäten kann zudem bestimmt werden, wie zwei Instrumente kombiniert werden müssen, um eine bestmögliche Messgenauigkeit zu erzielen. In diesem Zusammenhang wurde weiter gezeigt, dass die in der Literatur häufig genutzte Ratekorrektur von Testsummen einen Spezialfall der Testmischung darstellt. Eine Erweiterung der Methode auf mehr als zwei Teilkomponenten und auf andere Testunterteilungen (beispielsweise Skalen des Motivationsfragebogens) wird perspektivisch angestrebt.

Schließlich werden im Folgenden noch einige Forschungsperspektiven bzgl. der Wirkung des Aufgabentyps zusammengefasst, die in der vorliegenden Untersuchung entstanden sind, und über den Rahmen von deren Forschungsfragen hinausgehen

### Weiterführende Forschungsfragen

Aussichtsreiche empirische Folgefragen, die in dieser Arbeit nicht bearbeiten werden konnten, betreffen zum einen die Lernwirksamkeit videobasierter Aufgaben bzgl. weiterer Themen der Experimentalphysik 1 (beispielsweise Rotationsdynamik, Fluidmechanik). Neben einer Steigerung der Repräsentationskompetenz ist vor allem das physikalische Konzeptwissen bei diesen Themen von Interesse. Dazu müssen neu konstruierte oder angepasste Testinstrumente validiert werden. Die Hauptfragestellung dabei wäre, ob videobasierte Aufgaben auch bei inhaltlich komplexeren und theoretisch anspruchsvolleren Themengebieten lerneffektiver als traditionelle Aufgaben sind. Die Fragestellung überträgt sich auf die Inhalte weiterer Experimentalphysik-Lehrveranstaltungen (Optik, Elektromagnetismus, Atomphysik) und wird in der Habilitation von Dr. Sebastian Gröber (Technische Universität Kaiserslautern) aufgegriffen.

Eine weitere wichtige Variable, die in dieser Arbeit nicht berücksichtigt wurde, ist die experimentelle Kompetenz. Es besteht guten Grund zur Annahme, dass die Studierenden durch den Einsatz experimenteller Aufgabenstellungen ihr experimentelles Geschick beim Planen, Durchführen und Auswerten von Versuchen entwickeln bzw. verbessern. Die Ergebnisse haben nachdrücklich gezeigt, dass den Studierenden der experimentelle Bezug der Aufgaben bewusst war und sie erste Erfahrungen mit dem Erheben und Analysieren von Daten sammelten. In diesem Zusammenhang sind positive Auswirkungen auf das anschließende Laborpraktikum zu erwarten.

Wie in Kap. 12.5 diskutiert, liefert die Studie keine Aussage über eine Dosis-Wirkungs-Beziehung. Gerade für die Implementation des Aufgabenformats in andere Fachbereiche wäre es interessant zu wissen, in welchem Umfang die Aufgaben eingesetzt werden müssen, um erwünschte Lerneffekte zu erzielen. Die Beantwortung dieser Frage an nur einem Standort wäre allerdings sehr zeitaufwendig, da aufgrund der Stichprobengröße pro Semester nur eine bis maximal zwei verschiedene Variationen des Aufgabenverhältnisses umgesetzt werden könnten.

Schließlich weisen die Ergebnisse dieser Studie Anhaltspunkte für Gender-Effekte auf. Besonders in der Schule steht die Physikdidaktik vor der Aufgabe, für Schülerinnen attraktive Lernmöglichkeiten zu entwickeln. Zwar deuten die Ergebnisse dieser Studie darauf hin, dass die Leistung und das Interesse von Frauen durch videobasierte Aufgaben besser gefördert werden kann als durch traditionelle Aufgaben, aber es sind Untersuchungen mit größeren

Stichproben nötig, um dieses Ergebnis abzusichern. Qualitative Studien könnten Aufschluss über die Gründe des Gender-Effekts geben.

Es ist anzunehmen, dass die Videoanalyse als Physikmedium weiter Einzug in Schule und Hochschule nehmen wird, und dass die Verbreitung von mobilen Endgeräten in Lehr-/Lernszenarien zunimmt. Diese Arbeit hat gezeigt, dass eine zieladäquate Einbindung der Videoanalyse in ein aufgabenbasiertes Nutzungskonzept das Potential besitzt, das Lernen von Physik auf kognitiven und affektiven Ebenen zu fördern, ohne in Kauf nehmen zu müssen, dass bestehende Kursstrukturen grundlegend geändert werden müssen oder Studierende durch die Einführung eines neuen Mediums überlastet werden.



# Textbezogene Anhänge

In Anhang A-E befinden sich zusätzliche Daten, auf die im Hauptteil verwiesen wurde, sowie Instruktionsmaterialien und Testinstrumente. Weiterhin gibt es einen digitalen Anhang (CD-ROM), der neben dem Haupttext der Arbeit Recherche- und Analyse-Ergebnisse, Überarbeitungen der Instrumente sowie qualitatives Datenmaterial beinhaltet.

## A Zeitschriftenliste der Literaturrecherche

- PhyDid B - Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung
- Physik und Didaktik in Schule und Hochschule (Onlinezeitschrift)
- Unterrichtswissenschaft
- Zeitschrift für Didaktik der Naturwissenschaften (ZfDN; Onlinezeitschrift)
- American Journal of Physics
- Educational Studies
- Electronic Journal of Science Education
- European Journal of Physics
- European Journal of Physics Education
- Interdisciplinary Journal of Problem-based Learning
- International Journal of Computers for Mathematical Learning
- Journal of Educational Multimedia and Hypermedia
- International Journal of Environmental and Science Education
- International Journal of Math and Science Education
- International Journal of Mathematical Education in Science and Technology
- International Journal of Science Education
- International Journal of Research in Science Teaching
- Journal of Computers in Mathematics and Science Teaching
- Journal of Research in Science Teaching
- Journal of Science Education and Technology
- Journal of Science Teacher Education
- Journal of the Learning Sciences
- Learning and Instruction
- Physics Education
- Research in Science and Technological Education
- Research in Science Education
- School Science Review
- School Science and Mathematics
- Science and Education
- Science Education
- Science Education International
- Science Education Review
- Science in School
- Studies in Science Education

## B Testinstrumente

### Fragebögen zur Motivation, Neugierde, Belastung und zum Betreuer

**Tab. A-1:** Motivation-Items mit Quellenangaben

#	Item	Quelle
<b>Interesse / Engagement</b>		
<u>Gefühlsbezogene Valenzen</u>		
M12	Ich bearbeite gerne Übungsaufgaben.	-
M28	Beim Bearbeiten der Aufgaben passiert es, dass ich gar nicht merke, wie die Zeit vergeht.	MAI.25
M35	Auch wenn es anstrengend ist, Übungsaufgaben zu lösen, so bereitet mir die Beschäftigung damit doch Vergnügen.	MAI.17
M41	Das Lösen physikalischer Aufgaben macht mir Spaß.	MAI.01
<u>Wertbezogene Komponente</u>		
M10	Die Inhalte der Übungsaufgaben sind für mich von persönlicher Bedeutung.	FSL.12
M14	Ich lerne durch die Übungsaufgaben für mich selbst.	(FSL.15)
M20	Die Übungsaufgaben sind hilfreich, um die Inhalte besser zu verstehen.	-
M23	Die Aufgaben geben mir die Möglichkeit, neue Themengebiete selbstständig zu erschließen.	-
M40	Die Form der Aufgaben bringen mir persönlich das Thema näher.	-
M45	Den Dingen, die wir in den Physik-Übungen lernen, messe ich persönlich eine große Bedeutung zu.	FSL.16
M53*	Ich bin sicher, das Fach gewählt zu haben, welches meinen persönlichen Neigungen entspricht.	FSL.22
<u>Engagement</u>		
M02	Ich schaue zu Hause in Büchern, im Internet oder ähnlichem nach, um mehr zu den Übungsaufgaben zu erfahren.	MAI.04
M06	Ich beteilige mich aktiv an den Übungen.	MAI.06
M19	Ich beschäftige mich auch über Übungsaufgaben hinaus mit Themen, die mit den darin behandelten Inhalten zu tun haben.	MAI.09
M21	Ich diskutiere mit Anderen über die Übungsaufgaben.	MAI.19
M27	Ich strengte mich an, alle gestellten Übungsaufgaben zu bearbeiten.	(MAI.15)
M37	Unabhängig von Prüfungsanforderungen beschäftige ich mich intensiv mit interessanten Aufgaben.	FSL.21
M52*	Ohne äußeren Druck würde ich mich wohl nicht so regelmäßig mit den Inhalten und Problemen meines Studiums beschäftigen.	FSL.24
<b>Physikbezogenes Selbstkonzept</b>		
M01	Die Übungsaufgaben sind für mich verständlich.	MAI.03
M11	Ich erwarte, dass meine Leistungen in Zukunft gut sein werden.	MAI.08
M26	Es gelingt mir, die Übungsaufgaben zu meiner Zufriedenheit zu bearbeiten.	MAI.11
M30	Ich traue mir zu, eigene Ergebnisse in der Übungsstunde zu diskutieren.	-
M36	Meine Leistungen in den Übungen sind nach meiner Einschätzung gut.	MAI.05
M43	Ich kann den Physik-Übungen konzentriert folgen.	MAI.14
M50	Ich glaube, dass mich meine Kommilitonen für gut in den Übungen halten.	MAI.22
<b>Realitätsbezug / faktische Authentizität</b>		
M07	Die Übungsaufgaben zeigen mir, dass die physikalischen Themen mit der Realität zu tun haben.	-
M15	Die Übungsaufgaben sind hilfreich, um die Physik, die uns alltäglich betrifft, zu verstehen.	MAI.16
M18	Durch die Übungsaufgaben lerne ich, dass physikalische Beobachtungen und Messungen mit dem täglichen Leben zu tun haben.	-
M24	Die Themen der Physikübung sind für Dinge interessant, mit denen ich außerhalb der Universität zu tun habe.	MAI.24
M31	Durch die Übungsaufgaben lernen Dinge, die im Alltag nützlich sind.	MAI.02
M33	Die Übungsaufgaben sind auf Alltagsphänomene bezogen.	MAI.07
M38	Durch die Aufgaben kann ich einen Bezug zwischen physikalischer Theorie und Realität herstellen.	-
M46	In den Übungsaufgaben geht es um Dinge, die im Alltag vorkommen.	MAI.21
M48	Durch die Aufgaben wird mir klar, dass Physik mit dem Alltag zu tun hat.	-

Disziplinäre Authentizität / Professionsbezug		
M03	Ich finde die Übungsaufgaben sinnvoll zum Einüben physikalischer Denkweisen.	[A]
M04	In den Übungsaufgaben lerne ich Arbeitsmethoden kennen, die für Physiker wichtig sind.	[B]
M05	Die Aufgaben geben mir die Möglichkeit, wie ein Physiker zu arbeiten.	[B]
M09	Die Art und Weise, wie die Übungsaufgaben zu lösen sind, ermöglicht mir wissenschaftliches Arbeiten.	[B]
M17	Im Rahmen der Übungsaufgaben entwickle ich eigene weitergehende Fragen zu einem physikalischen Problem.	[C]
M22	Ich erkenne die fachspezifische Relevanz beim Bearbeiten der Übungsaufgaben.	[A]
M25	Durch die Übungsaufgaben lerne ich Strategien kennen, die Physiker zum Problemlösen nutzen.	[B]
M29	Im Rahmen der Übungsaufgaben formuliere ich eigene Hypothesen zu physikalischen Sachverhalten.	[C]
M34	Durch die Bearbeitung der Übungsaufgaben kann ich meine eigenen Fähigkeiten in Physik erweitern.	[D]
M39	Ich nutze eigene Experimente, um eine Aufgabe zu lösen.	[C]
M44	Ich erhebe eigene Daten, die der Klärung einer Fragestellung dienen.	[C]
M49	Im Rahmen der Übungsaufgaben führe ich eigene Experimente durch, um meine Hypothesen zu prüfen.	[E]
Autonomie		
M08	Ich kann selbst mitbestimmen, wie ich eine Übungsaufgabe bearbeite.	(SDS.3)
M16	Der Lösungsweg einer Aufgabe ist fest vorgeschrieben.	[F]
M32	Ich habe Freiheiten, wie ich mit den Aufgaben lernen und arbeiten kann.	(SDS.7)
M47	Ich habe Einfluss darauf, wie ich die Übungsaufgaben bearbeite.	(SDS.3)

*Bemerkungen zu den Quellen:* MAI kennzeichnen Items des „Modified Anchored Instruction“-Motivationsfragebogens (Kuhn, 2010, S. 332/333); FSI kennzeichnen Items aus dem „Fragebogen für Studieninteresse“ (Krapp et al., 1993, S. 350); SDS = Self-Determination Scale (Sheldon & Deci, o. J.). Referenzen in runden Klammern weichen von dem Originalitem sprachlich stark ab. Referenzen in eckigen Klammern stammen nicht aus einer Itemvorlage, sondern sind an die zugrunde liegenden Auffassungen des Authentizitätsbegriffs in den jeweiligen Publikationen angelehnt:

[A] „[...] courses would be effective in supporting learning if the student both recognized and valued the situation.“ (Gilbert, Bulte & Pilot, 2011, S. 821)

[B] „We define authentic activities in science as those that use tools - such as concepts, equations, or physical tools - in ways and for purposes that reflect how the disciplines build, organize, and assess knowledge about the world. [...] In the culture of physics, specific ways of thinking, building and evaluating knowledge, and communicating that knowledge are emphasized and valued. While these activities and evaluations have been and are continuing to be negotiated by members of the physics community, there exists a set of shared practices and values that are shared within the physics discipline and make it distinct even from the other sciences.“ (Watkins et al., 2012, S. 2, 3)

[C] „You have suggested experiments or new measurements (or computations, or simulations) in order to verify your hypotheses, even if you did not carry them out. [...] You have proposed and formulated new hypothesis, reflecting your own representations.“ (Cariou, 2010)

[D] „Students are agents in their own learning and relate own experiences to new knowledge, making learning their own.“ (Rainer & Matthews, 2002, S. 22)

[E] „The individual student ownership of learning refers to an individual students own question / idea that comes from own experiences, interests or anomalies of understanding; an idea / question that comes back several times and leads to new insights.“ (Enghag & Niedderer, 2007, S. 1)

[F] „Problems inherent in the activities are ill-defined and open to multiple interpretations rather than easily solved by the application of existing algorithms.“ (Herrington, Oliver & Reeves, 2003, S. 3)

\*Items wurden nur einem Teil der Stichprobe vorgelegt und bleiben in der Skalenbildung unberücksichtigt.

**Tab. A-2:** Neugierde-Items mit Quellenangaben

#	Item	Quelle
Neugierde als Personeneigenschaft (NP)		
NP1	Neues zu lernen fasziniert mich.	MTC.13
NP2	Dinge, auf die ich stoße, wecken meine Neugierde.	MTC.2
NP3	Es macht mir Spaß, bisherige Vorstellungen zu überdenken.	EC.8
NP4	Es kommt vor, dass ich in einer Sache versunken bin.	MTC.15
NP5	Ich freue mich, etwas über Themen zu lernen, die mir unbekannt sind.	EC.38
NP6	In meiner Freizeit gehe ich spannenden Aktivitäten nach.	MTC.5
NP7	Neue Ideen regen meine Phantasie an.	EC.47
NP8	Wenn ich etwas Neues lerne, möchte ich mehr darüber erfahren.	EC.48

NP9 Ich gehe Dingen, die ich nicht verstehe, auf den Grund. MTC\_19

#### Neugierde als aktueller Zustand (NZ)

NZ1	Durch die Übungsaufgaben (ÜA) wurde meine Phantasie angeregt.	EC_47
NZ2	Die ÜA regten dazu an, neue Inhalte zu entdecken.	EC_45
NZ3	Die ÜA weckten meine Neugierde, mich mit unbekanntem Themen zu befassen.	MSC_8
NZ4	Die ÜA brachten mich dazu, bisherige Vorstellungen zu überdenken.	EC_8
NZ5	Ich möchte die behandelten Inhalte der ÜA gerne weiter vertiefen.	MSC_6
NZ6	Die ÜA führten dazu, neue Ideen zu bekannten Themen zu entwickeln.	MSC_19
NZ7	Ich war überrascht, auf welche Art und Weise wir physikalische Probleme lösen.	MSC_2
NZ8	Die ÜA haben mich neugierig gemacht, mehr über Mechanik zu erfahren.	EC_48

*Bemerkungen zu den Quellen:* MTC / MSC bezeichnen Items des „Melbourne Trait / State Curiosity“ Inventars (Naylor, 1981) unter Verwendung der Itemkennung aus der Originalarbeit; EC bezeichnen Items des „Epistemic Curiosity“ Inventars (Litman & Spielberger, 2003).

**Tab. A-3:** Belastungs-Items mit Quellenangaben

#	Item	Quelle
<b>Beanspruchung durch Studieninhalte (PSQ)</b>		
PSQ1	Ich habe für mein Studium zu viele Dinge zu erledigen.	PSQ_4
PSQ2	Ich habe neben dem Studium genug Zeit für mich selbst.	PSQ_29
PSQ3	Ich fühle mich durch die Abgabetermine für die Übungsaufgaben unter Druck gesetzt.	PSQ_30
PSQ4	Ich bin wegen meines Studiums gehetzt.	PSQ_16
PSQ5	Ich habe das Gefühl, dass im Studium zu viel von mir verlangt wird.	PSQ_02
PSQ6	Ich bin wegen meines Studiums in Eile.	(PSQ_01)
<b>Aufgabenschwierigkeit (CL)</b>		
CL01	Die Übungsaufgaben (ÜA) waren schwierig.	AECL_2
CL02	Das Bearbeiten der ÜA lief für mich flüssig und glatt.	-
CL03	Ich musste mich anstrengen, um den Inhalt der ÜA zu verstehen.	AECL_5
CL04	Es war schwierig für mich, die ÜA zu lösen.	AECL_7
CL05	Ich wusste während der Bearbeitung der ÜA genau, was ich zu tun hatte.	-
CL06	Die ÜA waren eine Herausforderung für mich.	-
CL07	Es war schwierig, wichtige von unwichtigen Informationen in der Aufgabenstellung zu unterscheiden.	AECL_10
CL08	Es war schwierig, die richtigen Informationen zum Lösen der ÜA zu finden.	AECL_13
CL09	Ich musste mich anstrengen, um die ÜA lösen zu können.	AECL_15
CL10	Die ÜA zu verstehen war schwierig.	AECL_17

*Bemerkungen zu den Quellen:* AECL bezeichnen Items des Fragebogens zu „Ankereigenschaften und Cognitive Load“ aus Kuhn (2010) mit der Itemkennung auf S. 343; PSQ bezeichnet die Items aus Levenstein et al. (1993) mit dortiger Itemkennung. Items in Klammern wurden sprachlich stark verändert.

**Tab. A-4:** Betreuer-Items mit Quellenangaben

#	Item
B1	Der Übungsgruppenleiter spornt mich zum Bearbeiten der Übungsaufgaben an.
B2	Ich habe das Gefühl, dass unser Übungsgruppenleiter von Physik begeistert ist.
B3	Der Übungsgruppenleiter ist engagiert.
B4	Der Übungsgruppenleiter ist auf unsere Fragen eingegangen.
B5	Der Übungsgruppenleiter motiviert mich zum Mitarbeiten in der Übungsgruppe.

*Bemerkung:* Die Items B2 und B4 sind in Anlehnung an die Betreuer-Items aus Studien von Netzwerkcollegen formuliert (Hirth, 2016; Molz, 2016).

**Leistungstests: Repräsentationskompetenz in der Kinematik (KiRC) und Konzepttest**

**K1.** Es bezeichne  $x(t)$  den Ort,  $v(t)$  die Geschwindigkeit und  $a(t)$  die Beschleunigung eines Teilchens zum Zeitpunkt  $t$ . Das Teilchen sei zu Beginn der Bewegung in Ruhe, d.h. es sei  $v(0) = 0, x(0) = 0$ .

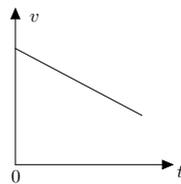
Außerdem gelte

$$\dot{v}(t) = b \cdot t, \quad b > 0.$$

Welche der Aussagen ist korrekt?

- (a) Der Körper bewegt sich mit konstanter Geschwindigkeit  $b$ , d.h.  $x(t) = b \cdot t$ .
- (b) Der Körper wird gleichmäßig beschleunigt, d.h.  $x(t) = 0,5 b t^2$ .
- (c) Der Körper bewegt sich ungleichmäßig beschleunigt, es gilt  $x(t) \propto t^3$ .

**K2.** Die Abbildung zeigt das Geschwindigkeit-Zeit-Diagramm einer Bewegung:

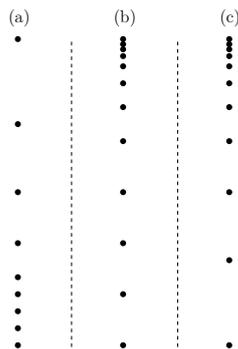


Welche Aussage stimmt?

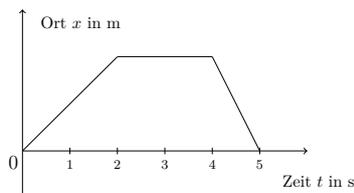
- (a) Der Körper bewegt sich mit konstanter Beschleunigung.
- (b) Der Körper bewegt sich mit gleichmäßig abnehmender Beschleunigung.
- (c) Der Körper bewegt sich vermutlich eine schiefe Ebene hinunter.

**K3.** Ein Gegenstand fällt zum Zeitpunkt  $t = 0$  von dem Dach eines Hochhauses. Die Luftreibung führt nach einiger Zeit zu einer konstanten Fallgeschwindigkeit.

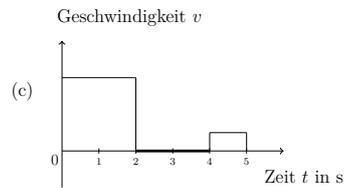
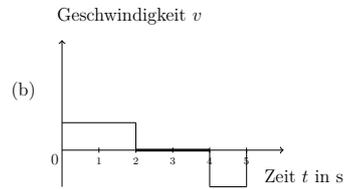
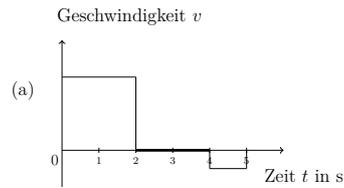
Welche Punktspur der Bewegung stimmt (jeweils gleiche Zeit zwischen den Punkten)?



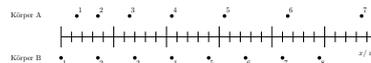
**K4.** Das folgende Diagramm zeigt die Bewegung eines Körpers:



Welches der folgenden Geschwindigkeit-Zeit-Diagramme beschreibt die Bewegung des Körpers korrekt?



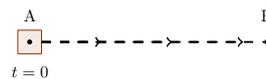
**K5.** Die Abbildung zeigt die Positionen zweier Körper A und B im zeitlichen Abstand von jeweils 0,2 s. Die Körper bewegen sich nach rechts.



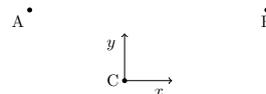
Haben die beiden Körper irgendwann einmal die gleiche Geschwindigkeit?

- (a) Nein
- (b) Ja, zum Zeitpunkt 2
- (c) Ja, irgendwann zwischen Zeitpunkt 3 und Zeitpunkt 4

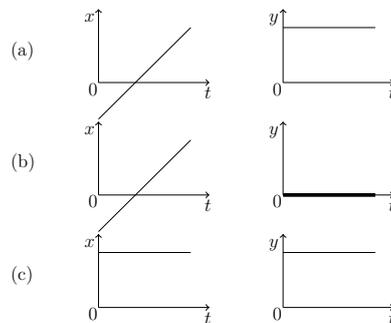
**K6.** Ein Objekt bewegt sich mit konstanter Geschwindigkeit entlang einer Geraden AB von A nach B. Für  $t = 0$  befindet sich das Objekt bei A.



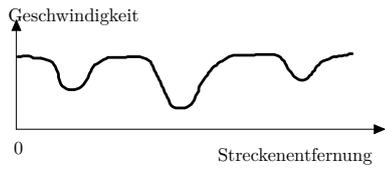
Betrachten Sie das kartesische (rechtwinklige) Koordinatensystem mit Ursprung in C:



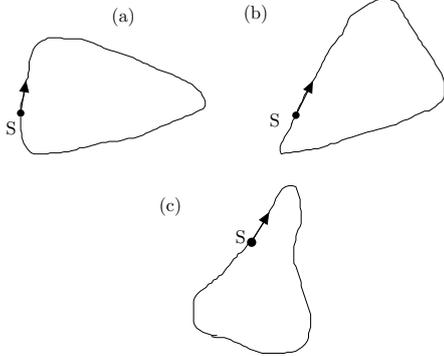
Welche Ort-Zeit-Diagramme beschreiben die Bewegung in dem Koordinatensystem korrekt?



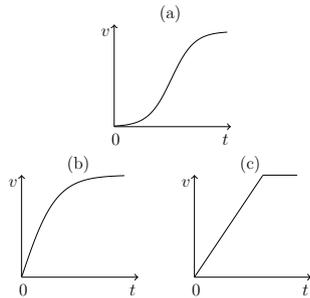
**K7.** Dieser Graph zeigt, wie die Geschwindigkeit eines Rennwagens während seiner zweiten Runde auf einer ebenen Rennstrecke variiert:



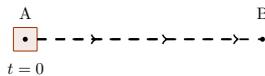
Auf welcher der abgebildeten Rennstrecken fuhr der Wagen, sodass der obige Graph entstand? S markiert die Startposition.



**K8.** Ein Gegenstand fällt zum Zeitpunkt  $t = 0$  von dem Dach eines Hochhauses. Die Luftreibung führt nach einiger Zeit zu einer konstanten Fallgeschwindigkeit. Welches Geschwindigkeit-Zeit-Diagramm beschreibt die Fallbewegung korrekt?



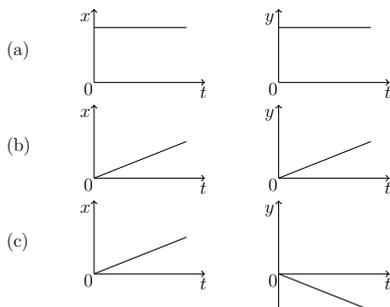
**K9.** Ein Objekt bewegt sich mit konstanter Geschwindigkeit entlang einer Geraden AB von A nach B. Zum Zeitpunkt  $t = 0$  befindet sich das Objekt bei A.



Betrachten Sie das kartesische (rechtwinklige) Koordinatensystem mit Ursprung in A:

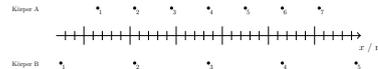


Welche Ort-Zeit-Diagramme beschreiben die Bewegung in dem Koordinatensystem korrekt?



**K10.** Die Abbildung zeigt die Positionen zweier Körper A und B in gleichen zeitlichen Abständen. Die Körper bewegen

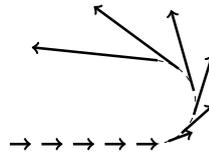
sich nach rechts.



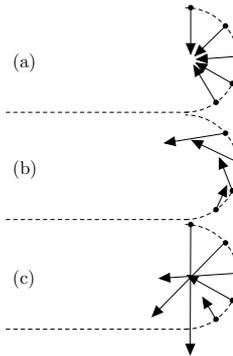
Welche Aussage zur Beschleunigung der beiden Körper ist korrekt?

- (a) Beschleunigung von A > Beschleunigung von B
- (b) Beschleunigung von B > Beschleunigung von A
- (c) Beschleunigung von A = Beschleunigung von B = 0

**K11.** Ein Auto fährt mit konstanter Geschwindigkeit auf eine Kurve zu, in der es gleichmäßig beschleunigt. Die Abbildung zeigt die Vektorspur der Geschwindigkeiten (jeweils gleiche Zeiten zwischen den Bildern):

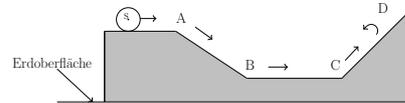


Welche Abbildung zeigt die Beschleunigungsvektoren korrekt?

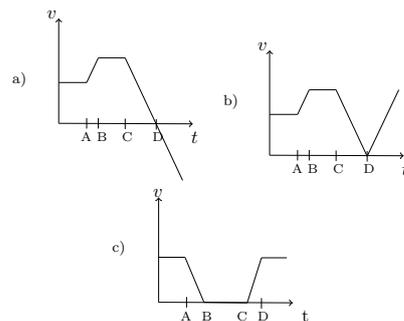


**K12.** Eine Kugel bewegt sich entlang der abgebildeten Bahn. Die Bewegung startet zum Zeitpunkt  $t = 0$  im Punkt S mit der positiven Anfangsgeschwindigkeit  $v_0$ . Am Ende des Pfades kehrt die Kugel die Bewegungsrichtung um. Die Koordinatenachse liegt entlang der Bahn (positive Richtung: nach rechts), sodass die Bewegung der Kugel eindimensional beschrieben wird.

Achtung: Sie schauen seitlich auf die Rampe, nicht von oben.

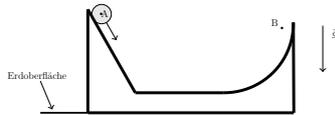


Welches Geschwindigkeit-Zeit-Diagramm beschreibt die Bewegung der Kugel?

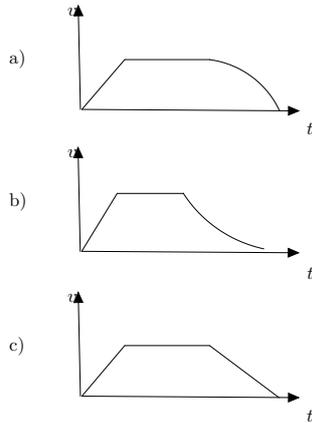


**K13.** Eine Kugel bewegt sich entlang der abgebildeten Bahn. Die Bewegung startet zum Zeitpunkt  $t = 0$  im Punkt A aus der Ruhe  $v(t = 0) = 0$  und endet im Punkt B. Die Koordinatenachse liegt entlang der Bahn (positive Richtung:

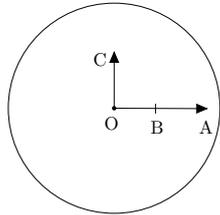
nach rechts), sodass die Bewegung der Kugel eindimensional beschrieben wird.



Welches Diagramm beschreibt die Bewegung der Kugel am ehesten?



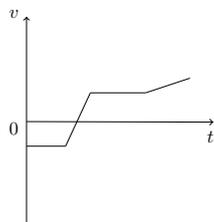
**K14.** Nachfolgende Abbildung zeigt eine Uhr mit Minuten- und Sekundenzeiger. Die Punkte A und C befinden sich auf dem Rand des jeweiligen Zeigers; Punkt B in der Mitte des Sekundenzeigers. Wir betrachten die gleichförmige Kreisbewegung der Punkte A, B und C um den Punkt O, wobei  $\omega_A, \omega_B$  und  $\omega_C$  die Beträge der Winkelgeschwindigkeiten der Punkte bezeichnen.



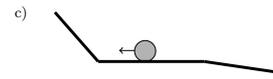
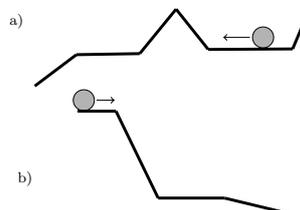
Welche Aussage ist korrekt?

- a)  $\omega_A > \omega_B > \omega_C$ .
- b)  $\omega_A = \omega_B = \omega_C$ .
- c)  $\omega_A = \omega_B > \omega_C$ .

**K15.** Gegeben ist das folgende Geschwindigkeit-Zeit-Diagramm einer Kugel:

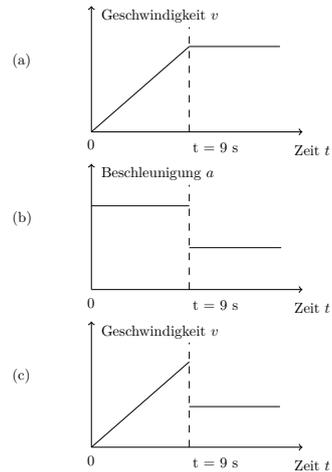


Auf welcher Bahn bewegt sich die Kugel, um diesen Graphen zu erzeugen? Die Koordinatenachse liegt entlang der Bahn (positive Richtung: nach rechts), sodass die Bewegung der Kugel eindimensional beschrieben wird. (Sie schauen stets seitlich auf die Rampe, nicht von oben!)

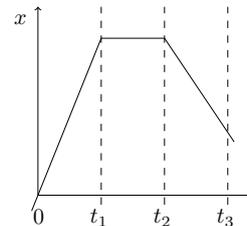


**K16.** Die Abbildung zeigt die Position eines Körpers, der sich aus der Ruhe von links nach rechts bewegt. Zwischen zwei Punkten liegt jeweils der zeitliche Abstand  $\Delta t = 1$  s.

Welches Diagramm beschreibt die Bewegung korrekt ?



**K17.** Betrachten Sie das folgende Weg-Zeit-Diagramm:



$v$  bezeichne die Geschwindigkeit. Welche der folgenden Aussagen sind richtig bzw. falsch?

- |  | richtig               | falsch                |
|--|-----------------------|-----------------------|
| (a) $v(t=0) = 0$ .                           | <input type="radio"/> | <input type="radio"/> |
| (b) $v(t)$ ist stetig in $t \in [0, t_3]$ .  | <input type="radio"/> | <input type="radio"/> |
| (c) $v(t) > 0$ für alle $t \in [0, t_3]$ .   | <input type="radio"/> | <input type="radio"/> |
| (d) $x(t=0) < x(t=t_3)$ .                    | <input type="radio"/> | <input type="radio"/> |
| (e) $\int_0^{t_3} x(t) dt = v(t_3) - v(0)$ . | <input type="radio"/> | <input type="radio"/> |

**K18.** Für die Trajektorien (Bahnkurven) zweier Teilchen (1 und 2) gilt:

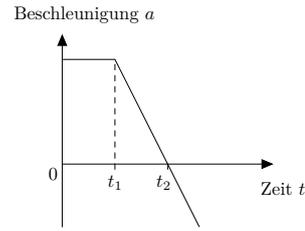
$$\vec{r}_1(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} -v_0 t \\ -\frac{1}{2} g t^2 + y_0 \\ z_0 \end{pmatrix}$$

$$\vec{r}_2(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} v_0 t \\ \frac{1}{2} g t^2 + y_0 \\ z_0 \end{pmatrix}$$

Entscheiden Sie, ob die jeweilige Aussage (a) - (e) richtig oder falsch ist:

**K19.** Die folgenden Diagramme zeigen den Ort  $x$ , die Geschwindigkeit  $v$  und die Beschleunigung  $a$  eines Körpers über der Zeit  $t$ . Kreuzen Sie jeweils an, ob die Diagramme (a) - (e)

- |     |  |                       |                       |
|-----|--|-----------------------|-----------------------|
| (a) | $\vec{r}_1(0) = \vec{r}_2(0).$                         | richtig               | falsch                |
|     |  | <input type="radio"/> | <input type="radio"/> |
| (b) | $\frac{d\vec{r}_1}{dt} = \frac{d\vec{r}_2}{dt}.$       | <input type="radio"/> | <input type="radio"/> |
| (c) | $\vec{r}_1(t) = -\vec{r}_2(t).$                        | <input type="radio"/> | <input type="radio"/> |
| (d) | $\vec{r}_1(t) = \vec{r}_2(-t).$                        | <input type="radio"/> | <input type="radio"/> |
| (e) | für $y_0 = 0$ : $\ \vec{r}_1(t)\  = \ \vec{r}_2(t)\ .$ | <input type="radio"/> | <input type="radio"/> |



Entscheiden Sie, ob die jeweilige Aussage (a) - (e) richtig oder falsch ist:

eine Bewegung mit konstanter, von Null verschiedener Beschleunigung beschreiben:

- |     |  |                          |                            |
|-----|--|--------------------------|----------------------------|
| (a) |  | ja <input type="radio"/> | nein <input type="radio"/> |
| (b) |  | ja <input type="radio"/> | nein <input type="radio"/> |
| (c) |  | ja <input type="radio"/> | nein <input type="radio"/> |
| (d) |  | ja <input type="radio"/> | nein <input type="radio"/> |
| (e) |  | ja <input type="radio"/> | nein <input type="radio"/> |

- |     |  |                       |                       |
|-----|--|-----------------------|-----------------------|
| (a) | Der Körper wird zunächst gleichmäßig beschleunigt.                       | richtig               | falsch                |
|     |  | <input type="radio"/> | <input type="radio"/> |
| (b) | Der Körper wird im Zeitintervall $[t_1; t_2]$ abgebremst.                | <input type="radio"/> | <input type="radio"/> |
| (c) | Der Körper erreicht seine maximale Geschwindigkeit zum Zeitpunkt $t_1$ . | <input type="radio"/> | <input type="radio"/> |
| (d) | Der Körper kehrt während der Bewegung seine Richtung um.                 | <input type="radio"/> | <input type="radio"/> |
| (e) | Der Körper befindet sich zum Zeitpunkt $t_2$ in Ruhe.                    | <input type="radio"/> | <input type="radio"/> |

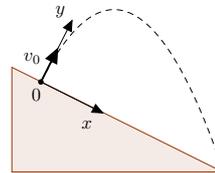
**K20.** Beim schiefen Wurf mit Abwurfwinkel  $\alpha$  und Anfangsgeschwindigkeit  $v_0$  gelte für die Bahnkurve  $\vec{r}(t)$  eines Objekts:

$$\vec{r}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} v_0 t \cos \alpha \\ -\frac{1}{2} g t^2 + v_0 t \sin \alpha \end{pmatrix}$$

Entscheiden Sie, ob die jeweilige Aussage (a) - (e) richtig oder falsch ist:

- |     |  |                       |                       |
|-----|--|-----------------------|-----------------------|
| (a) | Das Koordinatensystem, in dem die Bewegung beschrieben wird, wurde so gewählt, dass das Objekt eine Anfangshöhe $h \neq 0$ besitzt.  | richtig               | falsch                |
|     |  | <input type="radio"/> | <input type="radio"/> |
| (b) | Die Geschwindigkeit des Objekts in $x$ -Richtung ist konstant.   | <input type="radio"/> | <input type="radio"/> |
| (c) | Die Beschleunigung des Objekts ist zu jedem Zeitpunkt $t$ gegeben durch $\ddot{\vec{r}} = \begin{pmatrix} 0 \\ -g \end{pmatrix}.$  | <input type="radio"/> | <input type="radio"/> |
| (d) | Die Geschwindigkeit $\dot{\vec{r}}(t)$ und der Ortsvektor $\vec{r}(t)$ stehen zu allen Zeiten $t$ orthogonal zueinander, d.h. $\dot{\vec{r}}(t) \cdot \vec{r}(t) = 0$ für alle $t$ . | <input type="radio"/> | <input type="radio"/> |
| (e) | Die Geschwindigkeit des Objekts zum Zeitpunkt $t = 0$ ist gegeben durch $\dot{\vec{r}}(0) = \begin{pmatrix} v_0 \cos \alpha \\ v_0 \sin \alpha \end{pmatrix}.$                       | <input type="radio"/> | <input type="radio"/> |

**K22.** Eine Kugel wird zum Zeitpunkt  $t = 0$  s von einem Hang aus mit der Anfangsgeschwindigkeit  $v_0$  senkrecht zur Hangebene nach oben geworfen. Betrachten Sie das kartesische Koordinatensystem mit Ursprung  $(0, 0)$  im Abwurfpunkt, dessen  $x$ -Achse parallel zur Hangrichtung liegt:



$x_W$  bezeichne die Wurfweite,  $t_W$  die Wurfedauer. Betrachten Sie die folgenden Aussagen über die Bewegung der Kugel im eingezeichneten Koordinatensystem  $(x, y)$ :

Welche Aussagen sind richtig bzw. falsch?

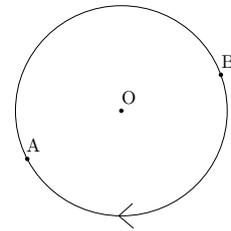
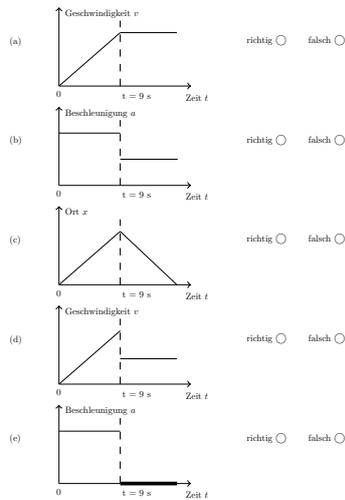
- |     |  |                       |                       |
|-----|--|-----------------------|-----------------------|
| (a) | Die Anfangsgeschwindigkeit $v_0$ besitzt nur eine Komponente in $y$ -Richtung. | richtig               | falsch                |
|     |  | <input type="radio"/> | <input type="radio"/> |
| (b) | Die maximale Höhe $y_{max}$ wird nach der halben Wurfedauer $t_W/2$ erreicht.  | <input type="radio"/> | <input type="radio"/> |
| (c) | Die Bahnkurve $y(x)$ ist eine Parabel.   | <input type="radio"/> | <input type="radio"/> |
| (d) | Die Bewegung der Kugel ist in $x$ -Richtung beschleunigt.                      | <input type="radio"/> | <input type="radio"/> |
| (e) | Die Starthöhe $y(x = 0)$ und die Endhöhe $y(x = x_w)$ sind identisch.          | <input type="radio"/> | <input type="radio"/> |

**K21.** Gegeben ist das folgende Diagramm zur Bewegung eines Körpers, der sich zunächst in Ruhe befindet,  $v(t = 0) = 0$ :

**K23.** Die Abbildung zeigt die Position eines Körpers, der sich aus der Ruhe von links nach rechts bewegt. Zwischen zwei Punkten liegt jeweils der zeitliche Abstand  $\Delta t = 1$  s.



Kreuzen Sie an, ob das jeweilige Diagramm die Bewegung **K25**. Ein Objekt bewegt sich mit gleichmäßig zunehmender Geschwindigkeit auf einer Kreisbahn um den Punkt O. Betrachten Sie die beiden ortsfesten Punkte A und B auf der Kreisbahn und entscheiden Sie, ob die jeweilige Aussage (a)-(e) jeweils wahr bzw. falsch ist.



**K24.** Ein Gegenstand fällt zum Zeitpunkt  $t = 0$  vom Dach eines Hochhauses. Die Luftreibung führt nach einiger Zeit zu einer konstanten Fallgeschwindigkeit in positive  $z$ -Richtung. Es bezeichne  $z$  die Position,  $v$  die Geschwindigkeit und  $a$  die Beschleunigung des Gegenstandes.

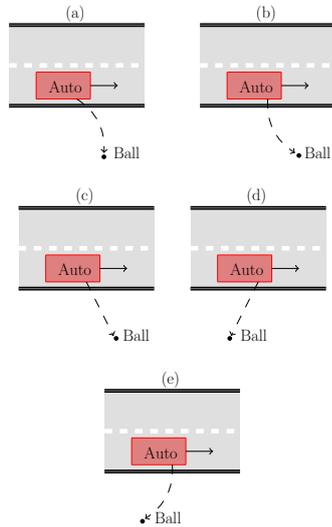
Entscheiden Sie, ob die jeweilige Aussage (a) - (e) richtig oder falsch ist:

- (a) Für  $t \ll 1$  gilt  $v(t) \approx gt$ . richtig falsch
- (b) Es gilt  $\frac{d}{dt}v(t) = const.$  für alle Zeiten  $t$ .
- (c) Für  $t \rightarrow \infty$  gilt  $z(t) = 0,5 g t^2$ .
- (d) Es gilt  $\lim_{t \rightarrow \infty} \frac{dv}{dt} = 0$ .
- (e) Die Beschleunigung ist zum Zeitpunkt  $t = 0$  maximal.

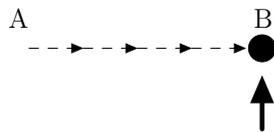
- (a) Der Betrag  $\omega = |\vec{\omega}|$  der Winkelgeschwindigkeit unterscheidet sich in den Punkten A und B. richtig falsch
- (b) Die Richtung der Winkelgeschwindigkeit  $\vec{\omega}$  ist in den Punkten A und B verschieden.
- (c) Der Vektor der Winkelgeschwindigkeit  $\vec{\omega}$  liegt in der Bewegungsebene (die Ebene, in der auch der Kreis liegt, also die Papierebene).
- (d) Die Winkelbeschleunigung  $\vec{\alpha} = \frac{d}{dt}\vec{\omega}$  ist in den Punkten A und B identisch (gleicher Betrag und gleiche Richtung).
- (e) Der Vektor der Winkelbeschleunigung  $\vec{\alpha}$  ist während der Bewegung von A nach B parallel zum Vektor der Winkelgeschwindigkeit  $\vec{\omega}$ .

**Kon1** Ein Ball wird aus dem Beifahrerfenster eines fahrenden Autos geworfen, welches sich konstant nach rechts bewegt (aus Sicht von oben, siehe Skizzen). Es wird angenommen, dass der Ball senkrecht zur Fahrtrichtung des Autos geworfen wird. Die Betrachtung erfolgt von oben! Vernachlässigen Sie die Luftreibung!

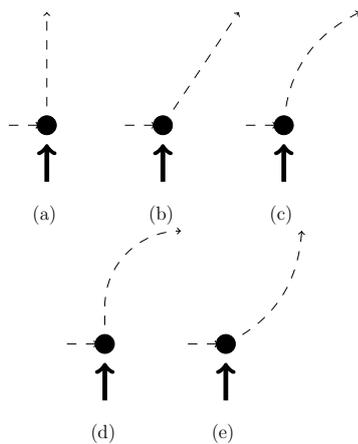
Welche Bahnkurve beschreibt die Bewegung des Balls korrekt?



**Kon2** Die Abbildung zeigt eine Scheibe, die mit konstanter Geschwindigkeit  $v_0$  auf einer reibungsfreien horizontalen Oberfläche (Tisch) von Punkt A nach Punkt B gleitet. Wenn die Scheibe den Punkt B erreicht, erhält sie einen kurzen Stoß senkrecht zur Bewegungsrichtung (dicker Pfeil).

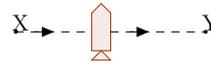


Welche der gezeigten Bahnkurven beschreibt die Bewegung der Scheibe nach dem Stoß korrekt?

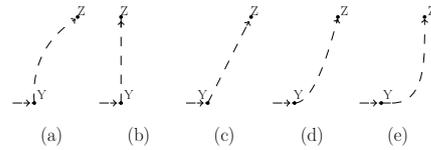


**Kon3** Ein Raumschiff bewegt sich im Weltraum ohne den Einfluss von äußeren Kräften von Position X nach Position Y. An Punkt Y wird das Raketentriebwerk gezündet und das Raumschiff bewegt sich auf einen Punkt Z zu. Das Raketentriebwerk erzeugt eine konstante Beschleunigung im rechten Winkel zur Linie XY.

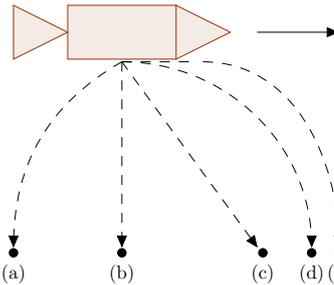
Z



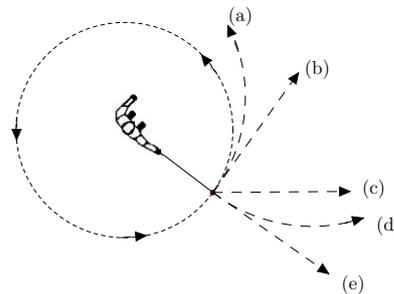
Auf welcher Bahnkurve bewegt sich das Raumschiff von Y nach Z?



**Kon4** Eine Kugel fällt aus dem Frachtraum eines Flugzeuges während das Flugzeug mit konstanter Geschwindigkeit in horizontaler Richtung fliegt. Vernachlässigen Sie Luftreibungskräfte! Der Vorgang wird von der Erde aus beobachtet. Welche Bahnkurve beschreibt die Kugel?



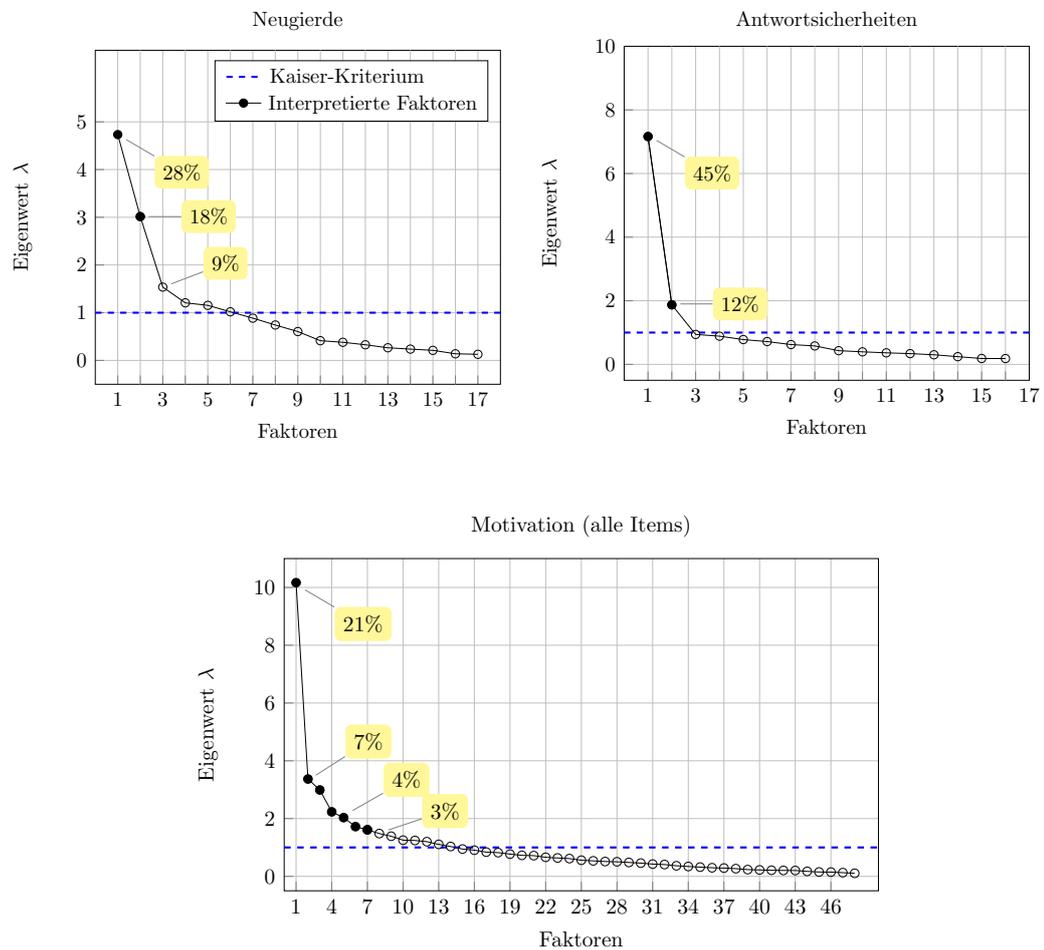
**Kon5** Ein schwerer Ball ist an einem Faden befestigt und wird, wie in der Abbildung gezeigt, im Kreis horizontal herumgeschwungen. An dem gekennzeichneten Punkt reißt plötzlich der Faden. Der Vorgang wird von oben betrachtet. Welchen Weg nimmt der Ball, nachdem der Faden gerissen ist?



## C Ergänzende Ergebnisberichte

### Ergebnisberichte der Testskalierung (KTT)

ScreepLOTS:



**Abb. A-1:** Screeplots zur Faktorenanalyse bzgl. den Neugierde-, Motivations- und Sicherheits-Items. Die erklärte Varianz einiger Faktoren ist hervorgehoben.

## Ergebnisberichte zur Rasch-Analyse

Zur Analyse des KiRC-Instruments:

**Tab. A-5:** Ausgewählte Ergebnisse des Wald-Tests

Item	Teilung „Ma-LK“			„Median-Split“		
	$z$	$p$	Sign. <sup>a</sup>	$z$	$p$	Sign. <sup>a</sup>
K2	2.06	.040	n.s.	-1.62	.106	n.s.
K3	-2.06	.040	n.s.	1.13	.261	n.s.
K4	2.20	.026	n.s.	-	-	-
K5	2.15	.032	n.s.	.54	.590	n.s.
K6	-1.48	.139	n.s.	-.88	.380	n.s.
K9	-1.30	.195	n.s.	-1.08	.282	n.s.
K10	1.50	.135	n.s.	-	-	-
K11	-.88	.378	n.s.	2.59	.010	n.s.
K17a	1.27	.204	n.s.	-.69	.492	n.s.
K17e	-.24	.811	n.s.	1.58	.115	n.s.
K18d	.86	.388	n.s.	.54	.590	n.s.
K20a	-.06	.954	n.s.	.80	.424	n.s.
K20c	.30	.763	n.s.	-.88	.379	n.s.
K20d	2.28	.022	n.s.	-.68	.497	n.s.
K20e	-1.27	.204	n.s.	.71	.475	n.s.
K21c	-.07	.944	n.s.	-1.65	.099	n.s.
K21d	-.07	.944	n.s.	.20	.841	n.s.
K21e	-.91	.363	n.s.	-1.06	.288	n.s.
K22a	-.54	.586	n.s.	-.18	.854	n.s.
K22c	-1.05	.291	n.s.	3.10	.002	n.s.
K22d	-1.03	.301	n.s.	1.52	.130	n.s.
K22e	-1.32	.187	n.s.	-.37	.715	n.s.

*Bemerkung:* Fehlende Werte (-) konnten aufgrund zu geringer Gruppengrößen infolge der Datensatzteilung nicht berechnet werden.

<sup>a</sup>Signifikanzniveau  $\alpha = 0.002$

Zur Analyse des Konzepttests:

**Tab. A-6:** Ergebnisse des Andersen LR-Tests und des Wald-Tests (Konzept-Items)

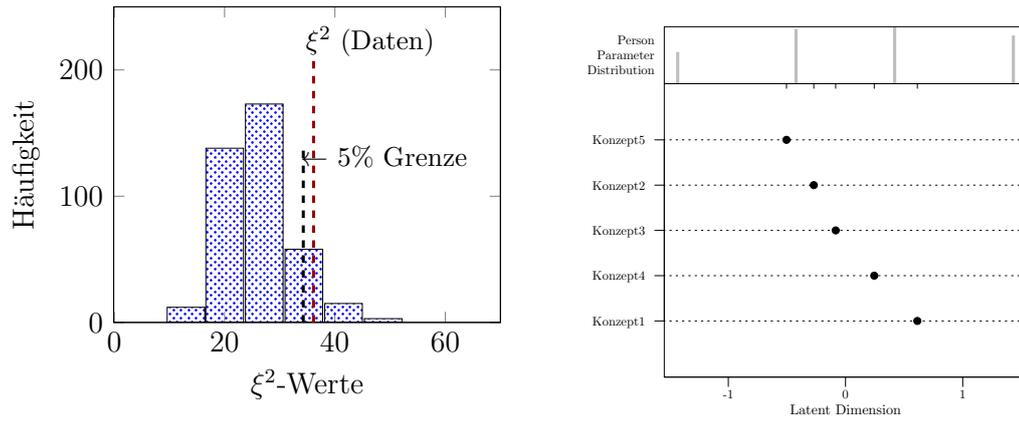
Teilungskriterium	Testwert	$p$ -Wert	Sign. ( $\alpha = 0.07$ )
Geschlecht	15.12	.004	!
MaLK	6.10	.192	n.s.
PhyLK	7.82	.098	n.s.
Vorleistung	7.24	.124	n.s.
Semester	5.32	.256	n.s.
Hauptfach	6.35	.174	n.s.
Median	10.53	.032	n.s.
Item	Testwert	$p$ -Wert	Sign. ( $\alpha = 0.01$ )
Kon1	1.07	.286	n.s.
Kon2	3.08	.002	!
Kon3	-1.44	.150	n.s.
Kon4	-1.50	.135	n.s.
Kon5	-1.61	.108	n.s.

*Bemerkung:* Da das Teilungskriterium „Geschlecht“ zu einem signifikanten Ergebnis führte (!), wurden anschließend Wald-Tests unter Verwendung dieses Teilungskriteriums auf Itemebene durchgeführt (untere Tabellenhälfte).

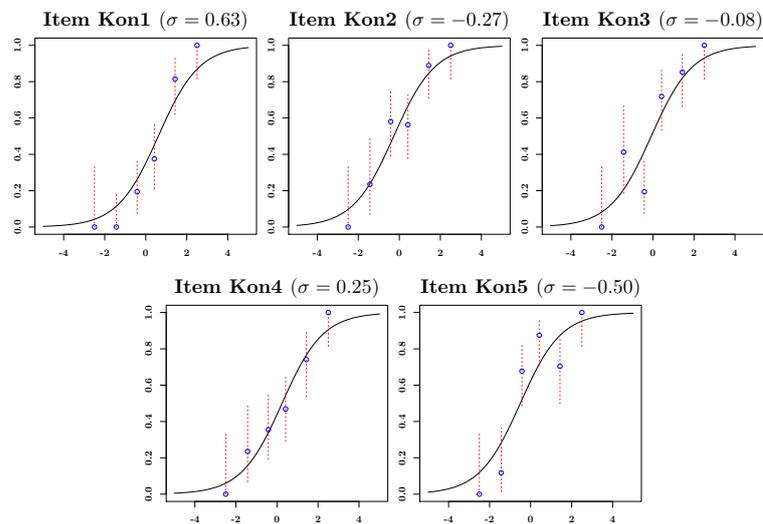
**Tab. A-7:** Fit-Statistiken der Konzept-Items und Korrelationsmatrix

Item	Itemparameter $\sigma$	Outfit	Infit	Fit-Typ	Tetrachorische Korrelationen			
					Kon1	Kon2	Kon3	Kon4
Kon1	0.63	.75	.81	(-) <sup>a</sup>	1			
Kon2	-0.27	1.00	1.03	ok	.51	1		
Kon3	-0.08	1.01	.98	ok	.41	.14	1	
Kon4	0.25	1.10	1.07	ok	.35	.07	.28	1
Kon5	-0.50	1.13	.99	ok	.25	.27	.28	.17

<sup>a</sup>Mixed-Fit mit Tendenz zum Overfit.



**Abb. A-2:** Globale Modellprüfung (Bootstrapping, links) sowie Verteilung der Personen- und Itemparameter auf einer gemeinsamen latenten Dimension (rechts).



**Abb. A-3:** Aufgabencharakteristische Kurven  $\mathcal{P}_\sigma(\eta)$  (ICCs) der Konzept-Items mit empirischen Daten ( $\circ$ ) und Konfidenzintervallen ( $\cdots$ ).



**Ergebnisberichte zur Hauptanalyse**

Voranalysen und Kovarianzmodelle:

**Tab. A-8:** Test auf Homogenität der Regressionssteigungen

AV	Interaktionseffekte „Gruppe × KV“ <sup>a</sup>							
	Abinote		Ma.-Vorleistung		Phy.-Vorleistung		AV <sub>0</sub>	
	F	p	F	p	F	p	F	p
KiRC <sub>1</sub>	.68	.413	.06	.803	.09	.771	.00	.980
Kon <sub>1</sub>	.57	.454	.39	.533	.01	.909	1.27	.264
RA <sub>1</sub>	.55	.462	.01	.918	.30	.588	.00	.951
IE <sub>1</sub>	.09	.763	.18	.677	.07	.794	2.64	.108
SK <sub>1</sub>	1.27	.263	2.27	.136	.30	.583	.24	.629
dA <sub>1</sub>	.44	.509	.44	.509	.90	.346	3.20	.078
Exp <sub>1</sub>	4.59	.036	2.23	.139	5.91	.018	.45	.507
Atnm <sub>1</sub>	1.68	.198	1.68	.198	2.35	.129	.02	.882
NZ <sub>1</sub>	2.23	.140	4.20	.044	1.00	.320	<sup>b</sup>	<sup>b</sup>
PSQ <sub>1</sub>			---	n.z.p. ---			17.43	.000
CL <sub>1</sub>			---	n.z.p. ---			.82	.369
Kon <sub>2</sub>			---	n.z.p. ---			.24	.623
RA <sub>2</sub>			---	n.z.p. ---			.16	.687
IE <sub>2</sub>			---	n.z.p. ---			.01	.920
SK <sub>2</sub>			---	n.z.p. ---			.25	.620
dA <sub>2</sub>			---	n.z.p. ---			.00	.960
Exp <sub>2</sub>			---	n.z.p. ---			2.24	.140
Atnm <sub>2</sub>			---	n.z.p. ---			.04	.851
PSQ <sub>2</sub>			---	n.z.p. ---			7.16	.010
CL <sub>2</sub>			---	n.z.p. ---			.25	.618

*Bemerkungen:* AV = abhängige Variable. Als Kovariaten (KV) wurden die drei Vorleistungsindikatoren geprüft sowie die jeweils zugehörige Prä-Variable. Zur Festlegung der ANCOVA Modelle (Kap. 11.1.5) sind für die AV der Zeitpunkte 2 (untere Tabellenhälfte) sowie für CL<sub>1</sub> und PSQ<sub>1</sub> nur die Interaktionseffekte mit den jeweiligen Prä-Variablen zu prüfen (vgl. Tab. 34 und 35). Die Homogenität der Regressionssteigungen bzgl. den Vorleistungsvariablen war bei den genannten Variablen nicht zu prüfen (n.z.p.). Außerdem musste nach Tab. 34 der Interaktionseffekt zwischen CL<sub>0</sub> und SK<sub>1</sub> überprüft werden — dieser konnte ausgeschlossen werden.

<sup>a</sup>df = 1, df<sub>F</sub> = 70 – 72. <sup>b</sup>NZ wurde zum Prä-Zeitpunkt nicht erhoben.

**Tab. A-9:** Kovarianzmodelle

Hyp	AV	KV <sup>b</sup> ( <i>r</i> >0)	Voraussetzungen ANCOVA <sup>a</sup>			Aufnahme <sup>c</sup>
			Normal	Unabh.	Homogen	
1.1	KiRC <sub>1</sub>	KiRC <sub>0</sub>	•	•	•	•
		AbiNote	•	•	•	•
1.2	Kon <sub>1</sub>	Kon <sub>0</sub>	○	•	•	•
		Ma.-Vorleistung	○	•	•	•
	Kon <sub>2</sub>	○	•	•	•	
	Kon <sub>2-1</sub>	-				
1.3	RA <sub>1</sub>	RA <sub>0</sub>	•	•	•	•
	RA <sub>2</sub>	RA <sub>0</sub>	•	•	•	•
	RA <sub>2-1</sub>	-				
	IE <sub>1</sub>	IE <sub>0</sub>	•	○	•	(•)
	IE <sub>2</sub>	IE <sub>0</sub>	•	○	•	(•)
	IE <sub>2-1</sub>	-				
SK <sub>1</sub>	SK <sub>0</sub>	•	•	•	•	
	CL <sub>0</sub>	•	•	•	•	
SK <sub>2</sub>	SK <sub>0</sub>	•	•	•	•	

	SK <sub>2-1</sub>	-				
	dA <sub>1</sub>	dA <sub>0</sub>	•	•	•	•
	dA <sub>2</sub>	dA <sub>0</sub>	•	•	•	•
	dA <sub>2-1</sub>	-				
	Exp <sub>1</sub>	Exp <sub>0</sub>	•	○	•	(•)
	Exp <sub>2</sub>	Exp <sub>0</sub>	•	○	•	(•)
	Exp <sub>2-1</sub>	-				
	Atnm <sub>1</sub>	Atnm <sub>0</sub>	•	•	•	•
	Atnm <sub>2</sub>	Atnm <sub>0</sub>	•	•	•	•
	Atnm <sub>2-1</sub>	-				
1.4	NZ <sub>1</sub>	-				
	NZ <sub>2-1</sub>	-				
2.1	CL <sub>1</sub>	CL <sub>0</sub>	•	•	•	•
	CL <sub>2</sub>	CL <sub>0</sub>	•	•	•	•
	CL <sub>2-1</sub>	-				
2.2	PSQ <sub>1</sub>	PSQ <sub>0</sub>	•	•	○	○
	PSQ <sub>2</sub>	PSQ <sub>0</sub>	•	•	○	○

*Bemerkungen:* AV = abhängige Variable.

<sup>a</sup>Gruppenunabhängigkeit gemäß Tab. 36; Homogenität der Regressionssteigungen gemäß Tab. A-8. Die Daten zur Analyse der Normalverteilung sind nicht berichtet. •/○ = erfüllt/nicht erfüllt. <sup>b</sup>Auswahl der potentiellen Kovariaten (= KV) gemäß Tab. 34 und 35. <sup>c</sup>Entscheidung über die Berücksichtigung der KV im Modell: • = ja, ○ = nein, (•) = Zwei Kovarianzanalysen werden durchgeführt (mit/ohne KV).

Deskriptive Daten:

**Tab. A-10:** Mittelwerte und Standardabweichungen der abh. Variablen

AV	KG-IG		IG-KG		AV	KG-IG		IG-KG	
	M	SD	M	SD		M	SD	M	SD
KiRC <sub>0</sub>	.67	.18	.67	.22	Atnm <sub>0</sub>	.68	.15	.72	.15
KiRC <sub>1</sub>	.75	.17	.86	.12	Atnm <sub>1</sub>	.60	.16	.60	.18
Konzept <sub>0</sub>	.54	.30	.53	.29	Atnm <sub>2</sub>	.62	.14	.63	.17
Konzept <sub>1</sub>	.63	.23	.72	.26	NZ <sub>1</sub>	.59	.13	.65	.13
Konzept <sub>2</sub>	.73	.24	.74	.28	NZ <sub>2</sub>	.60	.14	.63	.12
RA <sub>0</sub>	.63	.11	.68	.14	NP <sub>0</sub>	.79	.11	.79	.10
RA <sub>1</sub>	.58	.14	.67	.12	NP <sub>1</sub>	.75	.12	.75	.10
RA <sub>2</sub>	.58	.14	.65	.14	NP <sub>2</sub>	.72	.13	.75	.11
IE <sub>0</sub>	.72	.08	.78	.12	PSQ <sub>0</sub>	.55	.17	.59	.13
IE <sub>1</sub>	.68	.11	.72	.14	PSQ <sub>1</sub>	.68	.13	.67	.18
IE <sub>2</sub>	.69	.11	.73	.13	PSQ <sub>2</sub>	.61	.14	.69	.16
SK <sub>0</sub>	.60	.12	.62	.10	CL <sub>0</sub>	.43	.16	.42	.19
SK <sub>1</sub>	.59	.13	.64	.12	CL <sub>1</sub>	.56	.12	.53	.12
SK <sub>2</sub>	.66	.13	.67	.13	CL <sub>2</sub>	.49	.08	.52	.12
dA <sub>0</sub>	.73	.08	.74	.10	Exp <sub>0</sub>	.45	.14	.51	.15
dA <sub>1</sub>	.68	.09	.71	.13	Exp <sub>1</sub>	.42	.15	.57	.13
dA <sub>2</sub>	.66	.09	.69	.12	Exp <sub>2</sub>	.54	.15	.48	.18
ToT <sub>1</sub>	7.76	3.05	8.22	3.69					
ToT <sub>2</sub>	7.01	3.22	6.67	2.45					

Tab. A-11: Deskriptive Statistiken der KiRC-Items auf Gruppenebene

Item	RF	Inhalt	Prä-Test				Post-Test				Lernzuwächse					
			KG		IG		KG		IG		$P_j(t_1) - P_j(t_0)$			$g_H$		
			$P_j(t_0)$	$\sigma_N$	$P_j(t_0)$	$\sigma_N$	$P_j(t_1)$	$\sigma_N$	$P_j(t_1)$	$\sigma_N$	KG	IG	$\Delta_{IG-KG}$	KG	IG	$\Delta_{IG-KG}$
K1*	F	1	.60	.08	.39	.08	.68	.08	.47	.08	.08	.08	.01	.19	.14	-.05
K2	D	1	.78	.07	.75	.07	.73	.07	.83	.06	-.05	.08	.13	-.22	.33	.56
K3	B	2	.85	.06	.69	.08	.88	.05	.92	.05	.03	.22	.20	.17	.73	.56
K4	D	1	.75	.07	.89	.05	.83	.06	.92	.05	.08	.03	-.05	.30	.25	-.05
K5	B	1	.63	.08	.75	.07	.65	.08	.92	.05	.03	.17	.14	.07	.67	.60
K6	B↔D	3	.68	.08	.67	.08	.75	.07	.89	.05	.08	.22	.15	.23	.67	.44
K7*	B↔D	4	.88	.05	.97	.03	.93	.04	.97	.03	.05	.00	-.05	.40	.00	-.40
K8	D	2	.70	.07	.61	.08	.75	.07	.69	.08	.05	.08	.03	.17	.21	.05
K9	B↔D	3	.88	.05	.86	.06	.80	.06	.97	.03	-.08	.11	.19	-.60	.80	1.40
K10	B	1	.88	.05	.86	.06	.90	.05	.89	.05	.03	.03	.00	.20	.20	.00
K11	B	4,5	.45	.08	.33	.08	.50	.08	.67	.08	.05	.33	.28	.09	.50	.41
K12*	B↔D	4	.29	.11	.57	.11	.65	.12	.78	.09	.35	.22	-.14	.50	.50	.00
K13*	B↔D	4	.71	.11	.78	.09	.82	.10	.83	.08	.12	.04	-.07	.40	.20	-.20
K14*	B↔F	6	.59	.12	.65	.10	.71	.11	.78	.09	.12	.13	.01	.29	.38	.09
K15*	B↔D	4	.65	.12	.70	.10	.71	.11	.83	.08	.06	.13	.07	.17	.43	.26
K16*	B↔F	1	.76	.11	.74	.09	.88	.08	.91	.06	.12	.17	.06	.50	.67	.17
K17	D↔F	1	.69	.03	.81	.03	.76	.03	.85	.03	.07	.05	-.03	.24	.24	.00
K18	F	5,7	.88	.02	.86	.03	.88	.02	.87	.03	.00	.01	.02	-.04	.08	.12
K19*	D	1	.92	.03	.95	.03	.96	.02	.98	.02	.03	.03	.00	.43	.69	.26
K20	F	5,7	.73	.03	.75	.03	.82	.03	.86	.03	.10	.12	.02	.36	.46	.10
K21	D↔F	1	.70	.03	.71	.03	.84	.03	.90	.02	.14	.19	.05	.46	.64	.18
K22	B↔F	3,5,7	.49	.04	.50	.04	.59	.03	.73	.03	.11	.23	.13	.20	.46	.26
K23*	B↔F	1	.76	.04	.80	.05	.84	.03	.98	.02	.08	.18	.10	.34	.92	.58
K24*	F	2	.52	.05	.70	.04	.75	.05	.94	.02	.24	.24	.00	.49	.79	.30
K25*	B↔F	4,6	.42	.05	.48	.05	.65	.05	.87	.03	.22	.38	.16	.39	.74	.35

*Bemerkungen:* Für Kodierung der Inhalte und Repräsentationsformen (= RF) siehe Tab. 4.  $P_j(t_i)$  = Schwierigkeitsindex des Items  $j$  zum Zeitpunkt  $t_i$ ,  $P_j(t_1) - P_j(t_0)$  = Lernzuwachs auf Itemebene,  $\Delta$  = Differenzen der Lernzuwächse zwischen Interventionsgruppe (= IG) und Kontrollgruppe (= KG),  $\sigma_N$  = Standardfehler des Mittelwerts,  $g_H$  = normierter Lernzuwachs nach Hake, siehe Kap. 10.5.4.

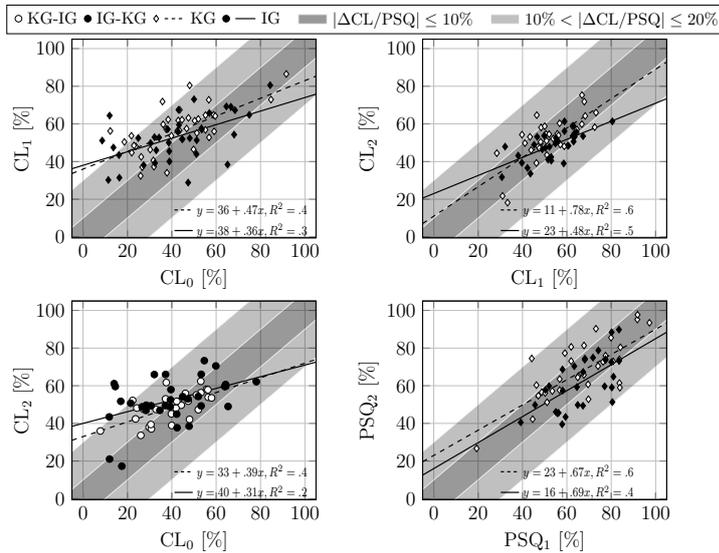
\*Items sind nicht Teil der optimierten Kern-Itemgruppe (welche Grundlage der ANCOVA ist).

**Tab. A-12:** Deskriptive Statistiken aus der Skalenbildung ausgeschlossener Motivations-Items

Item	Kurzformulierung	Kontrollgruppe					Interventionsgruppe				
		$t_0$		$t_1$		$\Delta$	$t_0$		$t_1$		$\Delta$
		$M$	$SD$	$M$	$SD$		$M$	$SD$	$M$	$SD$	
M01	Die ÜA sind verständlich	.76	.12	.67	.12	-.09	.77	.13	.71	.13	-.06
M02	Möchte mehr erfahren	.70	.24	.73	.21	.03	.70	.25	.74	.22	.04
M06	Beteiligung in den Übungen	.68	.20	.68	.19	.00	.61	.25	.67	.27	.06
M16	Lösungswege vorgeschrieben	.60	.22	.61	.18	.01	.64	.23	.56	.20	-.08
M19	Beschäftigung über ÜA	.58	.16	.49	.24	-.09	.60	.22	.61	.20	.01
M21	Diskussion über ÜA	.83	.22	.82	.23	-.01	.83	.23	.85	.21	.02
M23	Selbsttätigkeit	.69	.16	.61	.18	-.08	.66	.18	.67	.18	.01
M24	Interessante Themen	.46	.20	.45	.22	-.01	.49	.23	.53	.22	.03
M27	Bearbeitung aller Aufg.	.93	.14	.83	.16	-.09	.96	.11	.86	.19	-.10
M37	Beschäftigung unabh. von Prüfungsanforderungen	.70	.15	.56	.21	-.14	.72	.15	.65	.17	-.08
M40	Form der ÜA bedeutsam	.68	.17	.63	.19	-.05	.70	.18	.67	.18	-.03
M43	Konzentriert in Übungen	.75	.16	.68	.14	-.07	.78	.17	.70	.18	-.08
M45	Persönliche Bedeutung	.62	.21	.53	.22	-.10	.60	.23	.63	.18	.03

Bemerkung:  $\Delta$  = Differenzvariable zwischen Prä-Zeitpunkt  $t_0$  und Post-Zeitpunkt  $t_1$

Streudiagramme:



**Abb. A-4:** Ergänzende Streudiagramme der Belastungsvariablen (CL und PSQ)

zu Gender-Effekten:

**Tab. A-13:** Geschlechtsbezogene deskriptive Daten ausgewählter Variablen

	$\sigma(N = 56)$		$\varphi(N = 20)$	
	$M$	$SD$	$M$	$SD$
Abinote	2.2	.7	2.0	.7
Ma-Vorleistung	11.5	2.2	12.4	.9
Phy-Vorleistung	11.4	3.1	12.3	2.1
IE <sub>0</sub>	.74	.11	.77	.11
SK <sub>0</sub>	.61	.12	.63	.08
KiRC <sub>0</sub>	.71	.20	.57	.24

## D Berechnungen zur Testmischungsanalyse

### Invarianz des Hake-Index unter Ratekorrektur

Sei  $L : T \mapsto L(T)$  eine von der Identität verschiedene lineare Transformation der Testsumme  $T$ , d.h.  $L(T) = aT + b$  mit  $a \neq 1 \wedge b \neq 0$ . Dann ist der Hake-Index (aufgefasst als Funktion zweier Testsummen  $T_1$  und  $T_2$  mit maximaler Testsumme  $n > 1$ )

$$g_H(T_1, T_2) = \frac{T_2 - T_1}{n - T_1} \quad (50)$$

genau dann invariant unter  $L$ , wenn  $a + b/n = 1$  gilt.

*Beweis:*

Invarianz unter  $L$  bedeutet, dass gilt

$$g_H(L(T_1), L(T_2)) = g_H(T_1, T_2), \quad (51)$$

also nach der Definition von  $L$  und Gl. (50):

$$\frac{aT_2 + b - (aT_1 + b)}{n - aT_1 - b} = \frac{T_2 - T_1}{n - T_1}. \quad (52)$$

Dies ist aber äquivalent zu

$$na(T_2 - T_1) = n(T_2 - T_1) - b(T_2 - T_1) \quad (53)$$

$$\Leftrightarrow a + \frac{b}{n} = 1 \quad (T_2 \neq T_1) \quad \square \quad (54)$$

Für  $T_1 = T_2$  gilt Gl. (52) offensichtlich. Die Ratekorrektur von  $T$  nach Gl. (32),

$$L(T) = T' = \left(T - \frac{n}{q}\right) \cdot \frac{q}{q-1} = \frac{q}{q-1}T - \frac{n}{q-1}, \quad (55)$$

gehört zu der Klasse von linearen Funktionen, die die Bedingung  $a + b/n = 1$  erfüllt, denn

$$\frac{q}{q-1} - \frac{n}{q-1} \cdot \frac{1}{n} = 1. \quad (56)$$

### Reliabilität bei variabler Itemgewichtung

Nach Gl. (6) wird die Reliabilität eines Tests mit  $n$  Items gemäß

$$\alpha_c = \frac{n}{n-1} \left(1 - \frac{\sum_j SD_j^2}{SD^2}\right) \quad (57)$$

berechnet. Die Antworten auf die Items, die zur zweiten KiRC-Testhälfte gehören, werden nun mit dem Faktor  $\kappa$  gewichtet. Die Summe im Zähler kann dann in zwei Summen über die beiden Itemgruppen (MC und MTF) aufgespalten werden:

$$\sum_j SD_j^2 = \sum_{MC} SD_j^2 + \kappa^2 \sum_{MTF} SD_j^2. \quad (58)$$

Aus der Varianz der Testsummenwerte  $SD^2$  wird entsprechend zur Herleitung von Gl. (46)

$$SD^2(\kappa) = SD_1^2 + \kappa^2 SD_2^2 + 2\kappa\rho(T_1, T_2). \quad (59)$$

Die Indizes 1 und 2 stehen für die Testhälften 1 und 2 und  $\rho(T_1, T_2)$  bezeichnet die Korrelation zwischen den beiden Testsummen. Ohne die Terme weiter zu vereinfachen hat  $\alpha_c(\kappa)$  damit folgende strukturelle Gestalt ( $c_i > 0$ ):

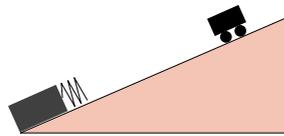
$$\alpha_c(\kappa) = \frac{n}{n-1} \left(1 - \frac{c_1 + \kappa^2 c_2}{c_3 + \kappa c_4 + \kappa^2 c_5}\right) \quad (60)$$

## E Instruktionmaterialien

VA-Aufgaben sind Videoanalyse-Aufgaben, deren Inhalt mit den traditionellen (T)-Aufgaben der Vergleichsgruppen übereinstimmen.

### Aufgabe 2.1: *Bewegung mit zeitweise konstanter Beschleunigung (T-Aufgabe)*

Ein Wagen rollt eine schiefe Ebene hinab, wird elastisch an einer Feder reflektiert und rollt die Ebene wieder hinauf. Reibungsverluste sollen vernachlässigt werden.



- Skizzieren Sie qualitativ den Verlauf von Ort  $x(t)$ , Geschwindigkeit  $v(t)$  und Beschleunigung  $a(t)$  des Wagens in einem gemeinsamen Diagramm!
- Betrachtet werde nur die Bewegung unter alleinigem Einfluss der Schwerkraft. Erklären Sie das Vorzeichen der Beschleunigung in den beiden Bewegungsphasen (hangaufwärts und hangabwärts)!
- Beschreiben Sie eine experimentelle Möglichkeit, wie die Beschleunigung des Wagens gemessen werden kann!

(9 Punkte)

### Aufgabe 2.1: *Bewegung mit zeitweise konstanter Beschleunigung (V-Aufgabe)*

Im Video gleitet ein Wagen annähernd reibungsfrei eine schiefe Luftkissenfahrbahn hinab, wird elastisch an einer Feder reflektiert und gleitet die Luftkissenfahrbahn wieder hinauf.

- Skizzieren Sie qualitativ den Verlauf der Beträge von Ort  $x(t)$ , Geschwindigkeit  $v(t)$  und Beschleunigung  $a(t)$  des Wagens in einem gemeinsamen Diagramm! Überprüfen Sie die Graphen durch eine Videoanalyse der Bewegung!
- Betrachtet werde nur die Bewegung unter alleinigem Einfluss der Schwerkraft. Erklären Sie das Vorzeichen der Beschleunigung und warum diese während der Bewegung gleich bleibt!
- Ermitteln Sie aus den Videoanalysedaten möglichst genau die Beschleunigung  $a(t)$  des Wagens während der Abwärtsbewegung durch lineare Regression! Kontrollieren Sie das Ergebnis durch eine davon unabhängige Methode!

*Hinweis:* Legen Sie den Koordinatenursprung in den Mittelpunkt des blauen Kreises im untersten Bahnpunkt der Bewegung, die  $x$ -Achse in Hangaufwärtsrichtung und den Zeitpunkt  $t = 0$  in den Frame, in dem der Wagen beginnt, hangabwärts zu fahren! Wählen Sie als Bildschrittwerte 5!

(9 Punkte)

**Aufgabe 2.2:***Kreisbewegungen (T-Aufgabe)*

Ein an einem Faden befestigtes Objekt rotiert waagrecht im Uhrzeigersinn auf einer Kreisbahn mit Radius  $r = 30$  cm.

Für die Bahngeschwindigkeit gelte

$$v(t) = \begin{cases} 5 \text{ cm/s} & \text{für } t \in [0; 10 \text{ s}] \\ -15 \text{ cm/s} + (2 \text{ cm/s}^2) \cdot t & \text{für } t > 10 \text{ s} \end{cases}.$$

- Bestimmen Sie betragsmäßig die Winkelgeschwindigkeit  $\omega = \omega(t)$  und die Winkelbeschleunigung  $\alpha = \alpha(t) = \dot{\omega}(t)$ ! Welche Orientierung besitzt der Vektor  $\vec{\alpha}(t)$ ?
- Beschreiben Sie die Bewegung des Objekts durch einen Ortsvektor  $\vec{r}(t)$  aus einem kartesischen Koordinatensystem mit Ursprung im Kreismittelpunkt!
- Stellen Sie für jeden der beiden Bewegungsabschnitte die Beschleunigungsvektoren des Objekts während einer Umdrehung qualitativ graphisch dar! Fertigen Sie hierzu zwei getrennte Zeichnungen an! Nennen Sie Gemeinsamkeiten und Unterschiede!

(11 Punkte)

**Aufgabe 2.2:***Kreisbewegungen (V-Aufgabe)*

Video I: Das Video *Blatt02\_Kreis1* zeigt die Rotation einer Kreisscheibe:

- Weisen Sie mit Hilfe der vier Markierungspunkte die Gültigkeit des Zusammenhangs

$$v = \omega r$$

mit  $v$  als Bahngeschwindigkeit,  
 $\omega$  als Winkelgeschwindigkeit und  
 $r$  als Abstand zur Drehachse

experimentell nach! Tragen Sie dazu die zeitlich gemittelte Bahngeschwindigkeit gegen  $r$  auf! Bestimmen Sie  $\omega$ !

**(Hinweise:** Zur automatischen Bewegungsanalyse empfiehlt sich eine Schrittweite von 10. Nutzen Sie für die Analyse der Punkte jeweils eine eigene Tabelle!)

- Blenden Sie die Beschleunigungsvektoren der ersten Umdrehung für den äußersten Punkt graphisch im Videobild ein! Weshalb ist auch eine Kreisbewegung mit konstanter Geschwindigkeit eine *beschleunigte* Bewegung?

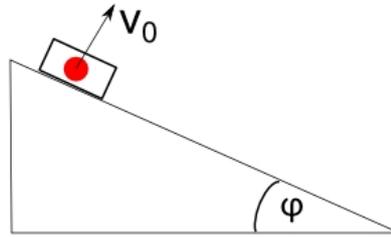
Video II: Im Video *Blatt02\_Kreis2* wird die Kreisscheibe über eine Umlenkrolle durch eine Masse gleichmäßig beschleunigt:

- Bestimmen Sie die Winkelbeschleunigung  $\dot{\omega}$ !
- Blenden Sie die Beschleunigungsvektoren der ersten Umdrehung graphisch im Videobild ein und vergleichen Sie das Ergebnis mit dem Ergebnis in (b)!

(11 Punkte)

**Aufgabe 3.1:***Schuss von schiefer Ebene (T-Aufgabe)*

Ein zunächst ruhender Wagen startet zum Zeitpunkt  $t = 0$  seine Fahrt auf einem Hang mit Neigungswinkel  $\varphi$ . Zeitgleich wird eine Kugel senkrecht zum Hang mit der Geschwindigkeit  $v_0$  geschossen. Vernachlässigen Sie Reibungseffekte!



- Zeigen Sie, dass die Kugel für beliebige Winkel  $0 \leq \varphi < 90^\circ$  wieder im Wagen landet und diskutieren Sie die Grenzfälle  $\varphi = 0^\circ$  und  $\varphi = 90^\circ$ !  
**Hinweis:** Mit einem geschickt gewählten Koordinatensystem vereinfachen sich die Rechnungen!
- Bleibt die Aussage aus (a) gültig, wenn der Wagen zum Zeitpunkt  $t = 0$  bereits eine Startgeschwindigkeit  $v_s$  in Richtung der Hangschräge besitzt? Diskutieren Sie in diesem Zusammenhang ebenso die Spezialfälle aus (a)!
- Wie könnte (mit möglichst einfachen Mitteln) prinzipiell eine experimentelle Prüfung des beschriebenen Szenarios umgesetzt werden?

(10 Punkte)

**Aufgabe 3.1:***Schuss von schiefer Ebene (V-Aufgabe)*

Das Video «Blatt 03 - Schuss von schiefer Ebene» zeigt, wie ein zunächst ruhender Wagen einen Hang mit Neigungswinkel  $\varphi$  hinunterrollt und bei Bewegungsbeginn ( $t = 0$ ) eine Kugel senkrecht zum Hang abschießt. Die Indizes K bzw. W stehen für Kugel und Wagen.

- Orientieren Sie das Koordinatensystem mit der  $x$ -Achse entlang des Hangs und der  $y$ -Achse senkrecht dazu! Nehmen Sie die Bewegung der Kugel und des Wagens auf und stellen Sie diese jeweils gemeinsam  $y(x)$ ,  $x(t)$  und  $y(t)$  in einem Diagramm dar!
- Zeigen Sie, dass die Beschleunigung  $a_x$  längs der  $x$ -Achse des Wagens konstant ist und finden Sie eine Formel für  $x_w(t)$ ! Messen Sie den Hangwinkel  $\varphi$  und vergleichen Sie  $a_x$  mit dem theoretisch zu erwartenden Wert!
- Zeigen Sie, dass es sich im Video nicht um einen Glücksschuss handelt, sondern dass die Kugel für beliebige Winkel  $0 \leq \varphi < 90^\circ$  wieder im Wagen landet und diskutieren Sie die Grenzfälle  $\varphi = 0^\circ$  und  $\varphi = 90^\circ$ !

(10 Punkte)

**Aufgabe 3.2:***Bungee-Springen (T-Aufgabe)*

Ein Bungee-Springer ( $m = 80 \text{ kg}$ ) springt von der Dammkrone des Verzasca-Staudamms (Schweiz, Dammhöhe 220 m) an einem Bungee-Seil (Länge  $l = 100 \text{ m}$ , Federkonstante  $D = 70 \text{ N/m}$ ) in die Tiefe. Masse und innere Reibung des Bungee-Seils sowie Luftreibung werden als vernachlässigbar angenommen.

- Für die Höhen-Ortskoordinate  $y$  des Bungee-Springers gelte  $y(t = 0) = 0$  beim Absprung und  $0 \leq y(t) \leq y_{\max}$ , wobei  $y_{\max}$  die Ortskoordinate des unteren Umkehrpunkts des Bungee-Springers ist ( $y$ -Achse nach unten!). Stellen Sie in einem Diagramm  $y(t)$ ,  $v(t)$  und  $a(t)$

qualitativ für die Abwärtsbewegung dar und begründen Sie die Verläufe!

Die potentielle Energie einer Feder ist

$$E_{\text{pot}} = \frac{1}{2}Dy^2$$

- b) Geben Sie die Teilkräfte und die resultierende Kraft auf den Bungee-Springer als Funktion von  $y$  an! Stellen Sie die Bewegungsgleichung auf (nicht lösen!)  
(**Hinweis:** Abschnittsweise definierte Funktionen: Seil ungespannt / Seil gespannt)
- c) Geben Sie ohne Rechnung, aber begründet, an, wann bzw. bei welcher Ortskoordinate  $y_0$  die maximale Geschwindigkeit  $v_{\text{max}}$  erreicht wird!  
Leiten Sie eine Formel für  $v(y)$  her! Berechnen Sie damit  $v_{\text{max}}$ ,  $y_0$  und  $y_{\text{max}}$  des Bungee-Springers!
- d) Welche Federkonstante  $D$  müsste das Bungee-Seil mindestens haben, damit der Bungee-Springer gerade noch nicht auf den Boden aufschlüge?  
Welche Kraft wirkt auf die Aufhängung des Bungee-Seils, wenn der fallende Bungee-Springer bei  $y = 50$  m und bei  $y = 130$  m ist?

(12 Punkte)

### Aufgabe 3.2:

### Bungee-Springen (V-Aufgabe)

Das Video « *Blatt 03 - Bungee-Springen* » zeigt in einem Modellexperiment die Fallbewegung eines Bungee-Springers (Kugel der Masse  $m = 133$  g) an einem Bungee-Feder-Seil (Länge  $l = 0,612$  m, Federkonstante  $D = 2,9$  N/m). Die Luftreibung und die Federmasse können vernachlässigt werden.

Die potentielle Energie einer Feder ist

$$E_{\text{pot}} = \frac{1}{2}Dy^2$$

- a) Legen Sie im Video eine skalierte  $y$ -Höhen-Achse mit  $y(t = 0) = 0$  bei Sprungbeginn und  $y(t) \geq 0$  an, wobei  $y_{\text{max}}$  die Ortskoordinate des unteren Umkehrpunkts des Bungee-Springers ist!  
Nehmen Sie das  $y(t)$ -Diagramm der Fallbewegung des Bungee-Springers mit Schrittweite 5 auf! Zeichnen Sie von Hand qualitativ und begründet  $v_y(t)$  und  $a_y(t)$  gemeinsam mit  $y(t)$  auf!
- b) Erzeugen Sie das  $a_y(y)$ -Diagramm! Erklären und bestätigen Sie den Verlauf der Messpunkte anhand von Funktionsvorschriften der Teilkräfte und der resultierenden Kraft auf den Bungee-Springer! Überprüfen Sie den angegebenen Wert der Federkonstanten!  
Stellen Sie die Bewegungsgleichung auf (nicht lösen!)  
(**Hinweis:** Abschnittsweise definierte Funktionen: Seil ungespannt / Seil gespannt)
- c) Geben Sie ohne Rechnung, aber begründet an, bei welcher Ortskoordinate  $y_0$  der Bungee-Springer die maximale Geschwindigkeit  $v_{\text{max}}$  erreicht.  
Leiten Sie eine Formel für  $v_y(y)$  her! Berechnen Sie damit  $v_{\text{max}}$ ,  $y_0$  und  $y_{\text{max}}$  des Bungee-Springers!  
Kontrollieren Sie die Ergebnisse für  $v_{\text{max}}$  und  $y_{\text{max}}$  experimentell (im Video)!
- d) Erzeugen Sie ein Diagramm, in dem die Teilenergien des Systems *Bungee-Springer mit Seil* über der Ortskoordinate  $y$  aufgetragen sind! Bleibt die Gesamtenergie erhalten?

(12 Punkte)

**Aufgabe 4.1:***Geschwindigkeit (T-Aufgabe)*

Eine Gewehr­kugel (Masse  $m = 0,53$  g) wird waagrecht auf eine Knetkugel der Punktmasse  $M = 108,3$  g, aufgehängt an einem masselosen Faden der Länge  $l = 0,5$  m, geschossen. Die Gewehr­kugel wird vollständig in der Knetkugel abgebremst und die Knetkugel in horizontaler Richtung um  $d = 10$  cm ausgelenkt. Die Zeitdauer des Abbremsens in der Knetkugelkugel ist klein gegenüber der Zeitdauer der Pendelauslenkung:

- a) Leiten Sie eine Formel zur Bestimmung der Geschwindigkeit  $v$  der Gewehr­kugel her und berechnen Sie sie!
- b) Welche Geschwindigkeit müsste die Gewehr­kugel für eine volle Umdrehung des Pendels mindestens haben?

(9 Punkte)

**Aufgabe 4.1:***Geschwindigkeit (V-Aufgabe)*

Im Video wird eine abgefeuerte Gewehr­kugel (Masse  $m = 0,53$  g) in einer Pendelmasse (Masse  $M = 108,3$  g) vollständig abgebremst. Die Zeitdauer des Abbremsens in der Pendelmasse ist klein gegenüber der Zeitdauer der Pendelauslenkung:

- a) Welche Größen können dem Video entnommen werden, um die Geschwindigkeit der Gewehr­kugel möglichst genau zu bestimmen? Leiten Sie die Formeln her und berechnen Sie die Geschwindigkeit!
- b) Welche Geschwindigkeit müsste die Gewehr­kugel für eine volle Umdrehung des Pendels mindestens haben?

(9 Punkte)

**Aufgabe 4.2:***Massenänderungen während der Fahrt (T-Aufgabe)*

Ein Schotterwaggon der Leermasse  $m = 1000$  kg bewegt sich mit konstanter Geschwindigkeit  $v_0 = 20$  m/s in  $x$ -Richtung. Zum Zeitpunkt  $t_0 = 10$  s wird er von oben schlagartig mit Schotter der Masse  $\Delta m = 500$  kg befüllt. Dabei unterscheiden wir zwei Szenarien:

- (i) der Schotter fällt aus einem fest stehenden Depot auf den Schotterwaggon,
- (ii) der Schotter fällt aus einem mit gleicher Geschwindigkeit  $v_0$  bewegten Depot auf den Schotterwaggon!

- a) Leiten Sie für beide Fälle eine Formel für die Geschwindigkeit  $v$  des Schotterwaggons nach dem Füllvorgang, d.h. für  $t > t_0$ , her! Zeichnen Sie das Geschwindigkeits-Zeit-Diagramm für beide Fälle!
- b) Im Szenario (ii) verliert das Depot schlagartig an Masse. Wie stellt sich formelmäßig nach dem Schüttvorgang die Geschwindigkeit des Depots unter der Annahme dar, dass (außer der Gewichtskraft) keine äußeren Kräfte an dem Depot angreifen?
- c) Nehmen Sie an, das Depot verliert im Fall (i) den Schotter nicht schlagartig, sondern gemäß der Ausflussrate

$$\lambda = 100 \text{ kg/s !}$$

Leiten Sie eine Formel für die Geschwindigkeit  $v(t)$  des Schotterwaggons her ( $t > t_0$ ) !

(12 Punkte)

**Aufgabe 4.2:***Massenänderungen während der Fahrt (mV-Aufgabe)*

Ein Objekt bewege sich mit konstanter Geschwindigkeit in  $x$ -Richtung, während sich seine Masse  $m$  zum Zeitpunkt  $t = t_0$  schlagartig ändert (neue Masse  $m^* = m + \Delta m$ ). Die Massenänderung  $\Delta m$  wird senkrecht zum Geschwindigkeitsvektor des Objekts herbeigeführt. In dieser Aufgabe stellen Sie dieses Szenario nach und werten das zugehörige Experiment aus!

- a) Formulieren Sie zunächst begründete Hypothesen, ob und wie sich die Geschwindigkeit des Objekts für  $t > t_0$  ändert! Unterscheiden Sie dabei zwei Fälle:
  - i) die Masse nimmt schlagartig zu ( $\Delta m > 0$ ),
  - ii) die Masse nimmt schlagartig ab ( $\Delta m < 0$ ) !
- b) Videographieren Sie das oben beschriebene Experiment! **Anregung:** Das sich bewegende Objekt könnte ein Skateboard oder ein Spielzeugwagen sein. An Schnüren befestigte Gewichte können die Massenänderung bewirken.
- c) Erstellen Sie das Geschwindigkeits-Zeit-Diagramm der beiden Bewegungen (i) und (ii) ihres Objekts und teilen Sie es dafür in Bewegungsabschnitte ein! Zeigen Sie zunächst, dass die Geschwindigkeit für  $t < t_0$  konstant ist! Bestätigt das Experiment ihre Vermutungen aus (a)?
- d) Begründen Sie, dass es für einen anderen Fall möglich ist, dass die Geschwindigkeit des Objekts trotz Massenzunahme unverändert bleibt! Videographieren Sie das zugehörige Experiment und erstellen Sie das Geschwindigkeits-Zeit-Diagramm! Falls Sie trotz richtiger Annahmen eine Geschwindigkeitsänderung messen, begründen Sie, warum!

(12 Punkte)

**Aufgabe 5.1:***Rollende Zylinder (T-Aufgabe)*

- a) Zeigen Sie, dass ganz allgemein für das Trägheitsmoment  $I$  von Voll- und Hohlzylinder (VZ/HZ) mit dem Zylinderaußenradius  $R$ , der Wandstärke  $d$  und der Masse  $m$  bezüglich der Zylinderachse gilt:

$$I_{VZ} = 0,5mR^2$$

$$I_{HZ} = m(R^2 - Rd + d^2/2) !$$

- b) Bestimmen Sie das Verhältnis  $d/R$ , so dass die Näherungsformel  $I_{HZ} \approx mR^2$  mit einem Fehler  $< 1\%$  angewendet werden kann!

Drei Hohlzylinder (Außenradius  $R$ , Wandstärke  $d$ , Höhe  $h$ ) 1-3 rollen ohne zu gleiten vom gleichen Startpunkt einen Hang mit Neigungswinkel  $\alpha$  hinunter. Zylinder 1 ist leer, Zylinder 2 ist vollständig wassergefüllt (keine Reibung mit Zylinderwand) und Zylinder 3 ist vollständig gefüllt mit Eis (fest an Zylinderwand). In den Fällen 2 und 3 soll die Zylinderwand die Masse  $m_1$  haben. Die Massen des Wassers und des Eises werden als gleich angenommen, nämlich  $m_2$ . Im Fall 1 des leeren Zylinders wird von einer Masse  $m_1 + m_2$  der Zylinderwand ausgegangen.

- c) Leiten Sie die Schwerpunktbeschleunigungen der drei Körper her und ermitteln Sie die Reihenfolge des Eintreffens der Zylinder am Hang-Ende!
- d) Das Experiment wird mit einem vierten Zylinder (form- und materialgleich wie im Fall 2), gefüllt mit einem kohlenensäurehaltigen Getränk, wiederholt! Vor dem Versuch wird der Zylinder stark geschüttelt! Formulieren Sie eine **begründete** Vermutung über den Betrag der Schwerpunktbeschleunigung (im Vergleich zu den anderen Zylindern)!

(16 Punkte)

**Aufgabe 5.1:***Rollende Zylinder (mV-Aufgabe)*

Verschiedene formgleiche Aluminiumzylinder rollen ohne zu gleiten vom gleichen Startpunkt einen Hang mit Neigungswinkel  $\alpha$  hinunter. Zylinder 1 ist leer, Zylinder 2 ist vollständig flüssigkeitsgefüllt und Zylinder 3 ist vollständig gefüllt mit Eis. Allgemein gilt für die Trägheitsmomente  $I$  von Voll- und Hohlzylinder (VZ/HZ) mit dem Zylinderaußenradius  $R$ , der Wandstärke  $d$  und der Masse  $m$  bezüglich der Zylinderachse:

$$I_{VZ} = 0,5mR^2$$

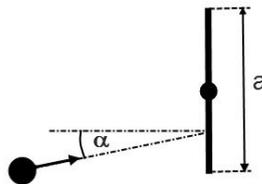
$$I_{HZ} = m(R^2 - Rd + d^2/2).$$

- Führen Sie das Experiment mit geeigneten Alltagsmaterialien durch! Dokumentieren und erklären Sie qualitativ die Reihenfolge des Eintreffens der Zylinder am Hang-Ende (mit Video-Beweis)!
- Leiten Sie eine Formel für die Schwerpunktbeschleunigung  $a$  der Zylinder her und berechnen Sie sie! Schätzen Sie dazu die Massen der Zylinder / des Inhalts ab! Vergleichen Sie die experimentell gewonnenen Schwerpunktbeschleunigungen mit den theoretischen Daten!
- Führen Sie das Experiment mit einer geschüttelten Getränkedose ( $\text{CO}_2$ -haltiges Getränk) und einer ungeschüttelten Getränkedose durch! Messen Sie Unterschiede? Bilden Sie eine Hypothese zur Erklärung der Beobachtung! *Hinweis:* Um möglichst glaubwürdige Daten zu erhalten, empfiehlt sich eine Messwiederholung!
- Skizzieren Sie qualitativ die Geschwindigkeits- und Beschleunigungsvektoren eines äußeren Zylinderpunktes auf halber Rollstrecke!

(16 Punkte)

**Aufgabe 5.2:***Schuss auf Platte (T-Aufgabe)*

Ein Knetklumpen (Masse  $m = 50$  g) trifft im Abstand  $r = 0,123$  m von der Drehachse unter dem Winkel  $\alpha = 8^\circ$  auf eine drehbar gelagerte rechteckförmige, homogene Platte (Kantenlänge  $a = 43,4$  cm und Höhe  $b = 9,4$  cm (senkrecht zur Zeichenebene), Dicke  $d = 8$  mm, Masse  $M = 917$  g) und bleibt dort haften. Die Platte wird dadurch in Rotation versetzt mit der Winkelgeschwindigkeit  $\omega = 2,79$  s $^{-1}$ .



- Leiten Sie das Trägheitsmoments der Platte bzgl. der Rotationsachse her! Berechnen Sie das Trägheitsmoment  $I_S$  des Systems *Platte mit Knetklumpen*! **Hinweis:** Nehmen Sie den Knetklumpen als Punktmasse an!
- Leiten Sie eine Formel zur Bestimmung der Geschwindigkeit  $v$  her, mit der der Knetklumpen auf die Platte trifft und berechnen Sie sie!

- c) Geben Sie begründete physikalische Antworten ohne Rechnungen: welche Größen verändern sich wie, wenn lediglich der Knetklumpen durch einen ideal-elastischen Gummiball ersetzt wird?

(12 Punkte)

**Aufgabe 5.2:**

*Schuss auf Platte (V-Aufgabe)*

Im Video «Schuss auf Platte» trifft ein Knetklumpen (Masse  $m = 50$  g) auf eine drehbar gelagerte, rechteckförmige und homogene Platte (Kantenlänge  $a = 43,4$  cm und Höhe  $b = 9,4$  cm, Dicke  $d = 8$  mm, Masse  $M = 917$  g). Das drehende Lager der Platte hat eine geringe Reibung und ein Trägheitsmoment (ohne Platte) von  $I_L = 1,85 \cdot 10^{-3}$  kg m<sup>2</sup>.

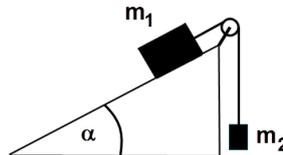
- a) Leiten Sie das Trägheitsmoments der Platte bzgl. der Rotationsachse her! (Beachten Sie das Bild im Video zum Versuchsaufbau!)  
Berechnen Sie das Trägheitsmoment  $I_S$  des Systems *Platte mit Knetklumpen und Drehlager* bezüglich der Rotationsachse!  
**Hinweis:** Nehmen Sie den Knetklumpen als Punktmasse an!
- b) Leiten Sie eine Formel zur Bestimmung der Geschwindigkeit  $v$ , mit der der Knetklumpen auf die Platte trifft, aus *gegebenen* und *im Video messbaren* Größen her! Berechnen Sie die Geschwindigkeit  $v$  mit dieser Formel!  
**Hinweis:** Bestimmen Sie die Winkelgeschwindigkeit  $\omega$  kurz nach dem Stoß!
- c) Kontrollieren Sie das Ergebnis aus (b) durch eine direkte Geschwindigkeitsmessung!

(12 Punkte)

**Aufgabe 6.1:**

*Haften und Gleiten am Hang (T-Aufgabe)*

Zwei durch ein Seil verbundene Massen  $m_1 = 345,4$  g (auf dem Hang) und  $m_2 = 110$  g (hängend) ruhen für Hangwinkel  $0 \leq \alpha < 42^\circ$ . Beim Hangwinkel  $\alpha_G = 42^\circ$  gleitet die Masse  $m_1$  in 1 s gleichmäßig beschleunigt eine Strecke von 48 cm den Hang hinunter. Die Luftreibung sowie Rollreibung und Masse der Umlenkrolle sind vernachlässigbar.



- a) Weshalb setzen sich die Massen erst bei einem Grenzwinkel  $\alpha_G > 0$  in Bewegung? Warum liegt für  $\alpha = \alpha_G$  eine gleichmäßig beschleunigte Bewegung vor?
- b) Leiten Sie eine Formel zur Bestimmung des Haftreibungskoeffizienten  $\mu_H$  her und berechnen Sie  $\mu_H$ ! Skizzieren Sie alle relevanten Kraftvektoren in einem Diagramm!
- c) Leiten Sie eine Formel zur Bestimmung des Gleitreibungskoeffizienten  $\mu_G$  her und überprüfen Sie die Formel an einem Spezialfall! Berechnen Sie  $\mu_G$ ! Skizzieren Sie alle relevanten Kraftvektoren in einem Diagramm!

(13 Punkte)

**Aufgabe 6.1:***Haften und Gleiten am Hang (V-Aufgabe)*

Das Video *Haften und Gleiten am Hang* zeigt das Verhalten zweier durch ein Seil verbundener Massen  $m_1 = 345,4$  g (auf dem Hang) und  $m_2 = 110$  g (hängend) bei Erhöhung des Hangwinkels  $\alpha$ . Die Luftreibung sowie Rollreibung und Masse der Umlenkrolle sind vernachlässigbar.

- Weshalb setzen sich die Massen erst bei einem Grenzwinkel  $\alpha_G > 0$  in Bewegung? Warum liegt für  $\alpha = \alpha_G$  eine gleichmäßig beschleunigte Bewegung vor? Bestimmen Sie experimentell die Beschleunigung  $a$ !
- Leiten Sie eine Formel zur Bestimmung des Haftreibungskoeffizienten  $\mu_H$  her! Leiten Sie eine Formel zur Bestimmung des Gleitreibungskoeffizienten  $\mu_G$  her und überprüfen Sie die Formel an einem Spezialfall! Skizzieren Sie alle relevanten Kraftvektoren in einem Diagramm!
- Mit Hilfe welcher gegebenen und im Video messbarer Größen lassen sich Haftreibungskoeffizient  $\mu_H$  bzw. Gleitreibungskoeffizient  $\mu_G$  experimentell bestimmen? Bestimmen Sie  $\mu_H$  und  $\mu_G$ !

(13 Punkte)

**Aufgabe 6.2:***Balkenbiegung (T-Aufgabe)*

Ein Aluminium-Vollbalken 1 mit quadratischem Querschnitt ( $b = 8$  mm) und ein Aluminium-Balken 2 mit Doppel-T-förmigem Querschnitt ( $b = 12$  mm,  $d = 2$  mm, siehe Abbildung 1) werden jeweils durch eine im Abstand  $x = l = 51,8$  cm von der Einspannstelle ( $x = 0$ ) in negativer  $z$ -Richtung angreifenden Kraft  $F$  um die Strecke  $s \sim F$  (Projektion auf die Vertikale) nach unten verbogen.

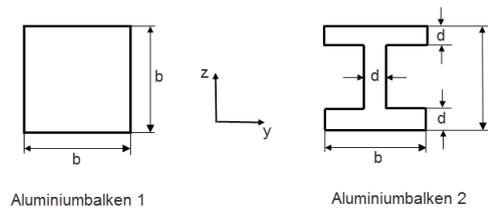


Abbildung 1: Querschnitte der betrachteten Aluminium-Balken

- Leiten Sie Formeln für die Biegemomente  $\Theta_F(b)$  bzw.  $\Theta_F(b, d)$  der beiden Balken her und berechnen Sie deren Werte. Weshalb werden in der Bautechnik Doppel-T-Träger eingesetzt? Hinweis:

$$\Theta_F = \int_A z^2 dA,$$

wobei die  $z$ -Richtung die Richtung der wirkenden Kraft ist.

- Beim Aluminium-Vollbalken 1 ist  $s(20 \text{ N}) = 4,0$  cm. Bestimmen Sie daraus den Elastizitätsmodul  $E$  und geben Sie die Funktionsgleichung  $z(x)$  der Biegelinie für  $F = 20$  N an.
- Aluminium-Balken 2 werde um  $90^\circ$  in der Querschnittsebene gedreht eingespannt. Zeigen Sie quantitativ und erklären Sie qualitativ, dass der gedrehte Aluminium-Balken sich stärker verbiegt.

(13 Punkte)

**Aufgabe 6.2:***Balkenbiegung (VA-Aufgabe)*

Ein Aluminium-Vollbalken 1 mit quadratischem Querschnitt ( $b = 8 \text{ mm}$ ) und ein Aluminium-Balken 2 mit Doppel-T-förmigem Querschnitt ( $b = 12 \text{ mm}$ ,  $d = 2 \text{ mm}$ , siehe Abbildung 1) werden jeweils durch eine im Abstand  $x = l = 51,8 \text{ cm}$  von der Einspannstelle ( $x = 0$ ) in negativer  $z$ -Richtung angreifenden Kraft  $F$  um die Strecke  $s \sim F$  (Projektion auf die Vertikale) nach unten verbogen.

- Nehmen Sie im Video *Balkenbiegung* den experimentellen Zusammenhang zwischen Strecke  $s$  und Kraft  $F$  beim Aluminium-Vollbalken 1 auf! Bestimmen Sie daraus die Federkonstante  $D$  des Balkens und den Elastizitätsmodul  $E$ !
- Nehmen Sie im Bild *Balkenbiegung - Aluminiumbalkens 1* die Biegelinie  $z(x)$  auf! Überprüfen Sie die Übereinstimmung von Theorie und Experiment!
- Leiten Sie Formeln für die Biegemomente  $\Theta_F(b)$  bzw.  $\Theta_F(b, d)$  der beiden Balken her und berechnen Sie deren Werte! Weshalb werden in der Bautechnik Doppel-T-Träger eingesetzt? Hinweis:

$$\Theta_F = \int_A z^2 dA,$$

wobei die  $z$ -Richtung die Richtung der wirkenden Kraft ist.

- Das Bild *Balkenbiegung - Aluminiumbalken 2* zeigt im Vergleich die Verbiegung des Doppel-T-Trägers und des um  $90^\circ$  in der Querschnittsebene gedreht eingespannten Doppel-T-Trägers bei gleicher Belastung. Erklären Sie qualitativ das Versuchsergebnis!

(13 Punkte)

#### Aufgabe 7.1:

*Beute über Bord (T-Aufgabe)*

Ein Räuber ( $m_R = 80 \text{ kg}$ ) versenkt sein Diebesgut ( $m_G = 90 \text{ kg}$ ,  $\rho_G = 1,1 \text{ g/cm}^3$ ) aus einem Boot (Masse  $m_B = 200 \text{ kg}$ ) in einem quaderförmigen See (Seefläche  $A = 1 \text{ km}^2$ ) konstanter Wassermenge ( $\rho_W = 1 \text{ g/cm}^3$ ). Das Diebesgut ist vor dem Eintauchen ins Wasser frei in der Luft, d.h. wird über Bord geworfen.

- Geben Sie zunächst eine qualitative Erklärung, ob der Wasserspiegel des Sees zwischen den Zuständen *Diebesgut im Boot* und *Diebesgut im Wasser* sinkt, steigt oder gleich bleibt!
- Leiten Sie Formel(n) für die Höhenänderung(en) des Wasserspiegels zwischen den Zuständen *Diebesgut im Boot*, *Diebesgut in der Luft* und *Diebesgut im Wasser* her! Geben Sie dazu eindeutige Bezeichnungen eingeführter Größen an! Berechnen Sie die Höhenänderungen!
- Kann durch Überbordwerfen von Gegenständen der Wasserspiegel eines Sees erhöht werden? Schreiben Sie strukturiert Überlegungen dazu auf!

(12 Punkte)

#### Aufgabe 7.1:

*Beute über Bord (VA-Aufgabe)*

Das Video *Beute über Bord* zeigt in einem Modellexperiment, wie ein Räuber (dargestellt durch Eisenplatte mit Masse  $m_R = 0,788 \text{ kg}$ ) sein Diebesgut (Eisenwägestück mit Masse  $m_G = 1 \text{ kg}$ , Dichte  $\rho_{Fe} = 7,87 \text{ g/cm}^3$ ) aus einem Boot (Plastikbox mit Masse  $m_B = 0,181 \text{ kg}$ ) in einem quaderförmigen See (Plastikbehälter mit Seefläche  $A = 38 \text{ cm} \times 17 \text{ cm}$ , die «38 cm» zur Videokalibrierung verwenden) konstanter Wassermenge (Dichte  $\rho_W = 1 \text{ g/cm}^3$ ) versenkt. Mit  $h_0$ ,  $h_1$  und  $h_2$  sind die vom Grund des Sees gemessenen Höhen des Wasserspiegels zu den Zeitpunkten  $t_0$  (Räuber, Diebesgut und Boot sind außerhalb des Sees),  $t_1$  (Räuber und Diebesgut im Boot auf dem See) und  $t_2$  (Räuber im Boot auf dem See, Diebesgut im See) bezeichnet:

- Messen Sie  $h_0$ ,  $h_1$  und  $h_2$  ! Erklären Sie qualitativ die Höhenänderung ( $h_2 - h_1$ ) des Wasserspiegels beim Ins-Wasser-Werfen des Diebesguts!
- Leiten Sie eine Formel für die Höhenänderung ( $h_2 - h_1$ ) des Wasserspiegels her! Geben Sie dazu eindeutige Bezeichnungen eingeführter Größen an! Vergleichen Sie die Ergebnisse mit den Messungen aus (a)!
- Kann durch Überbordwerfen von Gegenständen der Wasserspiegel eines Sees erhöht werden? Schreiben Sie strukturiert Überlegungen dazu auf!

(12 Punkte)

**Aufgabe 7.2:***Oberflächenspannung - Ringmethode (T-Aufgabe)*

Ein Aluminium-Ring (Dichte  $\rho = 2,7 \text{ g/cm}^3$ , Höhe  $h = 10 \text{ mm}$ , innerer Durchmesser  $d_1 = 50 \text{ mm}$ , äußerer Durchmesser  $d_2 = 52 \text{ mm}$ ) wird in horizontaler Lage vollständig in Wasser (Oberflächenspannung  $\sigma = 72,5 \text{ mN/m}$ ) eingetaucht und anschließend herausgezogen. Bevor sich eine dünne Lamelle an der Ringunterkante bildet, werden auch die Seitenflächen des Rings benetzt, siehe Abbildung 1. Der Ring sei an einem Kraftmesser aufgehängt, der die resultierende Kraft  $F_{\text{res}}$  zu jeder Zeit  $t$  misst.

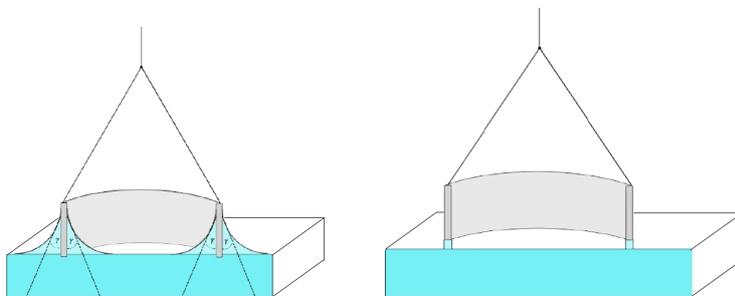


Abbildung 2: In Wasser getauchter Ring - Ansicht im Querschnitt

- Skizzieren Sie qualitativ einen möglichen Kraftverlauf  $F_{\text{res}}$  und teilen Sie diesen begründet in Abschnitte ein!
- Leiten Sie eine Formel zur Berechnung der Kraft  $F_{\sigma}$  her, die auf Grund der Oberflächenspannung an dem Ring angreift. Welche Kraft muss in vertikaler Richtung an dem Ring wirken, dass die Lamelle reißt?

(10 Punkte)

**Aufgabe 7.2:***Oberflächenspannung - Ringmethode (V-Aufgabe)*

Das Video *Ringmethode* zeigt einen Metallring (Masse  $m$ , Durchmesser  $d = 6,5 \text{ cm}$ ), der in Wasser getaucht und herausgezogen wird. Ein Kraftmesser misst zeitsynchron die resultierende Kraft  $F_{\text{res}}(t)$ , die am Ring zum Zeitpunkt  $t$  angreift.

- Skizzieren Sie den angezeigten Kraftverlauf und teilen Sie diesen begründet in **sechs** Abschnitte ein!  
*Hinweis:* Nutzen Sie zur Begründung unter anderem folgende Begriffe: Benetzung, Auftrieb, seitliche Ringfläche, Unterkante des Rings, Lamelle.

- b) Zeigen Sie, dass in diesem Versuch die Oberflächenspannung gemäß der Formel

$$\sigma = \frac{F_{\text{res}} - mg}{2\pi d}$$

bestimmt werden kann! Für welche Phase des Versuchs ist die Formel zur Berechnung der Oberflächenspannung  $\sigma$  gültig? Bestimmen Sie  $\sigma$ !

(10 Punkte)

**Aufgabe 8.1:**

*Magnus-Effekt (T-Aufgabe)*

Eine mit der Winkelgeschwindigkeit  $\vec{\omega} = \omega \vec{e}_z$  rotierende Kugel (Radius  $r$ , Dichte  $\rho_K$ ) fällt senkrecht (in  $y$ -Richtung) in ein Wasserbad (Dichte  $\rho_W$ ,  $x$ -Achse entlang der Wasseroberfläche). Zum Zeitpunkt  $t = 0$  des Eintauchens sei  $y(x = 0) = 0$ . Für die im Wasser beschriebene Bahnkurve der Kugel gilt  $y(x) \geq 0$ :

- a) Skizzieren Sie qualitativ die entstehende Bahnkurve  $y(x)$  der Kugel unter der Annahme, dass die Kugel mit konstanter Winkelgeschwindigkeit  $\omega$  rotiert! Nennen Sie alle auf die Kugel wirksamen Kräfte und zeichnen Sie diese in die Skizze ein!
- b) Zeigen Sie mit Hilfe der Bernoulli-Gleichung unter zu nennenden Idealbedingungen, dass für eine Kugel mit der Bahngeschwindigkeit  $v$  der Betrag  $F_M$  der Magnus-Kraft gegeben ist durch

$$F_M = 2\pi\rho_W r^3 v \omega \quad !$$

- c) Wie ändert sich die Bahnkurve  $y(x)$  aus (a) qualitativ, wenn die Winkelgeschwindigkeit  $\omega$  mit der Zeit abnimmt?

(10 Punkte)

**Aufgabe 8.1:**

*Magnus-Effekt (V-Aufgabe)*

Zwei Videos *Magnus-Kraft 1* und *Magnus-Kraft 2* zeigen, wie eine mit der Winkelgeschwindigkeit  $\omega$  rotierende Kugel und eine rotationsfreie Kugel (für beide gilt: Radius  $r$ , Dichte  $\rho_K$ ) senkrecht in ein Wasserbad (Dichte  $\rho_W$ ) eintauchen. Zum Zeitpunkt  $t = 0$  des Eintauchens sei  $y_i(x = 0) = 0$ ,  $i = 1, 2$  (Index 1 = rotierend, Index 2 = nicht rotierend). Für die im Wasser beschriebenen Bahnkurven der Kugeln gilt  $y_i(x) \geq 0$ ,  $i = 1, 2$ :

- a) Nehmen Sie je ein Stroboskopbild (Schrittweite 3) der beiden Bewegungen auf! Erklären Sie qualitativ die Bewegung der rotierenden Kugel! Zeichnen Sie in ein Bild kurz nach dem Eintauchen die Kraftpfeile aller Kräfte von Hand ein!
- b) Zeigen Sie mit Hilfe der Bernoulli-Gleichung unter zu nennenden Idealbedingungen, dass für eine Kugel mit der Bahngeschwindigkeit  $v$  die Magnus-Kraft  $F_M$  gegeben ist durch:

$$F_M = 2\pi\rho_W r^3 v \omega \quad !$$

- c) Rotiert die Kugel im Video im oder entgegen dem Uhrzeigersinn? Wie ändert sich die Bahnkurve  $y(x)$  qualitativ, wenn die Winkelgeschwindigkeit  $\omega$  mit der Zeit abnimmt?

(10 Punkte)

**Aufgabe 8.2:**

*Fallbewegungen unter Reibungseinfluss (T-Aufgabe)*

**Vorbemerkungen:**

Fällt ein Körper in positiver  $y$ -Richtung durch ein Fluid, z.B. eine Kugel durch Luft oder Öl, verursacht das Vorbeiströmen des Mediums am Körper eine Reibungskraft  $\vec{F}_R$ , die den Körper abbremst und der Bewegungsrichtung des Körpers entgegen gerichtet ist. Dies führt schließlich zu einer Endgeschwindigkeit  $\vec{v} = \vec{v}_E$  des Körpers (stationärer Zustand).

Bei kleinen Geschwindigkeiten ist diese Reibungskraft proportional zur Fallgeschwindigkeit (Stokesche Reibung):

$$\vec{F}_R = -\beta\vec{v} \quad (1)$$

mit  $\beta = 6\pi r\eta$ ,

wobei für eine Kugel  $r$  dem Radius entspricht und  $\eta$  die (dynamische) Viskosität des Mediums bezeichnet.

Wenn höhere Geschwindigkeiten erreicht werden (können), hängt die Reibungskraft quadratisch von der Fallgeschwindigkeit ab (Newtonsche Reibung):

$$\vec{F}_R = -\gamma v^2 \vec{e}_v \quad (2)$$

mit  $\gamma = \frac{1}{2} c_w \rho A$ ,

wobei  $\rho$  die Dichte des Mediums bezeichnet.  $A$  ist die auf eine Ebene senkrecht zur Bewegungsrichtung projizierte Querschnittsfläche des Körpers und  $c_w$  der als konstant angenommene Widerstandsbeiwert - eine von der Geometrie des fallenden Körpers abhängige Größe, die experimentell bestimmt werden muss.

**In Luft fallende Kugeln:**

Eine Styroporkugel (Masse  $m = 27,8$  g, Radius  $r = 7,5$  cm, Widerstandsbeiwert  $c_w = 0,45$ ) fällt in Luft (Dichte  $\rho_L = 1,3$  kg/m<sup>3</sup>) aus der Höhe  $h = 20$  m unter dem Einfluss der Luftreibung (Newtonscher Reibungsansatz!). Die Auftriebskraft kann vernachlässigt werden:

- a) Leiten Sie für beide Reibungsansätze (siehe Vorbetrachtung!) eine Formel für den Betrag der Endgeschwindigkeit  $v_E$  her! Schätzen Sie  $v_E$  ab (fehlende Größen schlagen Sie nach!) und folgern Sie, dass der Newtonsche Reibungsansatz der sinnvollere ist.
- b) Für  $y(t)$  der fallenden Styroporkugel soll  $y(0) = 0$ ,  $y(t) \geq 0$  und  $v(0) = 0$  sein:  
Zeichnen Sie qualitativ die Graphen von  $y(t)$ ,  $v(t)$  und  $a(t)$  (Beschleunigung) in das gleiche Diagramm ein!  
Erklären Sie qualitativ den Verlauf von  $v(t)$ .  
Welchen Wert hat  $a(0)$  ?
- c) Stellen Sie die Differentialgleichung für die zeitabhängige Geschwindigkeit  $v(t)$  des fallenden Papierkegels auf! Vernachlässigen Sie die Auftriebskraft in Luft!  
Zeigen Sie, dass

$$v(t) = v_E \tanh\left(\frac{g}{v_E} t\right)$$

mit der Erdbeschleunigung  $g$  und der Endgeschwindigkeit  $v_E$  eine spezielle Lösung der Differentialgleichung ist und berechnen Sie die Höhe  $y(t)$  des Kegels durch Integration - mit  $y(0) = 0$  !

- d) Berechnen Sie  $t_A$ , wobei dieser Zeitpunkt durch

$$v(t_A) = \frac{e^2 - 1}{e^2 + 1} v_E = 0,76 v_E$$

definiert ist!

Zeigen Sie, dass für  $t \ll t_A$  die Bewegung der Kugel näherungsweise als freier Fall betrachtet werden kann!

Hinweise:

$$\int \tanh(bx) dx = \frac{1}{b} \ln(\cosh(bx)) + C$$

$$\frac{d}{dx} \tanh(bx) = b(1 - \tanh^2(bx)) \quad \text{mit } b = \text{const}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$e^x \approx 1 + x \quad \text{für } |x| \ll 1$$

$$v_E := \sqrt{\frac{mg}{\gamma}} \quad \text{mit der Masse } m \text{ des Kegels}$$

(14 Punkte)

### Aufgabe 8.2:

Fallbewegungen unter Reibungseinfluss (mV-Aufgabe)

#### Vorbemerkungen:

Fällt ein Körper durch ein Fluid, z.B. eine Kugel durch Luft oder Öl, verursacht das Vorbeiströmen des Mediums am Körper eine Reibungskraft  $\vec{F}_R$ , die den Körper abbremst und der Bewegungsrichtung des Körpers entgegen gerichtet ist. Dies führt schließlich zu einer Endgeschwindigkeit  $\vec{v} = \vec{v}_E$  des Körpers (stationärer Zustand).

Bei kleinen Geschwindigkeiten ist diese Reibungskraft proportional zur Fallgeschwindigkeit (Stokesche Reibung):

$$\vec{F}_R = -\beta \vec{v} \quad (3)$$

mit  $\beta = 6\pi r \eta$ ,

wobei für eine Kugel  $r$  dem Radius entspricht und  $\eta$  die (dynamische) Viskosität des Mediums bezeichnet.

Wenn höhere Geschwindigkeiten erreicht werden (können), hängt die Reibungskraft quadratisch von der Fallgeschwindigkeit ab (Newtonsche Reibung):

$$\vec{F}_R = -\gamma v^2 \vec{e}_v \quad (4)$$

mit  $\gamma = \frac{1}{2} c_w \rho A$ ,

wobei  $\rho$  die Dichte des Mediums bezeichnet.  $A$  ist die auf eine Ebene senkrecht zur Bewegungsrichtung projizierte Querschnittsfläche des Körpers und  $c_w$  der als konstant angenommene Widerstandsbeiwert - eine von der Geometrie des fallenden Körpers abhängige Größe, die experimentell bestimmt werden muss.

#### Papierkegel fällt in Luft:

- a) Leiten Sie eine Formel für den Betrag  $v_E$  der Endgeschwindigkeit des Fallkegels her, der unter Einfluss der

- i) Stokeschen Reibungskraft,
- ii) Newtonschen Reibungskraft

mit der Spitze nach unten in Luft fällt. (Messdaten finden Sie am Ende des Übungsblattes! In Luft können Sie den Auftrieb vernachlässigen!)

- b) (mVA) Lassen Sie einen Papierkegel mit der Spitze nach unten aus ausreichender Höhe fallen, um eine Endgeschwindigkeit beobachten zu können! Videographieren Sie das Experiment und bestimmen Sie die Endgeschwindigkeit  $v_E$  ihres Fallkegels aus dem  $v(t)$ -Diagramm! Vergleichen Sie das Ergebnis mit ihren Vorhersagen aus Aufgabe (a) und entscheiden Sie, ob der Newtonsche Reibungsansatz der geeignete theoretische Ansatz für das Experiment ist!

- c) Stellen Sie die Differentialgleichung für die zeitabhängige Geschwindigkeit  $v(t)$  des fallenden Papierkegels auf! Vernachlässigen Sie die Auftriebskraft in Luft!  
Zeigen Sie, dass

$$v(t) = v_E \tanh\left(\frac{g}{v_E} t\right)$$

mit der Erdbeschleunigung  $g$  und der Endgeschwindigkeit  $v_E$  eine spezielle Lösung der Differentialgleichung ist und berechnen Sie die Höhe  $y(t)$  des Kegels durch Integration - mit  $y(0) = 0$ !

*Hinweise:*

$$\int \tanh(bx) dx = \frac{1}{b} \ln(\cosh(bx)) + C$$

$$\frac{d}{dx} \tanh(bx) = b(1 - \tanh^2(bx)) \quad \text{mit } b = \text{const}$$

$$v_E := \sqrt{\frac{mg}{\gamma}} \quad \text{mit der Masse } m \text{ des Kegels}$$

- d) (VA) Bestimmen Sie den  $c_w$ -Wert des Papierkegels!

### Zu den Experimenten:

Dichte der Luft:  $\rho_L = 1,25 \text{ kg/m}^3$

Viskosität der Luft:  $\eta_L = 1,7 \cdot 10^{-6} \text{ kg/(ms)}$

Papierkegel: Radius der Grundfläche: 7 cm (können Sie aber beliebig ändern!)

Masse des Papierkegels: für  $r = 7 \text{ cm}$  gilt  $m = 1,7 \text{ g}$  (für Ihren eigenen Kegel entsprechend umrechnen!)

$c_w$ -Wert des Papierkegels zur Abschätzung:  $c_w \approx 0,7$

### Aufgabe 9.1:

*Wasserpendel (T-Aufgabe)*

Wasser in einem U-Rohr (Schenkelabstand  $d = 8,8 \text{ cm}$ , Füllhöhe  $h = 32,3 \text{ cm}$ ) wird durch Unterdruck in einem der Schenkel ausgelenkt und einfach losgelassen:

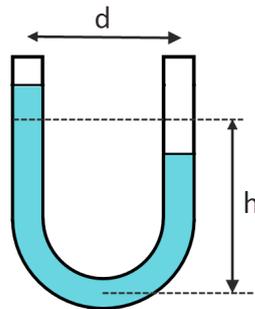


Abbildung 3: schematischer Versuchsaufbau - Wasser im U-Rohr ausgelenkt

- a) Erklären Sie mit Worten, weshalb eine harmonische Schwingung entsteht!

- b) Leiten Sie eine Formel für die Schwingungsdauer  $T_0$  des ungedämpften Wasserpendels her und berechnen Sie diese!
- c) Im Experiment wird eine gedämpfte Schwingung mit der Dämpfungskonstanten  $\gamma = 0,18 \text{ s}^{-1}$  beobachtet:  
Liegt die Dämpfungsverstimmung  $\frac{T-T_0}{T}$  der Schwingungsdauer unterhalb von 1% ?

(10 Punkte)

**Aufgabe 9.1:***Wasserpendel (V-Aufgabe)*

Im Video wird Wasser in einem U-Rohr durch einen Unterdruck ausgelenkt und einfach losgelassen:

- a) Erklären Sie, weshalb eine harmonische Schwingung entsteht!
- b) (VA) Leiten Sie eine Formel für die Schwingungsdauer  $T_0$  des ungedämpften Wasserpendels her!  
Vergleichen Sie die berechnete mit einer experimentell bestimmten Schwingungsdauer!
- c) (VA) Bestimmen Sie das logarithmische Dekrement  $\delta$  der Schwingung!

(10 Punkte)

**Aufgabe 9.2:***Fadenpendel (T-Aufgabe)*

Aus einem dünnen, homogenen Stab (Trägheitsmoment berücksichtigen!) der Länge  $l = 1 \text{ m}$  möchten Sie ein Sekundenpendel ( $T/2 = 1 \text{ s}$ ) bauen. Betrachten Sie dafür den Stab als physikalisches Pendel mit kleiner Auslenkung und ohne Reibung. An welcher Stelle oberhalb des Schwerpunktes müssen Sie den Stab aufhängen (Drehachse)?

(10 Punkte)

**Aufgabe 9.2:***Fadenpendel (mV-Aufgabe)*

Basteln Sie aus einem Faden und einem Pendelkörper (z.B. Knete) ein gedämpftes Fadenpendel! In den Aufgabenteilen (a) und (b) werden Reibungskräfte vernachlässigt.

- a) Bestimmen Sie durch Videoanalyse die Eigenfrequenz  $\omega_0$  ihres Pendels! Vergleichen Sie diesen mit dem theoretischen Wert

$$\omega_0 = \sqrt{\frac{g}{l}} = \frac{2\pi}{T}, \quad (5)$$

indem Sie  $g = 9,81 \text{ m/s}^2$  als gegeben annehmen! Es bezeichnet  $l$  die Fadenlänge und  $T$  die Schwingungsdauer.

- b) Verifizieren Sie den Zusammenhang zwischen  $T$  und  $l$ , wie er durch Gleichung (5) impliziert wird, durch Variation der Pendellänge  $l$ ! Bestimmen Sie die Erdbeschleunigung  $g$  mit Hilfe von fünf Messwertpaaren  $(l, T)$ !
- c) Veranschaulichen Sie eine gedämpfte Schwingung durch Variation des Versuchsmaterials (z.B. indem Sie die Angriffsfläche der Luftreibungskraft durch Anbringen eines Pappkartons vergrößern)! Bestimmen Sie die Dämpfungskonstante  $\gamma$ !

(10 Punkte)



# Literaturverzeichnis

- Abbott, D. (2009). The LivePhoto Physics videos and video analysis site. *The Physics Teacher*, 47 (6), 396.
- Adair, J. G., Sharpe, D. & Huynh, C.-L. (1989). Hawthorne control procedures in educational experiments: A reconsideration of their use and effectiveness. *Review of Educational Research*, 59 (2), 215–228.
- Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D. & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics — Physics Education Research*, 2 (1).
- Ahad, N., Yin, T., Othman, A. & Yaaco, C. (2011). Sensitivity of normality tests to non-normal data. *Sains Malaysiana*, 40 (6), 637–641.
- Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, 16 (3), 183–198.
- Albrecht, A. (2011). *Längsschnittstudie zur Identifikation von Risikofaktoren für einen erfolgreichen Studieneinstieg in das Fach Physik* (Dissertation). Freie Universität Berlin.
- Amelang, M. & Schmidet-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (Bd. 4). Heidelberg: Springer.
- Aslanides, J. S. & Savage, C. M. (2013). Relativity concept inventory: Development, analysis, and results. *Physical Review Special Topics — Physics Education Research*, 9 (1).
- Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart and Winston.
- Backhaus, K., Erichson, B. & Weiber, R. (2011). *Fortgeschrittene multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. Heidelberg: Springer.
- Baddeley, A. D. & Hitch, G. (1974). Working memory. In G. H. Bower (Hrsg.), *The psychology of learning and motivation* (Bd. 8, S. 47–89). New York: Academic press.
- Bandura, A., Barbaranelli, C., Caprara, G. & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children's aspirations and career trajectories. *Child Development*, 72 (1), 187–206.
- Barab, S. A., Squire, K. D. & Dueber, W. (2000). A co-evolutionary model for supporting the emergence of authenticity. *Educational Technology Research and Development*, 48 (2), 37–62.
- Barniol, P. & Zavala, G. (2014). Test of understanding of vectors: A reliable multiple-choice vector concept test. *Physical Review Special Topics — Physics Education Research*, 10 (1).
- Beichner, R. (1990). The effect of simultaneous motion presentation and graph generation in a kinematics lab. *Journal of Research in Science Teaching*, 27 (8), 803–815.
- Beichner, R. (1994). Testing student interpretation of kinematics graphs. *American Journal of Physics*, 62, 750–762.
- Beichner, R. (1996). The impact of video motion analysis on kinematic graph interpretation skills. *American Journal of Physics*, 64 (10), 1772–1777.

- Belloni, M., Christian, W. & Brown, D. (2013). Teaching astronomy using Tracker. *The Physics Teacher*, 51 (3), 149–151.
- Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57 (1), 289–300.
- Bennett, J., Lubben, F. & Hogarth, S. (2007). Bringing science to life: A synthesis of the research evidence on the effects of context-based and STS approaches to science teaching. *Science Education*, 91 (3), 347–370.
- Berliner, D. (1990). What's all the fuss about instructional time? In M. Ben-Peretz & R. Bromme (Hrsg.), *The nature of time in schools: Theoretical concepts, practitioner perceptions*. New York, London: Teachers College Press.
- Berlyne, D. E. (1978). Curiosity and learning. *Motivation and Emotion*, 2 (2), 97–175.
- Bieda, K. N. & Nathan, M. J. (2009). Representational disfluency in algebra: evidence from student gestures and speech. *International Journal of Mathematics Education*, 41 (5), 637–650.
- Boekaerts, M. & Rozendaal, J. (2006). Self-regulation in Dutch secondary vocational education: Need for a more systematic approach to the assessment of self-regulation. In D. Euler, M. Lang & G. Pätzold (Hrsg.), *Zeitschrift für Berufs- und Wirtschaftspädagogik* (Bd. 20, S. 49–77). Stuttgart: Franz Steiner Verlag.
- Bolte, C. & Stork, H. (1996). Entwicklung und Einsatz von Erhebungsinstrumenten zur Analyse der Schüler- und Lehrer-Interaktion im Chemieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, 2 (1), 21–37.
- Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2. Aufl.). New York: Routledge.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler*. Heidelberg: Springer-Verlag.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler. Lehrbuch mit Online-Materialien*. Heidelberg: Springer-Verlag.
- Bowden, J., Dall'Alba, G., Martin, E., Laurillard, D., Marton, F., Maters, G., ... Walsh, E. (1992). Displacement, velocity, and frames of reference: Phenomenographic studies of students' understanding and some implications for teaching and assessment. *American Journal of Physics*, 60 (3), 262–269.
- Boyd, A. & Rubin, A. (1996). Interactive video: A bridge between motion and math. *International Journal of Computers for Mathematical Learning*, 1, 57–93.
- Brassel, H. (1987). The effect of real-time laboratory graphing on learning graphic representations of distance and velocity. *Journal of Research in Science Teaching*, 24 (4), 385–395.
- Brown, D. & Cox, A. J. (2009). Innovative uses of video analysis. *The Physics Teacher*, 47 (3), 145–150.
- Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review*, 31, 21–32.
- Brungardt, J. & Zollman, D. (1995). Analysis on kinematics graphing skills of high-school physics students. *Journal of Research in Science Teaching*, 32 (8), 855–869.
- Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (Bd. 2). Pearson.
- Caleon, I. & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, 32 (7), 939–961.
- Cariou, J.-Y. (2010). Tentative de détermination de l'authenticité des démarches d'investigation. In *Actes des journées scientifiques DIES*. Lyon: INRP.
- Cluss, M., Laws, K., Martin, N., Nowicki, T. S. & Mira, A. (2006). The indirect measurement of biomechanical forces in the moving human body. *American Journal of Physics*, 74 (2), 102.
- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, 1 (3), 98–101.

- Coletta, V. P., Phillips, J. A. & Steinert, J. J. (2007). Interpreting Force Concept Inventory scores: Normalized gain and SAT scores. *Physical Review Special Topics — Physics Education Research*, 3 (1).
- Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.
- Crouch, C. H., Fagen, A. P., Callanc, J. P. & Mazur, E. (2004). Classroom demonstrations: Learning tools or entertainment? *American Journal of Physics*, 72, 835–838.
- Crouch, C. H. & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69 (9), 970–977.
- De Ambrosis, A., Malgieri, M., Mascheretti, P. & Onorato, P. (2015). Investigating the role of sliding friction in rolling motion: a teaching sequence based on experiments and simulations. *European Journal of Physics*, 36 (3), 35020.
- Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik*, 2, 223–239.
- De Cock, M. (2012). Representation use and strategy choice in physics problem solving. *Physical Review Special Topics — Physics Education Research*, 8 (2).
- Demtröder, W. (2012). *Experimentalphysik 1: Mechanik und Wärme* (6. Aufl.). Heidelberg: Springer.
- Dengler, R. & Luchner, K. (1990). Bewegungsabläufe — aufgenommen, dargestellt und analysiert durch Videokamer und Computer. *Mathematisch-Naturwissenschaftlicher Unterricht*, 43 (5), 285–290.
- Derby, N. (1999). Reality and theory in a collision. *The Physics Teacher*, 37 (1), 24–27.
- Dilber, R., Karaman, I. & Duzgun, B. (2009). High school students' understanding of projectile motion concepts. *Educational Research and Evaluation*, 15 (3), 203–222.
- Ding, L. & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics — Physics Education Research*, 5 (2).
- diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction*, 22 (3), 293–331.
- Dowd, J. E., Araujo, I. & Mazur, E. (2015). Making sense of confusion: Relating performance, confidence, and self-efficacy to expressions of confusion in an introductory physics class. *Physical Review Special Topics — Physics Education Research*, 11 (1).
- Düchs, G. & Ingold, G. (2015). Weiter auf hohem Niveau—Statistiken zum Physikstudium an den Universitäten in Deutschland 2015. *Physik Journal*, 14 (8/9), 28–33.
- Duit, R. & Mikelskis-Seifert, S. (2007). Kontextorientierter Unterricht — wie man es einbettet, so wird es gelernt. *Unterricht Physik*, 18 (98), 4–8.
- Dziarstek, C. & Hilscher, H. (1998). Videoanalyse von Bewegungen — Computerprogramm DIVA und 61 ausgewertete Videoclips auf CD-Rom. *Praxis der Naturwissenschaften — Physik in der Schule*, 47 (1), 40–41.
- Eadkhong, T., Rajsadorn, R., Jannual, P. & Danworaphong, S. (2012). Rotational dynamics with tracker. *European Journal of Physics*, 33 (3), 615–622.
- Echternacht, G. (1971). The use of confidence testing in objective tests. *ETS Research Bulletin Series*, 1971 (2), 1–30.
- Engeln, K. (2004). *Schülerlabore: Authentische aktivierende Lernumgebungen als Möglichkeit, Interesse an Naturwissenschaften und Technik zu wecken*. Berlin: Logos Verlag.
- Enghag, M. & Niedderer, H. (2007). Two dimensions of student ownership of learning during small-group work in physics. *International Journal of Science and Mathematics Education*, 6 (4), 629–653.
- Escalada, L., Grabhorn, R. & Zollman, D. (1996). Applications of interactive digital video in a physics classroom. *Journal of Educational Multimedia and Hypermedia*, 5 (1), 73–97.
- Escalada, L. & Zollman, D. (1997). An investigation on the effects of using interactive digital video on physics classroom on student learning and attitudes. *Journal of Research in Science Teaching*, 34 (5), 467–489.

- Etkina, E., Murthy, S. & Zou, X. (2006a). Using introductory labs to engage students in experimental design. *American Journal of Physics*, 74 (11), 979–986.
- Etkina, E., Van Heuvelen, A., White-Brahmia, S., Brookes, D. T., Gentile, M., Murthy, S., . . . Warren, A. (2006b). Scientific abilities and their assessment. *Physical Review Special Topics — Physics Education Research*, 2 (2).
- Filipp, S.-H. (2006). Entwicklung von Fähigkeitsselbstkonzepten. *Zeitschrift für Pädagogische Psychologie*, 20, 65–72.
- Finkelstein, N. D. & Pollock, S. J. (2005). Replicating and understanding successful innovations: Implementing tutorials in introductory physics. *Physical Review Special Topics — Physics Education Research*, 1 (1).
- Fischer, G. H. (2012). Derivations of the Rasch Model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5), 378–382.
- Fliege, H., Rose, M., Arck, P., Levenstein, S. & Klapp, B. F. (2001). Validierung des Perceived Stress Questionnaire (PSQ) an einer deutschen Stichprobe. *Diagnostika*, 47 (3), 142–152.
- Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, 7 (2), 33–38.
- Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice*, 11 (4), 21–26.
- Furtak, E. M., Seidel, T., Iverson, H. & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, 82 (3), 300–329.
- Gallis, M. R. (2010). Artificial video for video analysis. *The Physics Teacher*, 48 (1), 32–34.
- Garavalia, L. S. & Gredler, M. E. (2002). An exploratory study of academic goal setting, achievement calibration and self-regulated learning. *Journal of Instructional Psychology*, 29 (4), 221–230.
- Gerholz, K.-H. (2012). Selbstreguliertes Lernen in der Hochschule fördern — Lernkulturen gestalten. *Zeitschrift für Hochschulentwicklung*, 3, 60–73.
- Geßner, T., Benz, M. & Wilhelm, T. (2009). Einstein, Maxwell und Co. auf der Spur — Auswertung von Modellgasen mittels Videoanalyse. *Physik und Didaktik in Schule und Hochschule*, 8 (1), 1–10.
- Gil, S., Reisin, H. D. & Rodríguez, E. E. (2006). Using a digital camera as a measuring device. *American Journal of Physics*, 74 (9), 768.
- Gilbert, J. K., Bulte, A. M. & Pilot, A. (2011). Concept development and transfer in context-based science education. *International Journal of Science Education*, 33 (6), 817–837.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5 (10), 3–8.
- Goff, J. E. & Carré, M. J. (2009). Trajectory analysis of a soccer ball. *American Journal of Physics*, 77 (11), 1020–1027.
- GRE. (2011). *Graduate Record Examinaton - physics test: Practice book*. <http://www.ets.org/gre>: Graduate Record Examinations Board.
- Green, S., Lissitz, R. & Mulaik, S. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational Psychological Measurement*, 37, 827–838.
- Gröber, S., Klein, P. & Kuhn, J. (2014). Video-based problems in introductory mechanics physics courses. *European Journal of Physics*, 35 (5), 055019.
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66 (1), 64–74.

- Härtig, H. (2014). Der Force Concept Inventory — Vergleich einer offenen und einer geschlossenen Version. *Physik und Didaktik in Schule und Hochschule*, 13 (1), 53–61.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Haugland, O. A. (2001). Physics measurements for sports. *The Physics Teacher*, 39 (6), 350–353.
- Häußler, P. & Hoffmann, L. (1995). Physikunterricht — an den Interessen von Mädchen und Jungen orientiert. *Unterrichtswissenschaft*, 23, 107–126.
- Häußler, P. & Lind, G. (2000). Aufgabenkultur— Was ist das? *Praxis der Naturwissenschaft — Physik in der Schule*, 49 (4), 2–10.
- Heck, A., Uylings, P. & Kedzierska, E. (2009). Understanding the physics of bungee jumping. *Physics Education*, 45 (1), 63–72.
- Henning, T., Müller, R. & Strahl, A. (2012). Authentische Aufgaben an der Hochschule. *Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung*, 1–5.
- Herrington, J., Oliver, R. & Reeves, T. (2003). Patterns of engagement in authentic online learning environments. *Australasian Journal of Educational Technology*, 19 (1).
- Hestenes, D. & Halloun, I. (1995). Interpreting the Force Concept Inventory: A response to March 1995 critique by Huffman and Heller. *The Physics Teacher*, 33 (8), 502.
- Hestenes, D., Wells, M. & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, 30, 141–158.
- Heublein, U. & Wolter, A. (2011). Studienabbruch in Deutschland. Definition, Häufigkeit, Ursachen, Maßnahmen. *Zeitschrift für Pädagogik*, 57 (2), 214–236.
- Hill, M., Sharma, M. D. & Johnston, H. (2015). How online learning modules can improve the representational fluency and conceptual understanding of university physics students. *European Journal of Physics*, 36 (4), 045019.
- Hirth, M. (2016). *iAcoustics [Arbeitstitel]* (Unveröffentlichte Dissertation). Technische Universität Kaiserslautern, Fachbereich Physik.
- Hochberg, K. (2016). *iMechanics: Smartphones als Experimentiermittel im Physikunterricht der Sekundarstufe II — Wirkung auf Lernerfolg, Motivation und Neugier in der Mechanik* (Unveröffentlichte Dissertation). Technische Universität Kaiserslautern, Fachbereich Physik.
- Hockicko, P. (2012). Attractiveness of learning physics by means of video analysis and modeling tools. In *Sefi 40th annual conference*. Thessaloniki, Greece.
- Hockicko, P., Trpišová, B. & Ondruš, J. (2014). Correcting students' misconceptions about automobile braking distances and video analysis using interactive program Tracker. *Journal of Science Education and Technology*, 23 (6), 763–776.
- Höffler, T. N. & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, 17 (6), 722–738.
- Hoffmann, L., Häußler, P. & Lehrke, M. (1998). *Die IPN-Interessenstudie Physik*. IPN.
- Hollingsworth, H. H. (1980). An analytical investigation of the effects of heterogeneous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, 40 (3), 611–618.
- Holstermann, N. & Bögeholz, S. (2007). Interesse von Jungen und Mädchen an naturwissenschaftlichen Themen am Ende der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften*, 13, 71–86.
- Hsu, L., Brewe, E., Foster, T. M. & Harper, K. A. (2004). Resource letter rps-1: Research in problem solving. *American Journal of Physics*, 72 (9), 1147.
- Huffman, D. & Heller, P. (1995). What does the Force Concept Inventory actually measure? *The Physics Teacher*, 33, 138–143.
- Jamieson, J. (2004). Analysis of covariance (ANCOVA) with difference scores. *International Journal of Psychophysiology*, 52 (3), 277–283.
- Kaiser, M. K., Proffitt, D. R. & McCloskey, M. (1985). The development of beliefs about falling objects. *Perception and Psychophysics*, 38 (6), 533–539.

- Kanim, S. & Subero, K. (2010). Introductory labs on the vector nature of force and acceleration. *American Journal of Physics*, 78 (5), 461–466.
- Kasten, K. (2011). Mit Mathe nicht gerechnet — Jeder dritte Physikstudent gibt in den ersten Semestern auf. Viele unterschätzen die Anforderungen. *Die Zeit*. Zugriff auf <http://www.zeit.de/2011/22/C-Physik>
- Kelley, D. H. & Ouellette, N. T. (2011). Using particle tracking to measure flow instabilities in an undergraduate laboratory experiment. *American Journal of Physics*, 79 (3), 267–273.
- KFP. (2010). *Zur Konzeption von Bachelor- und Master-Studiengängen in der Physik*. Handreichung. Berlin. Zugriff auf [http://www.kfp-physik.de/dokument/KFP\\_Handreichung\\_Konzeption-Studiengaenge-Physik-101108.pdf](http://www.kfp-physik.de/dokument/KFP_Handreichung_Konzeption-Studiengaenge-Physik-101108.pdf)
- Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.
- Klein, P., Gröber, S., Kuhn, J., Fleischhauer, A. & Müller, A. (2015d). The right frame of reference makes it simple: an example of introductory mechanics supported by video analysis of motion. *European Journal of Physics*, 36 (1), 015004.
- Klein, P., Gröber, S., Kuhn, J., Fouckhardt, H., von Freymann, G., Oesterschulze, E., ... Müller, A. (2015a). physics.move — Videoanalyse-Aufgaben in der Experimentalphysik 1. *Physik und Didaktik in Schule und Hochschule*, 14 (1).
- Klein, P., Gröber, S., Kuhn, J. & Müller, A. (2013a). Video analysis of projectile motion using tablet computers as experimental tools. *Physics Education*, 49 (1), 37–40.
- Klein, P., Gröber, S., Kuhn, J. & Müller, A. (2013b). Experimentelle Aufgaben in den Übungen zur Experimentalphysik 1. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht* (S. 312–314). Beiträge zur GDGP-Jahrestagung in München 2013, Kiel: IPN.
- Klein, P., Gröber, S., Kuhn, J. & Müller, A. (2014b). Mobile Videoanalyse mit Tablets — Am Beispiel des freien Falls mit Luftreibung. *Praxis der Naturwissenschaften — Physik in der Schule*, 63 (5), 33–35.
- Klein, P., Groeber, S., Kuhn, J. & Wilhelm, T. (2015b). Bewegungen mit der App Video Analysis analysieren. *Physik in unserer Zeit*, 46 (2), 98–99.
- Klein, P., Kuhn, J. & Müller, A. (2015b). iMobile physics: A research and development project for teaching and learning with smartphones and tablets as mobile experimental tools. In C. Fazio & R. M. S. Mineo (Hrsg.), *Teaching and learning physics: Integrating research into practice* (S. 629–638). Proceedings of the GIREP-MPTL 2014 International Conference (Palermo, Juli 2014).
- Klein, P., Kuhn, J. & Müller, A. (2016). Experimente mit Smartphone und Tablet-PC: Analyse leistungsbezogener Antwortsicherheiten im Physikstudium. In S. Bastian & S. Aufenanger (Hrsg.), *Tablets in Schule und Unterricht. Forschungsmethoden und -perspektiven zum Einsatz digitaler Medien*. Wiesbaden: Springer VS. (in Druck).
- Klein, P., Kuhn, J., Müller, A. & Gröber, S. (2015c). Video analysis exercises in regular introductory mechanics physics courses: Effects of conventional methods and possibilities of mobile devices. In A. Kauertz, H. Ludwig, A. Müller, J. Pretsch & W. Schnotz (Hrsg.), *Multidisciplinary research on teaching and learning*. Basingstoke: Palgrave Macmillan.
- Klingerl, W. (1991). Wie schnell fliegt eine Spielzeugrakete? *Naturwissenschaft im Unterricht — Physik*, 2 (8), 20–25.
- Kohl, P. B. (2007). *Towards an understanding of how students use representations in physics problem solving* (Unveröffentlichte Dissertation). University of Colorado.
- Kohl, P. B. & Finkelstein, N. D. (2005). Student representational competence and self-assessment when solving physics problems. *Physical Review Special Topics — Physics Education Research*, 1 (1).
- Kohl, P. B. & Finkelstein, N. D. (2008). Patterns of multiple representation use by experts and novices during physics problem solving. *Physical Review Special Topics — Physics*

- Education Research*, 4 (1).
- Koleza, E. & Pappas, J. (2008). The effect of motion analysis activities in a video-based laboratory in students' understanding of position, velocity and frame of reference. *International Journal of Mathematical Education in Science and Technology*, 39 (6), 701–725.
- Koller, I., Alexandrowicz, R. & Hatzinger, R. (2012). *Das Rasch Modell in der Praxis: Eine Einführung in eRm*. Wien: facultas.wuv. (UTB-Band-Nr: 3786)
- Krapp, A. (1989). Der Stellenwert des Interessenkonzepts in der pädagogisch-orientierten Forschung. *Empirische Pädagogik*, 3, 233–255.
- Krapp, A., Schiefele, U., Wild, K. P. & Winteler, A. (1993). Der Fragebogen zum Studieninteresse (FSI). *Diagnostika*, 39 (4), 335–351.
- Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77 (6), 1121–1134.
- Kuhn, J. (2010). *Authentische Aufgaben im theoretischen Rahmen von Instruktions- und Lehr-Lern-Forschung*. Wiesbaden: Vieweg und Teubner Verlag.
- Kuhn, J. (2014). Relevant information about using a mobile phone acceleration sensor in physics experiments. *American Journal of Physics*, 82 (2), 94.
- Kuhn, J. & Vogt, P. (2015). Smartphone & co. in physics education: Effects of learning with new media experimental tools in acoustics. In W. Schnotz, A. Kauertz, H. Ludwig, A. Müller & J. Pretsch (Hrsg.), *Multidisciplinary research on teaching and learning* (S. 253–269). Basingstoke, UK: Palgrave Macmillan.
- Laws, P. (1998). Using digital video analysis in introductory mechanics projects. *The Physics Teacher*, 36 (5), 282.
- Leonhart, R. (2013). *Lehrbuch Statistik: Einstieg und Vertiefung*. Bern: Huber.
- Levenstein, S., Prantera, C., Varvo, V., Scribano, M., Berto, E., Luzi, C. & Andreoli, A. (1993). Development of the Perceived Stress Questionnaire: A new tool for psychosomatic research. *Journal of Psychosomatic Research*, 37 (1), 19–32.
- Linacre, J. M. (2003). Size vs. significance: Standardized chi-square fit statistic. *Rasch Measurement Transactions*, 17 (1), 918.
- Lindsey, B. A. & Nagel, M. L. (2015). Do students know what they know? Exploring the accuracy of students' self-assessments. *Physical Review Special Topics — Physics Education Research*, 11 (2).
- Litman, J. A. & Spielberger, C. D. (2003). Measuring epistemic curiosity and its diversive and specific components. *Journal of Personality Assessment*, 80 (1), 75–86.
- Lynn, R. (1994). Sex differences in intelligence and brain size: A paradox resolved. *Personality and Individual Differences*, 17 (2), 257–271.
- Mair, P. & Hatzinger, R. (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, 49, 26–43.
- Malgieri, M., Onorato, P., Mascheretti, P. & De Ambrosis, A. (2014). Pre-service teachers' approaches to a historical problem in mechanics. *Physics Education*, 49 (5), 500–511.
- Mandl, H. & Kopp, B. (2005). Situated learning: Theories and models. In P. Nentwig & D. Waddington (Hrsg.), *Context-based learning of science*. Münster, New York, München, Berlin: Waxmann.
- Marxen, M. (2013). *Untersuchung von Grenzflächenspannung an Flüssigkeiten durch Videoanalyse und Bildanalyse* (Masterarbeit im Fach Physik). Technische Universität Kaiserslautern, Kaiserslautern.
- Mashood, K. K. & Singh, V. A. (2012). An inventory on rotational kinematics of a particle: unravelling misconceptions and pitfalls in reasoning. *European Journal of Physics*, 33 (5), 1301–1312.
- Mayer, R. E. (1999). Multimedia aids to problem-solving transfer. *International Journal of Educational Research*, 31 (7), 611–623.
- Mayer, R. E. (2009). *Multimedia learning* (2. Aufl.). New York: Cambridge University Press.

- Mayer, R. E. & Anderson, R. B. (1991). Animations need narrations: an experimental test of a dual-coding hypothesis. *Journal of Educational Psychology*, 83 (4), 484–490.
- Mayer, R. E. & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38 (1), 43–52.
- Mayers, A. (2013). *Introduction to statistics and SPSS in psychology*. Pearson.
- McCloskey, M. (1983). Naive theory of motion. In D. Gentner & A. L. Stevens (Hrsg.), *Mental modals* (S. 299–324). Hillsdale, NJ: Erlbaum.
- McDermott, L. C., Rosenquist, M. & van Zee, E. (1987). Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics*, 55 (6), 503–513.
- Meltzer, D. E. (2005). Relation between students' problem-solving performance and representational format. *American Journal of Physics*, 73 (5), 463–478.
- Merton, R. K. (1968). The Matthew-effect in science: The reward and communication systems of science are considered. *Science*, 159 (3810), 56–63.
- Michel, M. & Wilhelm, T. (2010). Dynamik mit Hochgeschwindigkeitvideos. *Praxis der Naturwissenschaften — Physik in der Schule*, 59 (7), 23–30.
- Miles, J. & Shevlin, M. (2001). *Applying regression and correlation — a guide for students and researchers*. London: SAGE Publications.
- Miller, K., Lasry, N., Chu, K. & Mazur, E. (2013). Role of physics lecture demonstrations in conceptual learning. *Physical Review Special Topics — Physics Education Research*, 9 (2), 1–5.
- Mobalegh, A. & Barati, H. (2012). Multiple true-false (MTF) and multiple-choice (MC) test formats: A comparison between two versions of the same test paper of Iranian NUEE. *Journal of Language Teaching and Research*, 3 (5), 1027–1037.
- Molina-Bolívar, J. A. & Abella-Palacios, A. J. (2012). Magnetic braking: A video analysis. *The Physics Teacher*, 50 (7), 412–413.
- Molz, A. (2016). *Verbindung von Schülerlabor und Schulunterricht — Auswirkungen auf motivationale und kognitive Aspekte im Fach Physik [Arbeitstitel]* (Unveröffentlichte Dissertation). Technische Universität Kaiserslautern, Fachbereich Physik.
- Moosbrugger, H. & Augustin, K. (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.
- Moreno, R. (2005). Instructional technology: Promise and pitfalls. In L. PytlikZillig, M. Bodvarsson & R. Bruning (Hrsg.), *Technology-based education: Bringing researchers and practitioners together* (S. 1–19). Greenwich, CT: Information Age Publishing.
- Moreno, R. & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, 19 (3), 309–326.
- Mozafari, S., Bukhari, S., Klein, P., Kuhn, J. & Dengel, A. (2016). *An eye movement research on representational competence in physics*. (Conference Paper submitted to ETRA conference (Eye Tracking Research and Applications 2016))
- Müller, R. (2006). Kontextorientierung und Alltagsbezug. In H. F. Mikelskis (Hrsg.), *Physikdidaktik* (S. 102–118). Berlin: Cornelsen.
- Naylor, F. D. (1981). A state-trait curiosity inventory. *Australian Psychologist*, 16 (2), 172–183.
- Neumann, K. (2014). Rasch-Analyse naturwissenschaftsbezogener Leistungstests. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschafts-didaktischen Forschung* (S. 355–369). Berlin, Heidelberg: Springer.
- Nieminen, P., Savinainen, A. & Viiri, J. (2010). Force Concept Inventory-based multiple-choice test for investigating students' representational consistency. *Physical Review Special Topics — Physics Education Research*, 6 (2).
- Nieminen, P., Savinainen, A. & Viiri, J. (2012). Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning. *Physical Review Special Topics — Physics Education Research*, 8 (1).

- Oehme, W. & Rieger, P. (1996). Besseres Verstehen durch genaueres Sehen. *Physik in der Schule*, 34 (2), 67–69.
- Osborne, J. (2003). Attitudes towards science: a review of the literature and its implications. *International Journal of Science Education*, 25 (9), 1049–1079.
- Page, A., Moreno, R., Candelas, P. & Belmar, F. (2008). The accuracy of webcams in 2D motion analysis: sources of error and their control. *European Journal of Physics*, 29 (4), 857–870.
- Paivio, A. (1986). *Mental representations: A dual coding approach*. New York: Oxford University Press.
- Palazzo, D. & Schools, C. (2008). Video analysis: The next physics laboratory? In *Proceedings of the engineering education: 21st century frontiers*. NY: West Point.
- Pappas, J., Koleza, E., Rizos, J. & Skordoulis, C. (2002). Using interactive digital video and motion analysis to bridge abstract mathematical notions with concrete everyday experience. In *2nd International Conference on the Teaching of Mathematics*. Hersonissos.
- Phommarach, S., Wattanakasiwich, P. & Johnston, I. (2012). Video analysis of rolling cylinders. *Physics Education*, 47 (2), 189–196.
- Pieschl, S. (2008). Metacognitive calibration — an extended conceptualization and potential applications. *Metacognition Learning*, 4 (1), 3–31.
- PISA. (2007). *PISA 2006 — Die Ergebnisse der dritten internationalen Vergleichsstudie* (PISA-Konsortium Deutschland, Hrsg.). Münster: Waxmann.
- Planinic, M., Boone, W. J., Krsnik, R. & Beilfuss, M. L. (2006). Exploring alternative conceptions from Newtonian dynamics and simple DC circuits: Links between item difficulty and item confidence. *Journal of Research in Science Teaching*, 43 (2), 150–171.
- Planinic, M., Ivanjek, L. & Susac, A. (2010). Rasch model based analysis of the Force Concept Inventory. *Physical Review Special Topics — Physics Education Research*, 6 (1).
- Potgieter, M., Malatje, E., Gaigher, E. & Venter, E. (2010). Confidence versus performance as an indicator of the presence of alternative conceptions and inadequate problem-solving skills in mechanics. *International Journal of Science Education*, 32 (11), 1407–1429.
- Radinsky, J., Bouillion, L., Lento, E. M. & Gomez, L. M. (2001). Mutual benefit partnership: A curricular design for authenticity. *Journal of Curriculum Studies*, 33 (4), 405–430.
- Rainer, J. D. & Matthews, M. W. (2002). Ownership of learning in teacher education. *Action in Teacher Education*, 24 (1), 22–30.
- Ramm, M. (2008). *Das Studium der Naturwissenschaften. Eine Fachmonographie aus studentischer Sicht*. Bonn, Berlin.
- Rebello, N. S., Engelhardt, P. V. & Singh, C. (2012). How accurately can students estimate their performance on an exam and how does this relate to their actual performance on the exam? In *AIP conference proceedings* (Bd. 1413, S. 315). American Institute of Physics.
- Reinmann, G. & Mandl, H. (2006). Unterrichten und Lernumgebungen gestalten. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie* (S. 613–658). Weinheim: Beltz.
- Rickey, D. & Stacy, A. M. (2000). The role of metacognition in learning chemistry. *Journal of Chemical Education*, 77 (7), 915–920.
- Riera, J., Monsoriu, J. A., Giménez, M. H., Hueso, J. L. & Torregrosa, J. R. (2003). Using image recognition to automate video analysis of physical processes. *American Journal of Physics*, 71 (10), 1075–1078.
- Rodrigues, M. & Simeao-Carvalho, P. (2014). Teaching optical phenomena with tracker. *Physics Education*, 49 (6), 671–677.
- Rodrigues, S., Pearce, J. & Livett, M. (2001). Using video analysis or data loggers during practical work in first year physics. *Educational Studies*, 27 (1), 31–43.

- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24 (2), 3–13.
- Rosengrant, D., Etkina, E. & Van Heuvelen, A. (2006). An overview of recent research on multiple representations [Physics Education Research Conference. Proceedings held at Syracuse, New York]. In L. McCullough, L. Hsu & P. Heron (Hrsg.), *AIP conference proceedings* (Bd. 883, S. 149 – 152). Melville, NY: American Institute of Physics.
- Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion* (Bd. 2). Bern: Huber.
- Roth, W.-M. & Bowen, G. M. (2001). Professionals read graphs: A semiotic analysis. *Journal for Research in Mathematics Education*, 32 (2), 159–194.
- Ruiz-Primo, M. A., Briggs, D., Iverson, H., Talbot, R. & Shepard, L. A. (2011). Impact of undergraduate science course innovations on learning. *Science*, 331 (6022), 1269–1270.
- Ryan, R. M. & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25 (1), 54–67.
- Sacher, M. (2014). Dem Gefühl für Physik auf der Spur. *Physik Journal*, 13 (6), 3.
- Savinainen, A., Mäkynen, A., Nieminen, P. & Viiri, J. (2013). Does using a visual-representation tool foster students' ability to identify forces and construct free-body diagrams? *Physical Review Special Topics — Physics Education Research*, 9 (1).
- Sayre, E. C. & Wittmann, M. C. (2008). Plasticity of intermediate mechanics students' coordinate system choice. *Physical Review Special Topics — Physics Education Research*, 4 (2).
- Schecker, H. & Gerdes, J. (1999). Messung von Konzeptualisierungsfähigkeit in der Mechanik. *Zeitschrift für Didaktik der Naturwissenschaften*, 5 (1), 75–89.
- Scheid, J. (2013). *Multiple Repräsentationen, Verständnis physikalischer Experimente und kognitive Aktivierung — Ein Beitrag zur Entwicklung der Aufgabenkultur* (Bd. 151; H. Niedderer & H. Fischler, Hrsg.). Berlin: Logos.
- Schraw, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognition and Psychology*, 9 (4), 321–332.
- Schunk, D. H. & Pajares, F. (2004). Self-efficacy in education revisited: Empirical and applied evidence. In D. McInerney & S. Van Etten (Hrsg.), *Big theories revisited* (S. 115–138). Greenwich, CT: Information Age.
- Schüttler, M. & Wilhelm, T. (2011). Bewegungsanalyse im Freizeitpark. *Praxis der Naturwissenschaften — Physik in der Schule*, 60 (6), 18–24.
- Scott, T. F., Schumayer, D. & Gray, A. R. (2012). Exploratory factor analysis of a Force Concept Inventory data set. *Physical Review Special Topics — Physics Education Research*, 8 (2).
- Seifert, S., Mikelskis, H. & Winter, R. (1997). Der Computer — ein multimediales Werkzeug zum Lernen von Physik. *Physik in der Schule*, 35 (9), 306–311.
- Sharma, M. & Bewes, J. (2011). Self-monitoring: Confidence, academic achievement and gender differences in physics. *Journal of Learning Design*, 4 (3), 1–13.
- Shaughnessy, J. (1979). Confidence-judgement accuracy as a predictor of test performance. *Journal of Research in Personality*, 13, 505–514.
- Sheldon, K. & Deci, E. (o.J.). *The Self-Determination Scale*. Zugriff auf <http://www.selfdeterminationtheory.org/self-determination-scale/> (University of Rochester)
- Sim, J. & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85 (3), 257–268.
- Smith, L., Parker, D., Brown, C. & Ireson, G. (2010). Showing  $F = BIL$  using video analysis. *Physics Education*, 46 (1), 19–21.
- Spangenberg, S. (2015). *Videobasierte Lernaufgaben zur Experimentalphysik 1 — Drehimpuls und Drehimpulserhaltung mit dem Luftkissentisch* (Masterarbeit im Fach Physik). Technische Universität Kaiserslautern, Kaiserslautern.

- Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences* (5. Aufl.). London, New York: Routledge/Taylor & Francis.
- Stewart, J., Griffin, H. & Stewart, G. (2007). Context sensitivity in the Force Concept Inventory. *Physical Review Special Topics — Physics Education Research*, 3 (1).
- Stiensmeier-Pelster, J. & Schöne, C. (2008). Handbuch der Pädagogischen Psychologie. In W. Schneider & M. Hasselhorn (Hrsg.), (S. 62–73). Göttingen: Hogrefe.
- Stone, C. A. (1998). The metaphor of scaffolding: Its utility for the field of learning disabilities. *Journal of Learning Disabilities*, 31 (4), 344–364.
- Streiner, D. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80, 99–103.
- Strobl, C. (2012). *Das Rasch-Modell* (Bd. 2). Mering: Rainer Hampp Verlag.
- Struck, W. & Yerrick, R. (2010). The effect of data acquisition-probeware and digital video analysis on accurate graphical representation of kinetics in a high school physics class. *Journal of Science Education and Technology*, 19 (2), 199–211.
- Suleder, M. (2010). *Videoanalyse und Physikunterricht*. Aulis.
- Sullivan, P., Novak, J. & Sancilio, P. (2006). A block dragging a cart. *The Physics Teacher*, 44 (2), 114.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12 (2), 257–285.
- Tabachnick, B. G. & Fidell, L. S. (2001). *Using multivariate analysis*. California State University Northridge: Harper Collins College Publishers.
- Tavakol, M. & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55.
- Tesch, F. (2015). DPG-Lehrerfortbildung: Smartphone und Tablet-PC im Physikunterricht. *Physik Journal*, 14 (1), 63.
- Thornton, R. K. (1987). Tools for scientific thinking — microcomputer-based laboratories for physics teaching. *Physics Education*, 22, 230–238.
- Thornton, R. K. & Sokoloff, D. (1998). Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics*, 66 (4), 338–352.
- Tobin, K. G. & Capie, W. (1981). The development and validation of a group test of logical thinking. *Educational and Psychological Measurement*, 41 (2), 413–423.
- Trapmann, S., Hell, B., Weigand, S. & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs — eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie*, 21 (1), 11–27.
- Treagust, D. (1986). Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science Education*, 16 (1), 199–207.
- Trudel, L. & Metiouri, A. (2012). Effect of a video-based laboratory on high school pupils' understanding of constant speed motion. *International Journal of Advanced Computer Science and Application*, 3 (5), 79–86.
- Trudel, L. & Metiouri, A. (2013). Favouring the understanding of parabolic motion in a video-based laboratory. In C. P. Constantinou, N. Papadouris & A. Hadjigeorgiou (Hrsg.), *E-book proceedings of the ESERA 2013 conference: Science education research for evidence-based teaching and coherence in learning*. Nicosia, Cyprus.
- Trumper, R. (2003). The physics laboratory — a historical overview and future perspectives. *Science & Education*, 12, 645–670.
- Usman, B. A., Alam, J. & Anwar, M. S. (2015). Video motion analysis with automated tracking: an insight. *European Journal of Physics*, 36 (6), 065049.
- Van Heuvelen, A. (1991). Learning to think like a physicist: A review of research-based instructional designs. *American Journal of Physics*, 59, 891–897.
- Van Hook, S., Lark, A., Hodges, J., Celebrezze, E. & Channels, L. (2007). Playground physics: Determining the moment of inertia of a merry-go-round. *The Physics Teacher*, 45 (2), 85–87.

- Velden, M. (1982). *Die Signalentdeckungstheorie in der Psychologie*. Stuttgart: Kohlhammer.
- Victor, A., Elsäßer, A., Hommel, G. & Blettner, M. (2010). Wie bewertet man die p-Wert-Flut? *Deutsches Ärzteblatt*, 107 (4), 50–56.
- Vogt, P. (2010). *Werbeaufgaben im Physikunterricht: Motivations- und Lernwirksamkeit authentischer Texte*. Wiesbaden: Vieweg und Teubner.
- Vollmer, M. & Möllmann, K.-P. (2011). Exploding balloons, deformed balls, strange reflections and breaking rods: slow motion analysis of selected hands-on experiments. *Physics Education*, 46 (4), 472–485.
- Vollmer, M. & Möllmann, K.-P. (2012). Vapour pressure and adiabatic cooling from champagne: slow-motion visualization of gas thermodynamics. *Physics Education*, 47 (5), 608–615.
- von Davier, M. (2001). WINMIRA [Software-Handbuch]. St. Paul, MN. (Computer Software)
- von Stumm, S., Hell, B. & Chamorro-Premuzic, T. (2011). The hungry mind: Intellectual curiosity is the third pillar of academic performance. *Perspectives on Psychological Science*, 6 (6), 574–588.
- Watkins, J., Coffey, J. E., Redish, E. F. & Cooke, T. J. (2012). Disciplinary authenticity: Enriching the reforms of introductory physics courses for life-science students. *Physical Review Special Topics — Physics Education Research*, 8 (1).
- Wee, L. K. (2012). One-dimensional collision carts computer model and its design ideas for productive experiential learning. *Physics Education*, 47 (3), 301–308.
- Wee, L. K., Chew, C., Goh, G., Tan, S. & Lee, T. (2012). Using tracker as a pedagogical tool for understanding projectile motion. *Physics Education*, 47 (4), 448–455.
- Wee, L. K., Tan, K., Leong, T. & Tan, C. (2015). Using tracker to understand 'toss up' and free fall motion: a case study. *Physics Education*, 50 (4), 436–442.
- Wehrbein, W. M. (2001). Using video analysis to investigate intermediate concepts in classical mechanics. *American Journal of Physics*, 69 (7), 818–820.
- Weidig, J. (2013). *Videoanalyse in Übungen zur Experimentalphysik (Mechanik) — Modellexperimente und Aufgaben zu Impuls- und Energieerhaltung* (Masterarbeit im Fach Physik). Technische Universität Kaiserslautern, Kaiserslautern.
- Wilhelm, J., Thacker, B. & Wilhelm, R. (2007). Creating constructivist physics for introductory university classes. *Electronic Journal of Science Education*, 11 (2), 19–37.
- Wilhelm, T. (2009). Bierschaumzerfall — Modelle und Realität im Vergleich. *Praxis der Naturwissenschaften — Physik in der Schule*, 58 (8), 19–26.
- Wilhelm, T. (2011a). Möglichkeiten der Videoanalyse [Habilitationsschrift]. Würzburg.
- Wilhelm, T. (2011b). Sekt oder Selters? Wie steigen die Blasen auf? *Praxis der Naturwissenschaften — Physik in der Schule*, 60 (8), 28–34.
- Wilhelm, T., Gessner, T., Suleder, M. & Heuer, D. (2003). Sportaktivitäten vielseitig analysieren und modellieren. *Praxis der Naturwissenschaften — Physik in der Schule*, 52 (2), 23–29.
- Wilhelm, T. & Heuer, D. (2005). Verständnis der Newtonschen Mechanik bei bayerischen Elftklässlern — Ergebnisse beim Test Force Concept Inventory in herkömmlichen Klassen und im Würzburger Kinematik-/Dynamikunterricht. In V. Nordmeier & A. Oberländer (Hrsg.), *Physik und Didaktik in Schule und Hochschule*.
- Wolf, B. (1998). Effektstärkenmaße. In D. Rost (Hrsg.), *Handwörterbuch der Pädagogischen Psychologie* (S. 72–75). Weinheim: Beltz PVU.
- Zajkov, O. & Mitrevski, B. (2012). Video measurements: Quantity or quality. *European Journal of Physics Education*, 3 (4), 34–43.

# Sonstige Anhänge

## Lebenslauf

**Kontakt** Pascal Klein  
Davenportplatz 21  
67663 Kaiserslautern  
pascalklein@t-online.de



### Persönliche Daten

Geburtsdatum 05.03.1987  
Geburtsort Zweibrücken  
Zivilstand ledig

### Akademische Ausbildung

seit 12/2012 **Technische Universität Kaiserslautern**  
Wissenschaftlicher Mitarbeiter (Doktorand) in der AG  
Didaktik der Physik

04/2006 – 05/2012 **Technische Universität Kaiserslautern**  
Studium der Mathematik und Physik für das Lehramt  
an Gymnasien. Erstes Staatsexamen mit Auszeichnung  
bestanden (1.0)

03/2006 **Hofenfels Gymnasium, Zweibrücken**  
Allgemeine Hochschulreife (1.5)

### Eingeladene Vorträge und Beteiligung an der Durchführung von Lehrerfortbildungen (Auswahl)

03/2016 **Université de Genève**  
Continuous professional education for science teachers  
(Smartphones as experimental devices in the science classroom)

02/2015 **ETH Zürich**  
Public lecture (mit J. Kuhn): Experimentbasierte Aufgaben in den  
klassischen Übungen zur Experimentalphysik 1

10/2014 **Bad Honnef, DPG-Lehrerfortbildung**  
Lehrerfortbildung (Smartphones im Physik-Unterricht)

### Forschungsaufenthalte (Auswahl)

11/2015–12/2015 **Université de Genève** — Didactique de Sciences Physique  
09/2014–11/2014 **Université de Genève** — Didactique de Sciences Physique

## Publikationsliste

### Referierte Beiträge (peer-reviewed):

- Klein, P., Kuhn, J. & Müller, A. (2016). Experimente mit Smartphone und Tablet-PC: Analyse leistungsbezogener Antwortsicherheiten im Physikstudium. In: S. Bastian & S. Aufenanger (Hrsg.), *Tablets in Schule und Unterricht. Forschungsmethoden und -perspektiven zum Einsatz digitaler Medien*. Wiesbaden: Springer VS. (in Druck)
- Klein, P., Gröber, S., Kuhn, J., Fleischhauer, A. & Müller, A. (2015). The right frame of reference makes it simple — an example of introductory physics supported by video analysis of motion. *European Journal of Physics*, 36 (1), 015004.
- Klein, P., Gröber, S., Kuhn, J., Fouckhardt, H., von Freymann, G., Oesterschulze, E., Widera, A., Fleischhauer, A. & Müller, A. (2015). physics.move — Videoanalyse-Aufgaben in der Experimentalphysik 1. *Physik und Didaktik in Schule und Hochschule*, 14 (1).
- Klein, P., Kuhn, J. & Müller, A. iMobile Physics (2015): A research and development project for teaching and learning with smartphones and tablet PCs as mobile experimental tools. In C. Fazio & R. M. S. Mineo (Hrsg.), *Teaching and learning physics: Integrating research into practice* ( S. 629–638). Proceedings of the GIREP-MPTL 2014 International Conference (Palermo, Juli 2014).
- Gröber, S., Klein, P. & Kuhn, J. (2014). Video-based problems in introductory mechanics physics courses. *European Journal of Physics*, 35 (5), 055019.
- Klein, P., Hirth, M., Gröber, S., Kuhn, J. & Müller, A. (2014) Classical experiments revisited: Smartphone and tablets as experimental tools in acoustics and optics. *Physics Education*, 49 (4), S. 412–418.
- Klein, P., Kuhn, J., Müller, A. & Gröber, S. (2014). Video analysis exercises in regular introductory mechanics physics courses: Effects of conventional methods and possibilities of mobile devices. In: A. Kauertz, H. Ludwig, A. Müller, J. Pretsch & W. Schnotz (Hrsg.), *Multidisciplinary research on teaching and learning*, Basingstoke: Palgrave Macmillan.
- Klein, P., Gröber, S., Kuhn, J. & Müller, A. (2013). Video analysis of projectile motion using tablet computers as experimental tool. *Physics Education*, 49 (1), S. 37–40.

### Nicht-referierte / herausgeber-referierte Beiträge (Auswahl):

- Klein, P., Gröber, S., Kuhn, J. & Wilhelm, T. (2015). Bewegungen mit der App Video Analysis analysieren. *Physik in unserer Zeit*, 46 (2), S. 98–99.
- Klein, P., Gröber, S., Kuhn, J. & Müller, A. (2014). Mobile Videoanalyse mit Tablets am Beispiel des freien Falls mit Luftreibung. *Praxis der Naturwissenschaften — Physik in der Schule*, 63 (5), S. 33–35.
- Klein, P., Gröber, S., Kuhn, J. & Müller, A. (2013). Experimentelle Aufgaben in den Übungen zur Experimentalphysik 1. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht* (S. 312–314). Beiträge zur GDGP-Jahrestagung in München 2013, Kiel: IPN.